2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Learning Composable Skills by Discovering Spatial and Temporal Structure with Foundation Models

Neil Nie[1], Wenlong Huang[1], Jiayuan Mao[2], Li Fei-Fei[1], Weiyu Liu[1†], Jiajun Wu[1†]
[1]Stanford University [2]MIT

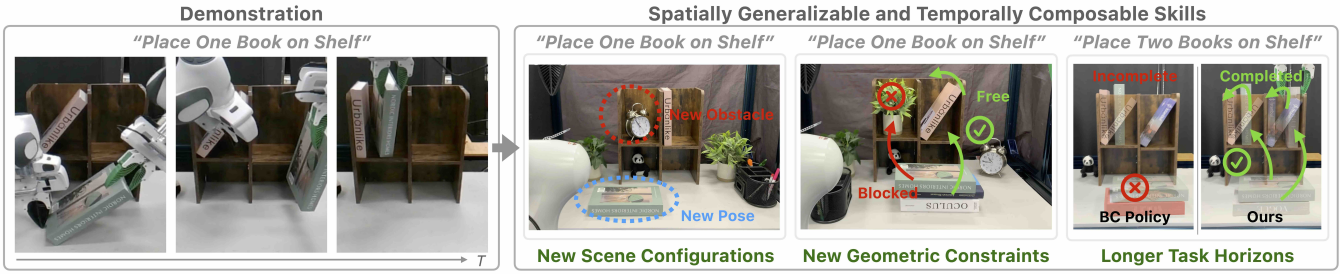

Fig. 1: STACK, a framework that discovers and learns composable skills with foundations models.

***Abstract*— We present STACK, a framework for discovering and learning composable manipulation skills from unsegmented demonstrations by leveraging *spatial and temporal structure* extracted from foundation models. STACK automatically extracts *temporal structure* by segmenting raw demonstrations into short-horizon skills using a video-language model, and *spatial structure* by identifying skill-relevant elements in 3D point cloud observations. For each discovered skill, we learn a diffusion-based trajectory sampler and a skill effect model, both of which operate in the reference frame of the relevant scene element. At test time, given a language goal, STACK segments the 3D scene, samples skill trajectories, and composes them by simulating geometric effects. This enables generalization to new scene configurations, geometric constraints, and longer task horizons beyond training across diverse real-world manipulation tasks. Project page: https://icra-stack.github.io**

I. Introduction

Humans effortlessly perform long-horizon manipulation tasks. Take the everyday task of storing books on a shelf: we can pick up books from the table regardless of their placements, adapt to new obstacles on the shelf by adjusting where books go, and plan ahead by mentally simulating different layouts to avoid clutter. In contrast, robots today cannot handle such diverse task variations.

A prominent way to enable diverse behaviors in long-horizon tasks is by training robots on human demonstration data spanning a range of task variations [1, 2]. However, collecting such data requires substantial effort [3]. To address this, prior work shows how incorporating *structure* can improve generalization from limited demonstrations. In particular, *temporal structure* helps by decomposing demonstrations into short-horizon *composable* skills (e.g., grasp, push, place), that can be recomposed into new sequences to solve novel task variations. *Spatial structure* helps learned skills generalize to new environmental states by focusing only on relevant elements of the environment. For example, picking up a book should be invariant to the book's placement and unaffected by the background. The challenge is that these structures are typically hand-designed, and most existing approaches do not automatically infer them from demonstrations [4–7].

Our insight is that video-language models can automatically extract *spatial and temporal structure* from demonstrations. They possess high-level semantic understandings of tasks, as well as low-level spatial and temporal understandings of how objects and robots interact. For example, these models can temporally localize the coordinated actions involved in picking up a book from a table with two robot arms, even when multiple contacts are involved. Furthermore, these models can identify relevant scene elements and help capture spatial invariance specific to each skill. For example, grasping is invariant to the pose of the book, while placing should account for the pose of the bookshelf.

Building on this insight, we introduce STACK, a framework that leverages spatial and temporal structure from foundation models for learning composable skills. For each discovered skill, we train a diffusion-based trajectory sampler to model a multimodal distribution of actions in the reference frame of the relevant scene element identified by the video-language model. We also train a skill effect model to predict the future state of the skill-relevant elements. These skills can be composed by simulating and updating the scene geometry based on the predicted effects. At test time, given a natural language goal and an initial scene observation, the system uses a vision-language model to generate candidate plans, segments the scene to localize relevant elements, samples trajectories from the learned skills, and composes them by reasoning over feasibility using the effect model. This enables generalizable behavior across a range of long-horizon manipulation tasks.

In summary, our contributions are:

- We show that foundation models can extract *spatial and temporal structure* from unsegmented demonstrations, eliminating the need for manual annotations.
- We introduce a framework for learning skill samplers and their geometric effects, enabling geometry-aware skill composition.

† denotes equal advising.

- We demonstrate that our method enables long-horizon, bimanual, and non-prehensile manipulation in real-world settings, with strong generalization to novel task variations.

II. Related Work

Visuomotor Skill Learning. Imitation learning is an prominent approach to learning visuomotor skills from demonstrations [8]. Diffusion-based BC improves multi-modal action modeling [9, 10] but tends to overfit to scene layouts seen in the demonstrations. Recent works tackle this by scaling up data [11] and generating synthetic demonstrations [12–14]. In parallel, zero-shot labeling with foundation models reduces annotation cost at scale [15], and large multi-embodiment datasets broaden task and object diversity [16]. These approaches often yield a single policy that struggles to generalize beyond the training horizon. To address this limitation, some methods use RL to discover and learn such skills [17–19], but are restricted to simulation. Other methods [2, 20–22] discover skills unsupervised from demonstration and play data. However, these methods lack explicit reasoning about geometric dependencies between skills, imperative for effective long-horizon, sequential manipulation.

Compositional Generalization for Robot Manipulation. Generalization to novel states and goals by recombining previously learned skills is essential for robot manipulation. Toward temporal compositionality, works introduce hierarchical structures [2, 23–25] and language-guided temporal segmentation [26]. However, they often fail to model spatial generalization, limiting them to following sequential goals. Symbolic action representations also compose skills, including manual abstractions [27–30], learned abstractions [31–33], and approaches with added segmentation and language annotations [5].

Recent works have also explored incorporating object-centric priors for improving spatial compositional generalization in learning-based policies and models [6, 34–38]. By contrast, we aim to construct spatiotemporal compositional action representations directly from data using the rich priors of foundation models. Crucially, it explicitly reasons about geometric constraints, thereby avoiding the limitations of purely discrete symbolic representations in physical domains.

Foundation Models for Robotics. Foundation models, pre-trained on Internet-scale data, have internalized rich semantics priors useful for robotics applications [39–42]. A recent body of literature adapts vision-language models (VLMs) as vision-language-action (VLA) models for end-to-end visuomotor learning [43–47].

