
Provable Ordering and Continuity in Vision-Language Pretraining for Generalizable Embodied Agents

Zhizhen Zhang

The University of Queensland
uqzzha39@uq.edu.au

Lei Zhu

Tongji University
leizhu0608@gmail.com

Zhen Fang

University of Technology Sydney
zhen.fang@uts.edu.au

Zi Huang

The University of Queensland
helen.huang@uq.edu.au

Yadan Luo

The University of Queensland
y.luo@uq.edu.au

Abstract

Pre-training vision-language representations on human action videos has emerged as a promising approach to reduce reliance on large-scale expert demonstrations for training embodied agents. However, prior methods often employ time contrastive learning based on goal-reaching heuristics, progressively aligning language instructions from the initial to the final frame. This overemphasis on future frames can result in *erroneous* vision-language associations, as actions may terminate early or include irrelevant moments in the end. To address this issue, we propose Action Temporal Coherence Learning (AcTOL) to learn *ordered* and *continuous* vision-language representations without rigid goal-based constraint. AcTOL treats a video as a continuous trajectory where it (1) contrasts semantic differences between frames to reflect their natural ordering, and (2) imposes a local Brownian bridge constraint to ensure smooth transitions across intermediate frames. Extensive imitation learning experiments on both simulated and real robots show that the pretrained features significantly enhance downstream manipulation tasks with high robustness to different linguistic styles of instructions, offering a viable pathway toward generalized embodied agents. Our project page is at <https://actol-pretrain.github.io/>.

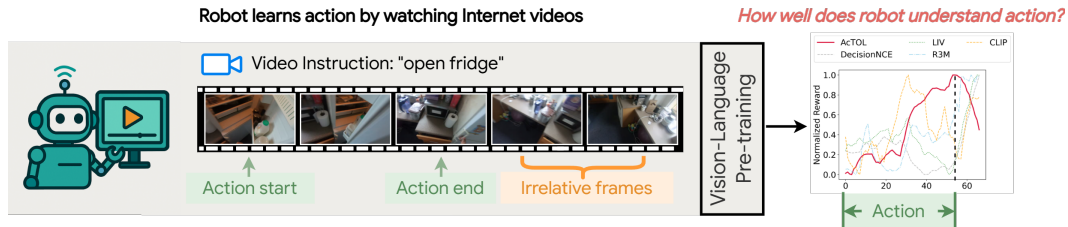


Figure 1: Pretraining on Internet human action videos for robot control, where the video-instruction pairs are noisy and often include irrelevant frames. The **red** vision-language reward curve demonstrates AcTOL learns to correctly align instruction with action, outperforming previous goal-reaching methods in the presence of distracting content.

1 Introduction

The long-term vision for embodied intelligence [29, 25] is to create systems that seamlessly perceive and interact with the world around them. Achieving this requires agents that integrate vision and language to understand their surroundings, interpret human instructions, and autonomously plan