In parallel, works introduce structural priors, leveraging foundation models for open-world perception [5, 48, 49], goal interpretation and reward specification [50–54], model or domain specification [55–57], (visual) dynamics prediction [58–61], and high-level decision-making [4, 5, 62–64]. VLMs have also been shown to generate meaningful subgoals for task and motion planning [65] and robotic affordance-grounded behaviors [4], highlighting their potential for temporal decomposition and long-horizon reasoning. Whereas prior works rely on carefully designed abstractions to apply off-the-shelf foundation models, we uniquely explore the question of *whether foundation models can provide such structural abstractions themselves* by discovering spatial and temporal structure from raw demonstrations.

III. Problem Formulation

Formally, the environment is modeled as a tuple $\langle \mathcal{X}, \mathcal{U}, \mathcal{T}, \mathcal{O} \rangle$ where $\mathcal{X}$ is the inaccessible underlying state space, $\mathcal{U}$ is the low-level action space of 6-DoF end-effector poses and gripper actions, $\mathcal{T}$ is the unknown transition function, and $\mathcal{O}$ denotes the partial point cloud observation space of the environment. We consider the learning from demonstrations setting, where a dataset $\mathcal{D} = \left\{(\ell^i, \mathbf{o}^i, \mathbf{u}^i)\right\}_{i=1}^{N}$ consisting of N unsegmented demonstrations is given, each with a natural language task description ℓ^i, observation sequence $\mathbf{o}^i = (o_0^i, \ldots, o_{T_i}^i)$, and action sequence $\mathbf{u}^i = (u_1^i, \ldots, u_{T_i}^i)$.

From $\mathcal{D}$ we aim to automatically discover a library of *skills*, denoted by Π. Importantly, these skills operate on *factorized observation*. Specifically, each o is decomposed into a set of entities $\{e_1, ..., e_n\}$, denoted by $\mathcal{E}$. Each entity $e \in \mathcal{E}$ refers to a task-relevant object or part in the scene, represented by its name and its segmented point cloud p_e. Each skill s is represented as a tuple $\langle \mathrm{name}, \mathrm{args}, \pi_s, f_s \rangle$. Here, name is a natural language description summarizing the skill's intended behavior. The arguments args specifies the entities involved in the skill, defining a subset $\mathcal{E}_s \subseteq \mathcal{E}$; The trajectory sampler $\pi_s : \mathcal{O}_{\mathcal{E}_s} \to \mathcal{U}^{H_s}$ operates on the factorized observation and outputs a sequence of H_s low-level actions; if no valid trajectory exists, it returns an empty sequence; The skill effect model $f_s \colon \mathcal{O}_{\mathcal{E}_s} \times \mathcal{U}^{H_s} \to \mathcal{O}_{\mathcal{E}_s}$ predicts the resulting entities' observations based on the executed trajectory.

At deployment time, given a language goal ℓ for a novel long-horizon task and the initial observation $\bar{o}_0$, we are interested in obtaining the low-level action sequence $u_{1:T}$, represented by a sequence of K skills, each with a trajectory τ_s sampled from its learned trajectory sampler π_s, while subjecting to transition feasibility, kinematics and collision constraints $\mathcal{G}$, and the initial observation. Formally,

$$\min_{K \in \mathbb{N},\ \pi_{0:K-1} \in \Pi,\ \tau_{0:K-1}} \mathcal{L}_{\mathrm{goal}}(\ell, o_K)$$

$$\begin{aligned} \text{s.t.} \quad & \tau_k \sim \pi_k(\cdot \mid o_k), && \forall\, k = 0, \ldots, K-1, \\ & o_{k+1} = f_{\pi_k}(o_k, \tau_k), && \forall\, k = 0, \ldots, K-1, \\ & \mathcal{G}(o_k, \tau_k) \le 0, && \forall\, k = 0, \ldots, K-1, \\ & o_0 = \bar{o}_0 && \text{(initial condition)} \end{aligned}$$

where $\mathcal{L}_{\mathrm{goal}}$ is a goal function consumed implicitly by the method. In this work, we particularly focus on novel states and goals that are outside of $\mathcal{D}$, which necessitates both a deliberate and reusable choice of Π and a careful consideration of their geometric dependencies enforced by the learned effect model f, which collectively enables skill compositions towards completing unseen goals.

IV. Method

As shown in Fig. 2, STACK extracts *spatial and temporal structure* from demonstrations to learn composable manipulation skills. It uses a video-language model for *temporal*

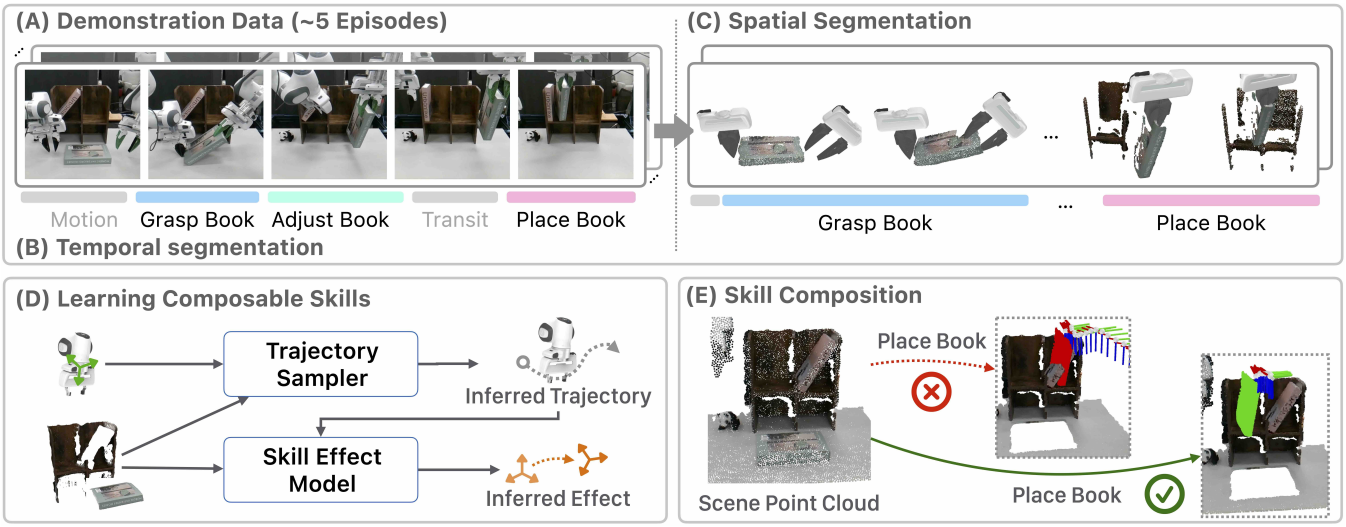

Fig. 2: Overview of STACK. (A) Given a small number of demonstrations, (B) a video-language model segments them into individual skills, (C) vision foundation models extracts point clouds for entities relevant to each skill, (D) a trajectory sampler and an effect model are trained for each skill, and (E) at test time, the learned skills are composed using predicted effects to solve novel long-horizon tasks.

segmentation, dividing demonstrations into skill segments and generating natural language descriptions. It then performs *spatial segmentation* with vision foundation models to extract point clouds of skill-relevant entities from the descriptions. A trajectory sampler and effect model are trained for each skill, conditioned on the segmented entities. At test time, these skills are composed to solve novel long-horizon tasks.

A. Discovering Skills using Foundation Models

For each raw demonstration, STACK first extracts temporal skill boundaries. Then, it identifies relevant entities for each skill, producing a name and segmented object point cloud for each entity e. Finally, the entity names are lifted to form the arguments args for each skill s.

Temporal segmentation. We use a video-language model (Video-LM) to extract temporal segments corresponding to skills from unsegmented demonstrations. For each demonstration, the Video-LM receives the task goal ℓ, video frames from the demonstration, and proprioceptive cues (i.e., gripper width and joint torques). For bimanual scenes, we overlay transparent color-coded arm masks on the video frames. The Video-LM then returns a set of skills with start and end timestamps for each discovered skill in `mm:ss` format, and a natural-language summary of the skills' intended behaviors. To improve accuracy, we perform skill segmentation in two stages. The proposal stage produces initial coarse candidate segments. In the refinement stage, the proposed skill timestamps, along with all demonstration videos, are re-supplied to the model for further refinement.

Task-relevant entity discovery. For every discovered skill segment, we automatically extract the name of the relevant entities $\mathcal{E}_s$ and extract their point clouds $\{p_e \ \forall \ e \in \mathcal{E}_s\}$. This process involves (i) extracting entity names using the Video-LM, and (ii) extracting corresponding 3D point clouds using vision foundation models. We query the Video-LM with each proposed skill segment to obtain the names of relevant entities for that skill. The model also provides coarse attributes such as color and spatial location, which are included in the entity name. This allows us to automatically derive object descriptions without manual labeling. Each name then prompts open-vocabulary detectors [66, 67] on the segment's first frame to extract entities' bounding boxes. These bounding boxes then prompt SAM2 [68], which segments and tracks the entity through the video. Entity point clouds are reconstructed from RGB-D images and the generated segmentation masks, and aligned to a common world frame. We apply outlier removal and farthest point sampling to obtain a clean, downsampled point cloud.

B. Learning Composable Skills

Given each discovered skill s and the proposed argument set $\mathcal{E}_s$, we learn trajectory samplers that generalize across spatial configurations and can be composed to solve long-horizon tasks. Each sampler produces a short trajectory that completes a local skill.

Learning trajectory samplers. Each trajectory sampler $\pi_s : \mathcal{O}_{\mathcal{E}_s} \rightarrow \mathcal{U}^{H_s}$ in a skill maps the segmented point clouds $\{p_e \ \forall \ e \in \mathcal{E}_s\}$ of the relevant entities $\mathcal{E}_s$ to a sequence of H_s low-level actions, conditioned on robot proprioception. The output trajectory is represented as a sequence of end-effector poses. Many skills allow multiple valid trajectories. Therefore, we parameterize π as a generative diffusion model that can sample diverse, valid trajectories. The model is designed to be general across skill types. The model input supports different numbers of entities, different numbers of arms involved, and whether the skill depends on the previous end-effector state. Additional details are provided in the appendix.

Learning skill effect models. The skill effect model, $f_s : \mathcal{O}_{\mathcal{E}_s} \times \mathcal{U}^{H_s} \rightarrow \mathcal{O}_{\mathcal{E}_s}$, predicts how the execution of a skill transforms the given entities. Given a sampled trajectory and the point clouds $\{p_e \ \forall \ e \in \mathcal{E}_s\}$, the model predicts a rigid-body transform $T \in SE(3)$ for each $e \in \mathcal{E}_s$ such that the updated point cloud is given by $p'_e = Tp_e$. This prediction allows us to perform test-time filtering of infeasible trajectories by considering collision and kinematics constraints, enabling long-horizon planning. For computing

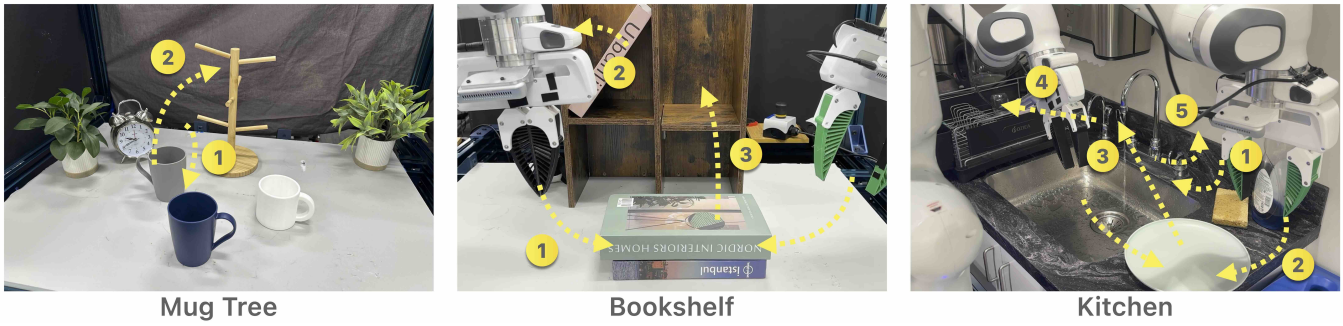

Fig. 3: Task Domains. Three domains involve non-prehensile and bimanual manipulation tasks. The numbered yellow arrows illustrate the sequence of steps in the human demonstrations. The robot must handle 8 different object types across these domains.

the labels, we take the extracted object point clouds at the initial and final frames of each skill segment and compute a rigid transform in $SE(3)$ using Iterative Closest Point (ICP) [69] alignment, which is used to supervise the effect model's prediction of this transformation from the initial point clouds and the sampled trajectory. To improve robustness, we further apply data augmentations during training to simulate geometric variations. Details are provided in the appendix.