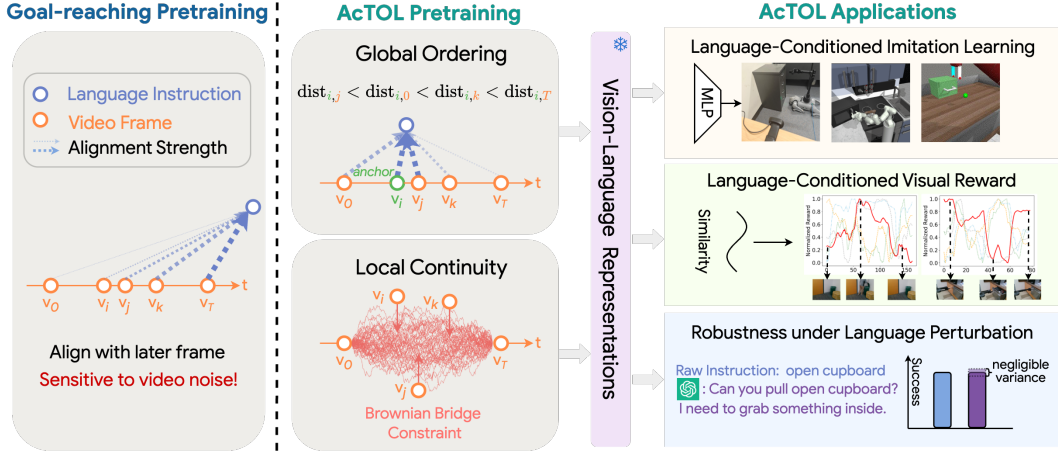


Figure 2: Comparison of existing *goal-reaching* pre-training strategies and the proposed AcTOL approach. Our learned multi-modal representations can be effectively transferred to downstream language-conditioned robot manipulation tasks, exhibiting robustness to diverse instruction and linguistic variations.

actions for complex tasks. Current end-to-end approaches achieve policy learning through direct vision-language-action mapping [48, 13, 7, 21, 3]. However, the inherent unpredictability of physical environments, including unseen scenarios and dynamic object interactions, constrains these solutions by requiring massive, high-quality robotic trajectories with action annotations, which are *costly* to collect. To mitigate this, recent research has leveraged large-scale, readily available egocentric human action videos [14, 10, 15] for *pre-training*. Although these out-of-domain videos often lack low-level action details and contain noise, their diverse human-object interactions and task instructions provide valuable prior knowledge. This enables the pre-trained representations to be more effectively transferred to novel tasks with fewer demonstrations, reducing reliance on large-scale robotic datasets while preserving strong generalization capabilities.

A promising approach for vision-language pre-training from human action videos leverages the concept of *time contrastive learning* [37] to capture temporally consistent visual representations, where language serves as the guiding goal, with semantic alignment between the language and chronologically later frames in the video [30, 26, 23]. However, this *goal-reaching* semantic alignment approach relies on a rigid assumption that action videos adhere to a specific principle: *actions progressively approach the target instruction from the initial frame to the final one*. Such assumption can be easily violated in egocentric human action videos, which are typically annotated at a coarse-grained level and riddled with noise. Figure 1 shows an example video-instruction pair, where the end of the video clip does not correspond to the actual end of the action. As a result, existing methods suffer from misleading semantic alignment, which hampers their ability to learn accurate vision-language relationships.

Given the challenges outlined above, a more natural and flexible pre-training strategy without rigid assumptions is needed to enhance vision-language representations for better policy learning. Building solely on the intrinsic temporal consistency of human action videos, we argue that the *ordering* and *continuity* of pre-trained vision-language representations play a crucial role in ensuring the effectiveness of policy learning. Ordering refers to the need for visual features to align with the underlying action logic required by the language instruction. For instance, as the task progresses, visual representations closer to the completion of the action should exhibit stronger alignment with the language instruction. This ensures that each step in the sequence is meaningfully associated with the corresponding instruction, enabling the model to effectively capture the dynamic progression of the task. Continuity, on the other hand, emphasizes that both visual features and their alignment with the language should evolve smoothly over time, with gradual transitions rather than abrupt changes. This is crucial because actions in the real world are not discrete but unfold continuously in time. Moreover, the alignment between visual and instruction should also be fluid, ensuring that as the action progresses, the visual representations consistently align with the target language instruction.

To address the aforementioned issues, as illustrated in Figure 2, we propose Action Temporal Coherence Learning (AcTOL), a novel approach designed to implicitly capture the ordering and continuity of video actions without relying on rigid assumptions, while providing strong theoretical

guarantees. Unlike previous approaches that focus on goal-directed semantic alignment, AcTOL introduces a Vision-Language Ordering (VLO) loss. This loss leverages the intrinsic temporal coherence of videos, contrasting frames against each other based on their relative temporal distance, theoretically ensuring that the semantic alignment between frames reflects their temporal ordering and continuity throughout the entire sequence. However, the VLO loss does not explicitly enforce the continuity of the visual features themselves, and under conditions with variations in frame content and noise, it can lead to suboptimal local consistency of the visual features. To address this, AcTOL introduces a Brownian bridge constraint over the video, treating video frames as a Brownian bridge process. This approach imposes a structured, continuous flow on the visual representations, ensuring that the model learns more consistent and stable intermediate states, further enhancing the continuity of the visual representations and improving the stability of their alignment with language instruction. Further theoretical analysis suggests that these properties also contribute to the model’s resilience to language perturbations, a crucial trait for real-world applications. To evaluate the generalization ability of AcTOL on embodied agents, we conducted extensive language-conditioned imitation learning experiments using both the real-world Unitree D1 robotic arm and two simulation environments. The results demonstrate that AcTOL significantly outperforms prior methods with a limited number of expert demonstrations. Additionally, AcTOL can generate language-conditioned visual rewards from real-world robot videos and remains robust to complex linguistic perturbations, highlighting its potential as a generalizable solution for real-world embodied agents.

2 Preliminaries

We first set up notations and mathematically formulate tasks.

Language-Conditioned Imitation Learning (LC-IL). The task of LC-IL aims to train an agent to mimic expert behaviors from a given robot demonstration set $\mathcal{D}_{\text{robot}} = \{(\tau_i, l_i)\}_{i=1}^{N_r}$, where $l_i \in \mathcal{L}$ represents a task-specific language instruction. Each trajectory $\tau_i \in \mathcal{T}$ consists of a sequence of state-action pairs $\tau_i = \{(\mathbf{s}_t, \mathbf{a}_t)\}_{t=1}^T$ of the horizon length T . In robot manipulation tasks, action $\mathbf{a}_t \in \mathcal{A}$ corresponds to the control commands executed by the agent and state $\mathbf{s}_t = [\mathbf{p}_t; \mathbf{o}_t] \in \mathcal{S}$ records proprioceptive data $\mathbf{p}_t$ (e.g., joint positions, velocities) and visual inputs $\mathbf{o}_t$ (e.g., camera images) at the time step t . The objective of LC-IL is to find an optimal language-conditioned policy $\pi^*(\mathbf{a}|\mathbf{s}, l) : \mathcal{S} \times \mathcal{L} \mapsto \mathcal{A}$ via solving the supervised optimization as follows,

$$\pi^* \in \arg \min_{\pi} \mathbb{E}_{(\tau_i, l_i) \sim \mathcal{T}} \left[\frac{1}{T} \sum_{(\mathbf{s}_t, \mathbf{a}_t) \sim \tau_i} \ell(\pi(\hat{\mathbf{a}}_t, \mathbf{s}_t | l_i), \mathbf{a}_t) \right],$$

where $\ell(\cdot, \cdot)$ is a task-specific loss, such as mean squared error or cross-entropy. Training the policy π_{θ} in an end-to-end fashion may require *hundreds* of high-quality expert demonstrations to converge, primarily due to the high variance of visual inputs $\mathbf{o}$ and language instructions l .

Vision-language Pre-training. Address such scalability issues can be achieved by leveraging large-scale, easily accessible human action video datasets [10, 15] $\mathcal{D}_{\text{human}} = \{(\mathcal{O}_i, l_i)\}_{i=1}^{N_h}$, where $\mathcal{O}_i = \{\mathbf{o}_t\}_{t=1}^T$ represents a video clip with T frames and l_i the corresponding description. Pretraining on such datasets enables policies to rapidly learn visual-language correspondences with minimal expert demonstrations. Mainstream pretraining methods employ *time contrastive learning* [37] to fine-tune a visual encoder ϕ and a text encoder φ , which project frames and descriptions into a shared d -dimensional embedding space, i.e., $\mathbf{v}_t = \phi(\mathbf{o}_t) \in \mathbb{R}^d$ and $\mathbf{l}_i = \varphi(l_i) \in \mathbb{R}^d$. To provide a unified perspective on various pretraining approaches, we formulate them within the objective $\mathcal{L}_{\text{tNCE}}(\phi, \varphi)$:

$$\mathcal{L}_{\text{tNCE}} = -\mathbb{E}_{\mathbf{o}^+ \sim \mathcal{P}(\mathcal{O}_i)} \log \frac{\exp(\mathfrak{R}(\mathbf{v}^+, \mathbf{l}_i))}{\mathbb{E}_{\mathbf{o}^- \sim \mathcal{N}(\mathcal{O}_i)} \exp(\mathfrak{R}(\mathbf{v}^-, \mathbf{l}_i))},$$