Spatial augmentation. Many skills and their effects are spatially invariant. For example, if a bookshelf is moved by a particular transform, the trajectory for placing a book and its final pose can be adjusted by applying the same transform. Simply using point clouds for the entities is not sufficient to capture this spatial invariance. To address this, we perform fully automatic augmentation of the demonstration data by transforming the entity point clouds, skill trajectories, and final entity poses simultaneously. Specifically, we first transform the observations and actions into a common entity reference frame, defined by the anchor entity (e.g., the mug tree for the *hang* skill). The anchor entity is identified by the Video-LM during the entity discovery process. We then sample an augmentation transformation to apply in this frame. The same set of augmentation hyperparameters is applied across all tasks and domains without any task-specific tuning. To simulate out-of-distribution point cloud observations caused by extreme object placements at test time, we additionally remove points that are not within the direct line of sight of any camera, either due to occlusion or being outside the field of view.

C. Planning with Learned Skills

We use a vision-language model to propose candidate skill sequences for a new scene given a task goal ℓ in natural language and the learned skills. The names of the entities relevant to each skill are used to extract corresponding point clouds via open-vocabulary detection and segmentation. To instantiate each skill sequence, we perform a tree search beginning with the first skill. For each skill s, we generate M trajectories using the learned sampler π_s and predict their effects using the effect model f_s. Collisions are detected by computing pairwise distances between entities' partial point clouds with a fixed threshold via a k-d tree [70]. For each valid trajectory, we update the scene geometry based on the predicted effect; the next skill then uses this updated observation to generate its own trajectories and effects.

TABLE I: Temporal Segmentation Results. STACK outperforms all baselines with better alignment to ground truth and fewer spurious segments. Performance drops without proprioception or our prompts. For **ES** and **MoF**, higher values indicate better performance; for **#IS**, lower values are better.

Method	Mug Tree			Bookshelf			Kitchen		
	ES	**MoF**	**#IS**	**ES**	**MoF**	**#IS**	**ES**	**MoF**	**#IS**
UVD	0.55	0.29	2.6	0.42	0.18	4	0.41	0.19	7.2
Contact Heuristic	0.50	0.53	1.2	0.50	0.31	0.36	0.70	0.57	0.16
Video-LM ZS	**0.76**	**0.68**	**0**	0.52	0.47	0.1	0.59	0.31	1
STACK (ours)	**0.76**	**0.68**	**0**	**0.84**	**0.76**	**0**	**0.77**	**0.72**	**0**

To transition between the learned skills, we introduce a special *transit* skill for free-space arm motion and a *move* skill for transporting a grasped object, assuming rigid attachment between the object and the gripper during the transition. These transition skills produce a sequence of robot actions and update the scene geometry accordingly. In practice, we first sample the learned skills and use cuRobo [71] to compute collision-free paths connecting them.

V. Experiments

A. Task Domains and Generalization Axes

As shown in Fig. 3, we evaluate STACK on three domains involving single-arm or bimanual configurations of Franka Emika robots. Our robot setup includes one or two calibrated Intel RealSense RGB-D cameras and a GELLO [72] teleoperation interface for collecting human demonstrations. For each domain, we design test cases to investigate three types of generalization: new scene configurations (**SC**), new geometric constraints (**GC**), and longer task horizons (**LH**).

In the *Mug Tree* domain, we collect five demonstrations of hanging a mug on each of five pegs. The *Bookshelf* domain includes ten bimanual demonstrations of inserting a book into an occupied compartment of the bookshelf by pushing existing books. The *Kitchen* domain includes ten demonstrations of washing a plate and placing it on a rack, involving articulated manipulation of the faucet lever and handling liquids. Across all domains, **SC** varies relevant object poses, **GC** adds obstacles or kinematic infeasibility, and **LH** introduces multiple objects, requiring generalization beyond the training horizon.

TABLE II: Generalization Results. Average partial success rate across 10 trials for baselines and ablations on three generalization cases: new scene configurations (**SC**), new geometric constraints (**GC**), and longer task horizons (**LH**) in three domains.

Method	Mug Tree			Bookshelf			Kitchen		
	SC	GC	LH	SC	GC	LH	SC	GC	LH
DP3-LONG [10]	0	0	0	0	0	0	0	0	0
DP3-SHORT [10]	16.7	10	0	0	0	0	0	0	0
BLADE [5]	86.7	70	67.1	**90.0**	46.7	35.0	**85.0**	65.0	68.8
STACK (ours)	**90.0**	**80.0**	**77.4**	**90.0**	**93.3**	**86.1**	**85.0**	**75.0**	**71.6**
w/o Spatial Aug.	50.0	50.0	14.6	50.0	40.0	23.6	80.0	**75.0**	64.5
w/o Effect	83.3	75.0	56.7	85.0	47.9	67.5	70.0	60.0	67.9

B. Evaluating Skill Segmentation

To evaluate the accuracy of *temporal segmentation*, we compare predicted segment labels with manually annotated ground truth using three metrics. Edit Score (ES) [73] measures the normalized edit distance between predicted and ground-truth segment sequences, evaluating structural alignment. Mean over Frames (MoF) [74] computes the proportion of frames with correct segment labels, evaluating frame-wise accuracy. We also report #IS, the average number of missing or extra segments relative to ground truth. In addition, we evaluate *entity point cloud segmentation* against expert-labeled ground truths and use standard metrics including Mean Class Accuracy (mAcc) and Mean Intersection over Union (mIoU).

Baselines. For *temporal segmentation*, we compare against two baselines and a variant of our method. UVD [75] extracts skill segments from demonstration videos by detecting phase shifts in the embedding space of a pretrained visual representation. Contact Heuristic is a common strategy in prior work that segments demonstrations based on proprioceptive signals to identify keyframes or skill boundaries. We also evaluate STACK variant Video-LM ZS, which uses the same video-language model for temporal segmentation in zero shot.

Results. As shown in Table I, our method outperforms both baselines across all metrics and domains for *temporal segmentation*. Performance drops significantly in the *Bookshelf* and *Kitchen* domains when removing proprioceptive signals and two-stage processing, validating that our method is crucial for accurately extracting skill segments in complex and realistic manipulation scenarios. Compared to the baselines, our temporal segmentation method achieves an #IS score of 0, indicating no missing or extra segments, which is important for learning all necessary skills. Our trajectory samplers and skill effect models are robust to imprecise temporal boundaries, which is validated in the following robot experiments, where all skills are automatically discovered by video-language models rather than relying on ground-truth segments. For *entity point cloud segmentation*, our method achieves 97% mAcc and 91% mIoU across all frames and demonstrations, showcasing robust performance.

C. Evaluating Generalization in Manipulation

We evaluate all methods using partial success rates, which measure the fraction of key steps completed in a trial. For example, in the task of hanging four mugs, each successfully hung mug contributes 25%. For each generalization type in each domain, we design two test cases and run five trials per case. In total, we conducted 540 real-world trials across domains and methods.