where $\mathbf{v}^{+/-} = \phi(\mathbf{o}^{+/-})$. Different pretraining strategies differ in their selection of (1) the positive frame set $\mathcal{P}(\mathcal{O}_i)$, (2) negative frame set $\mathcal{N}(\mathcal{O}_i)$; and (3) the semantic alignment scoring function $\mathfrak{R}(\mathbf{v}, \mathbf{l}_i)$ measuring the gap of VL similarities.

As motivated by goal-conditioned RL [1], current approaches *explicitly* select future frames (e.g., R3M [30], DecisionNCE [23]) or the last frame (e.g., LIV [26]) as the goal within the positive frame

set, enforcing their visual embedding to align with the semantics. Likewise, the scoring functions $\mathfrak{R}$ are often designed to maximize this transition direction. However, the pretraining action videos are *noisy* as actions may terminate early or include irrelevant subsequent actions, which may mislead the encoders and result in inaccurate vision-language association. As detecting precise action boundaries is non-trivial, we argue for a more flexible approach that leverages *intrinsic* characteristics of actions to guide pertaining.

3 Our Approach: AcTOL

We introduce an action temporal coherence learning (AcTOL) to capture two temporal properties of video actions: *ordering* and *continuity*. *Ordering* was ensured in the vision-language ordering loss (Section 3.1), where the semantic difference between frames reflects their temporal distance, with closer frames exhibiting smaller differences than those further apart. *Continuity* requires smooth visual transitions between adjacent frames, avoiding abrupt changes and high variance. To achieve this, we model sampled frame intervals as a Brownian bridge process (Section 3.2), penalizing deviations from the expected trajectories. Different from prior works that relies on setting explicit goal frames, the proposed approach *implicitly* explore the global and local structure of actions without imposing rigid constraints.

3.1 Visual-Language Ordering

To capture the temporal coherence of video actions, we first propose a vision-language ordering (VLO) loss that ensures the semantic alignment between frames reflects their temporal order. Since the VLO loss is applied within each video individually, we henceforth write $\mathcal{O}_i, l_i$ as $\mathcal{O}, l$ for simplicity. Consider an anchor frame $o_i \in \mathcal{O}$ with an index $n(i)$ corresponding to its position in the original video. For any given frame pair (o_i, o_j) , we first define the semantic alignment score $\mathfrak{R}$ to quantify differences in their VL similarities *w.r.t* a language description l as:

$$\mathfrak{R}(\mathbf{v}_i, \mathbf{v}_j, \mathbf{l}) = -\|\text{sim}(\mathbf{v}_i, \mathbf{l}) - \text{sim}(\mathbf{v}_j, \mathbf{l})\|_2, \quad (1)$$

where $\mathbf{v}_i = \phi(o_i), \mathbf{l} = \varphi(l)$. The function $\text{sim}(\cdot, \cdot)$ computes the VL similarity using cosine similarity. To ensure the proposed $\mathfrak{R}$ adhere to the temporal ordering of frames, we construct a negative set $\mathcal{N}_{i,j}$ by selecting $o_k \in \mathcal{O}$ correspond to frames that are temporally more *distant* from o_i than o_j :

$$\mathcal{N}_{i,j} = \{o_k \mid k \neq i, |n(i) - n(k)| \geq |n(i) - n(j)|\},$$

This formulation allows us to reformulate $\mathcal{L}_{\text{tNCE}}$ by enforcing that the VL similarity difference between frames i and j should be smaller than that between frame i and any negative frame k within the video $\mathcal{O}$:

$$\mathcal{L}_{\text{VLO}} = -\mathbb{E}_{(o_i, o_j) \sim \mathcal{O}} \log \frac{\exp(\mathfrak{R}(\mathbf{v}_i, \mathbf{v}_j, \mathbf{l}))}{\sum_{o_k \in \mathcal{N}_{i,j}} \exp(\mathfrak{R}(\mathbf{v}_i, \mathbf{v}_k, \mathbf{l}))}.$$

Notably, our VLO loss does not strictly require o_j to be from a future timestep for goal-reaching. Instead, we leverage the inherent temporal dynamics in videos, allowing the model to learn the natural ordering in an unsupervised manner.