Baselines. We compare against end-to-end visuomotor policies and a planning method. For the visuomotor baselines, we adapt prior work [10] that uses 3D point cloud observations for spatial generalization. DP3-LONG is trained on all demonstrations for each domain and rolled out continuously until timeout to handle tasks with longer horizons than seen during training. DP3-SHORT is trained on demonstrations for individual skill segments extracted by our method; since it lacks planning capabilities, we provide the ground-truth skill sequence. We also compare with a planning baseline BLADE adapted from recent work [5], which achieves strong performance on long-horizon tasks. This method defines symbolic preconditions and effects for each of our learned skills, grounds them on image observations via a VLM, and performs task planning with an off-the-shelf planner.

We study several ablations of our method. w/o Spatial Aug. removes the spatial augmentations applied during training of the skill samplers and skill effect models to evaluate their impact on generalization. w/o Effect removes the skill effect models at test time, executing skills sequentially without predicting future states, to isolate the role of rejection sampling based on learned effects.

STACK achieves strong spatial generalization. In the **SC** setting shown in Fig 4 (top), STACK achieves significantly better performance than the end-to-end visuomotor policies DP3-LONG and DP3-SHORT. DP3-LONG, trained on unsegmented demonstrations, fails across all domains. DP3-SHORT achieves non-zero performance on the *Mug Tree* domain by separately learning the *grasp* and *hang* skills that were discovered through our segmentation method. The ablation without spatial augmentation (w/o Spatial Aug.) shows that spatial segmentation alone enables our sampler to outperform DP3-SHORT by 54%. When combined with spatial augmentation, our full method gains an additional 28% improvement. The spatial generalization results suggest that spatial segmentation and augmentation offer complementary benefits for generalizing skills to novel scene configurations.

STACK can handle new geometric constraints. In the **GC** setting, end-to-end visuomotor policies perform poorly. STACK outperforms the planning baseline BLADE, which relies on symbolic representations of the environment and therefore fails to capture the geometric feasibility of skills. For example, in the *Bookshelf* domain, BLADE fails because the vision-language model cannot reliably classify the predicate *is-compartment-blocked*(compartment) when a book is partially blocking the compartment. In contrast, STACK solves these cases by sampling diverse trajectories and predicting their effects for each skill. This allows the model to reject trajectories kinematically infeasible or that lead to future collisions (e.g., placing a book in a partially blocked compartment Fig. 4 (middle)). Removing the skill effect model (w/o Effect) removes the feasibility check and leads to a 22% average drop in performance. Spatial augmentation is also crucial for

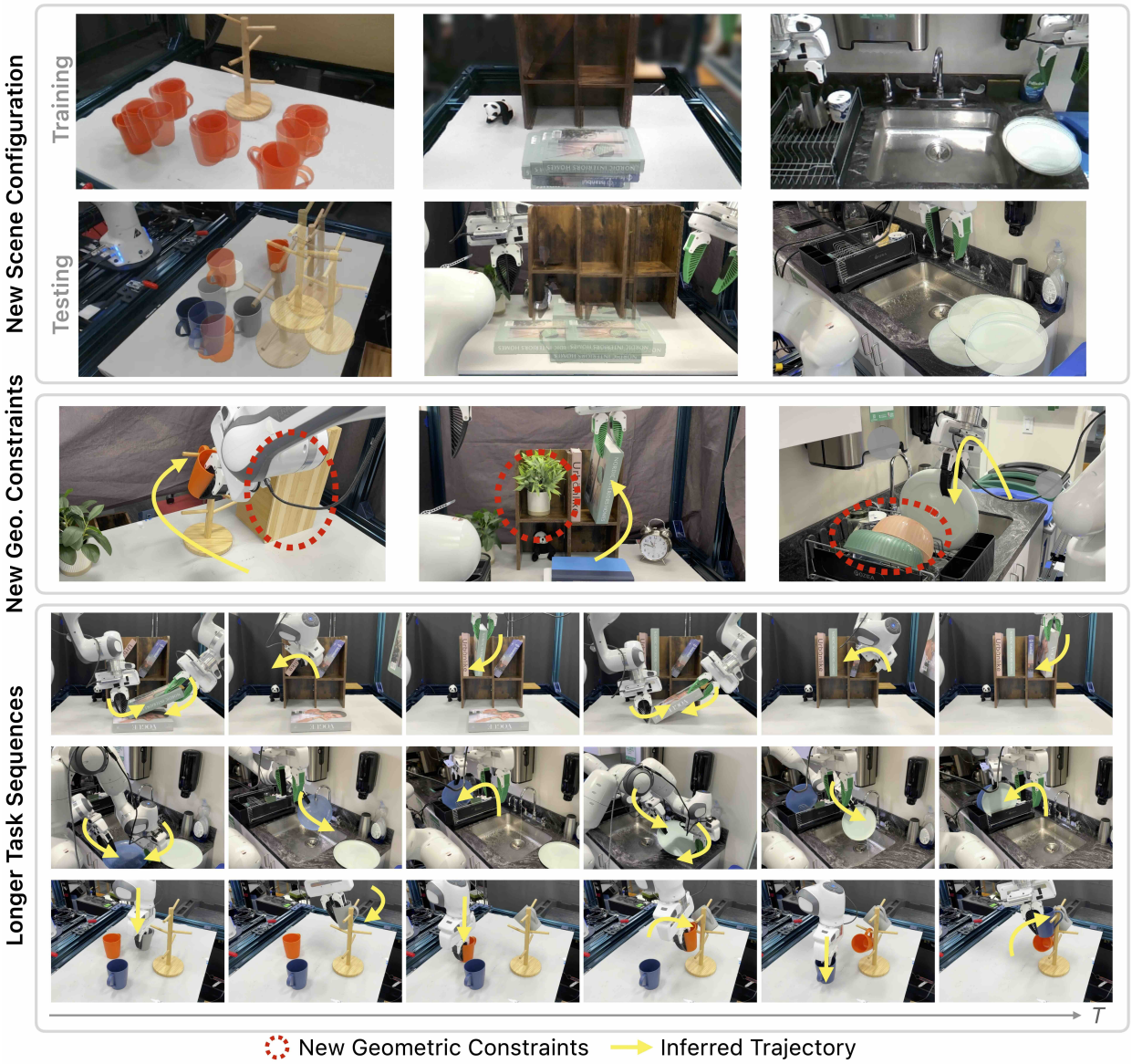

Fig. 4: Real-World Generalization. In the top row, we visualize the distribution shift in object poses between training and testing by showing overlaid examples from each set. In the middle row, yellow arrows indicate trajectories generated by STACK that successfully avoid test-time geometric constraints, highlighted with red dashed outlines. Compared to Fig. 3, the bottom row highlights generalization to longer task horizons.

learning samplers and effect models that generalize beyond training object poses. Without it, performance drops by 28%.

STACK can complete long-horizon tasks by composing learned skills. **LH** tasks combine the challenges of both **SC** and **GC** generalization. As shown in Fig. 4 (bottom), when multiple objects are involved, the scenes often contain novel object poses, and successful execution requires anticipating future interactions to avoid collisions. For example, both BLADE, which models symbolic dependencies between skills, and the w/o Effect variant fail by attempting to hang two mugs on the same peg midway through the task, resulting in a collision. These limitations prevent them from solving longer-horizon tasks. As a result, they fall behind our method by 21% and 14%, respectively. Our method demonstrates graceful performance degradation as the task horizon increases. While **LH** tasks can involve up to six skill steps, we maintain a minimum of 70% partial success rate, showing the robustness of our approach in handling increased task complexity.

STACK works on a diverse set of tasks in the real-world . Across all real-world domains, *Mug Tree*, *Bookshelf*, and *Kitchen* as shown in Fig. 4, STACK achieves strong performance using a handful of demonstrations per task, highlighting its data efficiency. Despite substantial differences in embodiment (single-arm vs. bimanual), object types (rigid, articulated), and skill requirements (e.g., in-place insertion, contact-rich washing), our method performs well without domain-specific tuning. This is enabled by leveraging *spatial and temporal structure*, generative trajectory sampler, and learned effect predictions. These components allow STACK to robustly acquire and compose reusable skills across structurally and semantically distinct tasks.

Failure analysis. Finally, we conduct a detailed failure

analysis. Most failures arise from skill execution (51.5%), followed by object detection (27.3%) and segmentation (21.2%). The skills *turn off faucet* and *hang mug* are particularly challenging, since the former requires applying force and the latter has low error tolerance. Given the multi-step nature of our tasks, 13.7% of failures are caused by error accumulation. These results highlight key challenges and suggest directions for improvement.

VI. Conclusion and Discussion

STACK is a novel framework that discovers and learns *composable* manipulation skills from a handful of unsegmented demonstrations by leveraging *spatial and temporal structure* extracted using foundation models. STACK automatically segments long-horizon demonstrations into composable skills by combining temporal structure from video-language models with spatial structure from vision foundation models. For each skill, it learns a conditional diffusion policy and an effect predictor, enabling flexible and robust skill composition under geometric constraints. Our experiments demonstrate that STACK generalizes to new scene configurations, can handle new geometric constraints, and can solve longer horizons tasks beyond those in training, outperforming end-to-end policies and approaches based on symbolic planners.

Limitations. STACK currently assumes rigid-body and articulated object interactions, and does not handle deformable objects such as cloth. Our evaluations focus on open-loop skill composition. While the learned trajectory samplers operate at around 2–5 Hz and STACK supports replanning after each execution, extending the approach to dynamic tasks is left for future work. STACK can also be extended to handle stochasticity by leveraging diffusion-based effect models to capture multiple outcomes, and by incorporating probabilistic planners to reason under uncertainty. As task horizons increase, planning requires solving complex constraint satisfaction problems, which becomes computationally expensive. To ensure scalability, hierarchical planning or mixed planning-and-execution strategies [76] may be necessary.

Acknowledgment

This work is in part supported by Analog Devices, AFOSR YIP FA9550-23-1-0127, ONR N00014-23-1-2355, ONR YIP N00014-24-1-2117, ONR MURI N00014-24-1-2748, and NSF RI #2211258.

References

[1] A. Mandlekar, D. Xu, R. Martín-Martín, S. Savarese, and L. Fei-Fei, "Learning to generalize across long-horizon tasks from human demonstrations," *arXiv preprint arXiv:2003.06085*, 2020. 1

[2] C. Wang, L. Fan, J. Sun, R. Zhang, L. Fei-Fei, D. Xu, Y. Zhu, and A. Anandkumar, "Mimicplay: Long-horizon imitation learning by watching human play," in *CoRL*, 2023. 1, 2

[3] T. Z. Zhao, J. Tompson, D. Driess, P. Florence, K. Ghasemipour, C. Finn, and A. Wahid, "Aloha unleashed: A simple recipe for robot dexterity," *arXiv preprint arXiv:2410.13126*, 2024. 1

[4] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes *et al.*, "Do as i can and not as i say: Grounding language in robotic affordances," in *arXiv preprint arXiv:2204.01691*, 2022. 1, 2

[5] W. Liu, N. Nie, R. Zhang, J. Mao, and J. Wu, "Blade: Learning compositional behaviors from demonstration and language," in *CoRL*, 2024. 1, 2, 5

[6] S. Cheng, C. Garrett, A. Mandlekar, and D. Xu, "Nod-tamp: Multi-step manipulation planning with neural object descriptors," *arXiv preprint arXiv:2311.01530*, 2023. 1, 2

[7] A. Simeonov, Y. Du, A. Tagliasacchi, J. B. Tenenbaum, A. Rodriguez, P. Agrawal, and V. Sitzmann, "Neural descriptor fields: Se (3)-equivariant object representations for manipulation," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 6394–6400. 1

[8] C. Finn, S. Levine, and P. Abbeel, "Guided cost learning: Deep inverse optimal control via policy optimization," in *ICML*. PMLR, 2016, pp. 49–58. 2

[9] C. Chi, S. Feng, Y. Du, Z. Xu, E. Cousineau, B. Burchfiel, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," in *Proceedings of Robotics: Science and Systems (RSS)*, 2023. 2

[10] Y. Ze, G. Zhang, K. Zhang, C. Hu, M. Wang, and H. Xu, "3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations," in *RSS*, 2024. 2, 5

[11] A. Khazatsky, K. Pertsch, S. Nair, A. Balakrishna, S. Dasari, S. Karamcheti, S. Nasiriany, M. K. Srirama, L. Y. Chen, K. Ellis *et al.*, "Droid: A large-scale in-the-wild robot manipulation dataset," *arXiv preprint arXiv:2403.12945*, 2024. 2

[12] A. Mandlekar, S. Nasiriany, B. Wen, I. Akinola, Y. Narang, L. Fan, Y. Zhu, and D. Fox, "Mimicgen: A data generation system for scalable robot learning using human demonstrations," in *7th Annual Conference on Robot Learning*, 2023. 2

[13] Z. Xue, S. Deng, Z. Chen, Y. Wang, Z. Yuan, and H. Xu, "Demogen: Synthetic demonstration generation for data-efficient visuomotor policy learning," in *RSS*, 2025. 2

[14] Z. Jiang, Y. Xie, K. Lin, Z. Xu, W. Wan, A. Mandlekar, L. Fan, and Y. Zhu, "Dexmimicgen: Automated data generation for bimanual dexterous manipulation via imitation learning," in *ICRA*, 2025. 2

[15] N. Blank, M. Reuss, M. Rühle, Ömer Erdinç Yağmurlu, F. Wenzel, O. Mees, and R. Lioutikov, "Scaling robot policy learning via zero-shot labeling with foundation models," 2024. 2

[16] K. Wu, C. Hou, J. Liu, Z. Che, X. Ju *et al.*, "Robomind: Benchmark on multi-embodiment intelligence normative data for robot manipulation," 2025. 2

[17] G. Konidaris and A. Barto, "Skill discovery in continuous reinforcement learning domains using skill chaining," *Advances in neural information processing systems*, vol. 22, 2009. 2

[18] A. Sharma, S. Gu, S. Levine, V. Kumar, and K. Hausman, "Dynamics-aware unsupervised discovery of skills," 2020. 2

[19] S. Rho, L. Smith, T. Li, S. Levine, X. B. Peng, and S. Ha, "Language guided skill discovery," *arXiv preprint arXiv:2406.06615*, 2024. 2

[20] Y. Zhu, P. Stone, and Y. Zhu, "Bottom-up skill discovery from unsegmented demonstrations for long-horizon robot manipulation," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 4126–4133, 2022. 2

[21] S. Nair and C. Finn, "Hierarchical foresight: Self-supervised learning of long-horizon tasks via visual subgoal generation," in *ICLR*, 2020. 2

[22] M. Xu, Z. Xu, C. Chi, M. Veloso, and S. Song, "Xskill: Cross embodiment skill discovery," in *CoRL*, 2023. 2

[23] L. X. Shi, Z. Hu, T. Z. Zhao, A. Sharma, K. Pertsch, J. Luo, S. Levine, and C. Finn, "Yell at your robot: Improving on-the-fly from language corrections," *arXiv:2403.12910*, 2024. 2

[24] J. Luo, C. Xu, X. Geng, G. Feng, K. Fang, L. Tan, S. Schaal, and S. Levine, "Multi-stage cable routing through hierarchical imitation learning," *IEEE Transactions on Robotics*, 2024. 2

[25] O. Mees, J. Borja-Diaz, and W. Burgard, "Grounding language with visual affordances over unstructured data," in *ICRA*, 2023. 2

[26] D. Raj, O. Patil, W. Gu, C. Baral, and N. Gopalan, "Learning temporally composable task segmentations with language," in *IROS*, 2024. 2

[27] S. Srivastava, E. Fang, L. Riano, R. Chitnis, S. Russell, and P. Abbeel, "Combined Task and Motion Planning through an Extensible Planner-Independent Interface Layer," in *ICRA*, 2014. 2

[28] M. Toussaint, "Logic-Geometric Programming: An optimization-based approach to combined task and motion planning," in *IJCAI*, 2015. 2

[29] A. Curtis, X. Fang, L. P. Kaelbling, T. Lozano-Pérez, and C. R. Garrett, "Long-horizon manipulation of unknown objects via task and motion planning with estimated affordances," in *ICRA*, 2022. 2

[30] Z. Yang, C. R. Garrett, T. Lozano-Pérez, L. Kaelbling, and D. Fox, "Sequence-based plan feasibility prediction for efficient task and motion planning," in *RSS*, 2023. 2

[31] G. Konidaris, L. P. Kaelbling, and T. Lozano-Perez, "From skills to symbols: Learning symbolic representations for abstract high-level

planning," *Journal of Artificial Intelligence Research*, vol. 61, pp. 215–289, 2018. 2

[32] T. Silver, R. Chitnis, N. Kumar, W. McClinton, T. Lozano-Pérez, L. Kaelbling, and J. B. Tenenbaum, "Predicate invention for bilevel planning," in *AAAI*, 2023. 2

[33] A. Ahmetoglu, E. Oztop, and E. Ugur, "Symbolic manipulation planning with discovered object and relational predicates," *arXiv preprint arXiv:2401.01123*, 2024. 2

[34] Y. Zhu, A. Joshi, P. Stone, and Y. Zhu, "Viola: Imitation learning for vision-based manipulation with object proposal priors," *6th Annual Conference on Robot Learning*, 2022. 2

[35] P. Battaglia, R. Pascanu, M. Lai, D. Jimenez Rezende *et al.*, "Interaction networks for learning about objects, relations and physics," *Advances in neural information processing systems*, vol. 29, 2016. 2

[36] E. Jang, C. Devin, V. Vanhoucke, and S. Levine, "Grasp2vec: Learning object representations from self-supervised grasping," *arXiv preprint arXiv:1811.06964*, 2018. 2

[37] F. Locatello, D. Weissenborn, T. Unterthiner, A. Mahendran, G. Heigold, J. Uszkoreit, A. Dosovitskiy, and T. Kipf, "Object-centric learning with slot attention," *Advances in neural information processing systems*, vol. 33, pp. 11 525–11 538, 2020. 2

[38] 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," in *ICRA*. IEEE, 2023, pp. 9515–9522. 2

[39] Y. Hu, Q. Xie, V. Jain, J. Francis, J. Patrikar, N. Keetha, S. Kim, Y. Xie, T. Zhang, Z. Zhao *et al.*, "Toward general-purpose robots via foundation models: A survey and meta-analysis," *arXiv:2312.08782*, 2023. 2

[40] R. Firoozi, J. Tucker, S. Tian, A. Majumdar, J. Sun, W. Liu, Y. Zhu, S. Song, A. Kapoor, K. Hausman *et al.*, "Foundation models in robotics: Applications, challenges, and the future," *arXiv preprint arXiv:2312.07843*, 2023. 2

[41] K. Kawaharazuka, T. Matsushima, A. Gambardella, J. Guo, C. Paxton, and A. Zeng, "Real-world robot applications of foundation models: A review," *arXiv preprint arXiv:2402.05741*, 2024. 2

[42] S. Yang, O. Nachum, Y. Du, J. Wei, P. Abbeel, and D. Schuurmans, "Foundation models for decision making: Problems, methods, and opportunities," *arXiv preprint arXiv:2303.04129*, 2023. 2

[43] A. B. et al, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," 2023. 2

[44] Octo Model Team, D. Ghosh, H. Walke, K. Pertsch, K. Black, O. Mees, S. Dasari, J. Hejna, C. Xu, J. Luo, T. Kreiman, Y. Tan, L. Y. Chen, P. Sanketi, Q. Vuong, T. Xiao, D. Sadigh, C. Finn, and S. Levine, "Octo: An open-source generalist robot policy," in *RSS*, Delft, Netherlands, 2024. 2

[45] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi *et al.*, "Openvla: An open-source vision-language-action model," *arXiv preprint arXiv:2406.09246*, 2024. 2

[46] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter *et al.*, "π_0: A vision-language-action flow model for general robot control," *arXiv preprint arXiv:2410.24164*, 2024. 2

[47] H. Zhen, X. Qiu, P. Chen, J. Yang, X. Yan, Y. Du, Y. Hong, and C. Gan, "3d-vla: A 3d vision-language-action generative world model," *arXiv preprint arXiv:2403.09631*, 2024. 2

[48] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 000–16 009. 2

[49] J. Gao, B. Sarkar, F. Xia, T. Xiao, J. Wu, B. Ichter, A. Majumdar, and D. Sadigh, "Physically grounded vision-language models for robotic manipulation," in *ICRA*. IEEE, 2024, pp. 12 462–12 469. 2

[50] W. Yu, N. Gileadi, C. Fu, S. Kirmani, K.-H. Lee, M. G. Arenas, H.-T. L. Chiang, T. Erez, L. Hasenclever, J. Humplik *et al.*, "Language to rewards for robotic skill synthesis," *arXiv preprint arXiv:2306.08647*, 2023. 2

[51] Y. J. Ma, W. Liang, G. Wang, D.-A. Huang, O. Bastani, D. Jayaraman, Y. Zhu, L. Fan, and A. Anandkumar, "Eureka: Human-level reward design via coding large language models," *arXiv preprint arXiv:2310.12931*, 2023. 2

[52] W. Huang, C. Wang, Y. Li, R. Zhang, and L. Fei-Fei, "Rekep: Spatio-temporal reasoning of relational keypoint constraints for robotic manipulation," in *CoRL*, 2024. 2

[53] Y. Wang, T.-H. Wang, J. Mao, M. Hagenow, and J. Shah, "Grounding language plans in demonstrations through counterfactual perturbations," 2024. 2

[54] T. Xie, S. Zhao, C. H. Wu, Y. Liu, Q. Luo, V. Zhong, Y. Yang, and T. Yu, "Text2reward: Automated dense reward function generation for reinforcement learning," *arXiv preprint arXiv:2309.11489*, 2023. 2

[55] B. Liu, Y. Jiang, X. Zhang, Q. Liu, S. Zhang, J. Biswas, and P. Stone, "Llm+ p: Empowering large language models with optimal planning proficiency," *arXiv preprint arXiv:2304.11477*, 2023. 2

[56] J. Hsu, J. Mao, and J. Wu, "Ns3d: Neuro-symbolic grounding of 3d objects and relations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2614–2623. 2

[57] A. Athalye, N. Kumar, T. Silver, Y. Liang, T. Lozano-Pérez, and L. P. Kaelbling, "Predicate invention from pixels via pretrained vision-language models," *arXiv preprint arXiv:2501.00296*, 2024. 2

[58] Y. Du, M. Yang, P. Florence, F. Xia, A. Wahid, B. Ichter, P. Sermanet, T. Yu, P. Abbeel, J. B. Tenenbaum *et al.*, "Video language planning," *arXiv preprint arXiv:2310.10625*, 2023. 2

[59] S. Li, Y. Gao, D. Sadigh, and S. Song, "Unified video action model," *arXiv preprint arXiv:2503.00200*, 2025. 2

[60] B. Chen, D. Martí Monsó, Y. Du, M. Simchowitz, R. Tedrake, and V. Sitzmann, "Diffusion forcing: Next-token prediction meets full-sequence diffusion," *Advances in Neural Information Processing Systems*, vol. 37, pp. 24 081–24 125, 2024. 2

[61] Y. Du, S. Yang, B. Dai, H. Dai, O. Nachum, J. Tenenbaum, D. Schuurmans, and P. Abbeel, "Learning universal policies via text-guided video generation," *Advances in neural information processing systems*, vol. 36, pp. 9156–9172, 2023. 2

[62] J. Liang, W. Huang, F. Xia, P. Xu, K. Hausman, B. Ichter, P. Florence, and A. Zeng, "Code as policies: Language model programs for embodied control," *arXiv preprint arXiv:2209.07753*, 2022. 2

[63] J. Duan, W. Yuan, W. Pumacay, Y. R. Wang, K. Ehsani, D. Fox, and R. Krishna, "Manipulate-anything: Automating real-world robots using vision-language models," in *CoRL*, 2024. 2

[64] C. Wang, F. Xia, W. Yu, T. Zhang, R. Zhang, C. K. Liu, L. Fei-Fei, J. Tan, and J. Liang, "Chain-of-modality: Learning manipulation programs from multimodal human videos with vision-language-models," *arXiv preprint arXiv:2504.13351*, 2025. 2

[65] Z. Yang, C. Garrett, D. Fox, T. Lozano-Pérez, and L. P. Kaelbling, "Guiding long-horizon task and motion planning with vision language models," in *ICRA*, 2024. 2

[66] M. Minderer, A. Gritsenko, and N. Houlsby, "Scaling open-vocabulary object detection," in *NeurIPs*, 2023. 3

[67] L. Medeiros, "Lang-Segment-Anything: Language-guided instance segmentation," 2023. 3

[68] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollár, and C. Feichtenhofer, "Sam 2: Segment anything in images and videos," in *ICLR*, 2025. 3

[69] P. Besl and N. D. McKay, "A method for registration of 3-d shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 2, pp. 239–256, 1992. 4

[70] J. L. Bentley, "Multidimensional binary search trees used for associative searching," *Communications of the ACM*, vol. 18, no. 9, pp. 509–517, 1975. 4

[71] B. Sundaralingam, S. K. S. Hari, A. Fishman, C. Garrett, K. V. Wyk, V. Blukis, A. Millane, H. Oleynikova, A. Handa, F. Ramos, N. Ratliff, and D. Fox, "curobo: Parallelized collision-free minimum-jerk robot motion generation," 2023. 4

[72] P. Wu, Y. Shentu, Z. Yi, X. Lin, and P. Abbeel, "Gello: A general, low-cost, and intuitive teleoperation framework for robot manipulators," 2023. 4

[73] C. Lea, A. Reiter, R. Vidal, and G. Hager, "Segmental spatiotemporal cnns for fine-grained action segmentation," in *ECCV*, 2016. 5

[74] G. Ding, F. Sener, and A. Yao, "Temporal action segmentation: An analysis of modern techniques," 2023. 5

[75] Z. Zhang, Y. Li, O. Bastani, A. Gupta, D. Jayaraman, Y. J. Ma, and L. Weihs, "Universal visual decomposer: Long-horizon manipulation made easy," in *ICRA*. IEEE, 2024, pp. 6973–6980. 5

[76] L. P. Kaelbling and T. Lozano-Pérez, "Hierarchical task and motion planning in the now," in *ICRA*, 2011, pp. 1470–1477. 7