3.2 Vision-Language Continuity

While the VLO property provides a strong global constraint on the structural alignment of VL pretraining, optimizing triplet relationships alone can be *unstable*. Variations in frame content and noise often lead to *suboptimal* local consistency. To mitigate this, we introduce an additional local continuity constraint inspired by the *Brownian bridge* [36]. This stochastic process models transitions between two fixed endpoints over by any sampled local video interval $[n(i), n(j)]$. For any time step $t \in [n(i), n(j)]$ within this interval, the transition density of Brownian Bridge process $\mathbf{B}(t)$ follows a time-dependent Gaussian distribution:

$$\mathcal{N}\left(\mathbf{v}_i + \frac{t - n(i)}{n(j) - n(i)}(\mathbf{v}_j - \mathbf{v}_i), \frac{t(n(j) - n(i)) - t^2}{n(j) - n(i)}\right),$$

where $\mathbf{v}_i, \mathbf{v}_j \in \mathbb{R}^d$ are the visual embeddings of the first and last frames in the sampled interval. The mean trajectory $\mathbb{E}[\mathbf{B}(t)]$ linearly interpolates between the two endpoints, while the variance

$\text{Var}[\mathbf{B}(t)]$ provides uncertainty modeling that peaks in the middle of the interval. To enforce this local continuity, the Brownian bridge loss $\mathcal{L}_{\text{BB}}$ is formulated as,

$$\mathcal{L}_{\text{BB}} = \frac{1}{T} \sum_{t=1}^T \frac{1}{2\text{Var}[\mathbf{B}(t)]} \|\mathbf{v}_t - \mathbb{E}[\mathbf{B}(t)]\|_2^2. \quad (2)$$

This loss encourages local consistency by penalizing deviations from expected trajectories, ensuring consistency across short temporal spans.

Overall Objective. The final training objective integrates both global and local constraints to achieve temporal coherence simultaneously:

$$\mathcal{L}_{\text{AcTOL}} = \mathcal{L}_{\text{VLO}} + \lambda \mathcal{L}_{\text{BB}}, \quad (3)$$

where λ is empirically set to balance two components.

4 Theoretical Analysis

In this section, we theoretically prove the vision-language ordering and continuity, as well as extend the robustness of linguistic perturbations of representations learned by AcTOL. All proofs are provided in Appendix C for reference.

Vision-Language Ordering. Ordering and sorting properties are well-established in self-supervised learning [39, 18, 45]. Building upon these insights, we formalize the concept of vision-language ordering (VLO) below.

Definition 1 (VLO Representations). Let $\{o_i\}_{i \in [T]}$ be a sequence of video frames and l the corresponding language description. The representations of the frames are said to satisfy the VLO property for any $0 < \delta < 1$ if $\forall i \in [T]$, and distinct frames $j, k \in [T] \setminus \{i\}$, the following conditions hold:

$$\begin{cases} \mathfrak{R}_{i,j,l} > \mathfrak{R}_{i,k,l} + 1/\delta, & \text{if } d_{i,j} < d_{i,k}, \\ |\mathfrak{R}_{i,j,l} - \mathfrak{R}_{i,k,l}| < \delta, & \text{if } d_{i,j} = d_{i,k}, \\ \mathfrak{R}_{i,j,l} < \mathfrak{R}_{i,k,l} - 1/\delta, & \text{if } d_{i,j} > d_{i,k}, \end{cases}$$

where $\mathfrak{R}_{i,j,l}$ denotes $\mathfrak{R}(\mathbf{v}_i, \mathbf{v}_j, \mathbf{l})$ and $d_{i,j}$ denotes $|n(i) - n(j)|$.

Implications of the VLO Property. The VLO property enforces a structured representation of video frames, ensuring that temporally adjacent frames have consistent and predictable semantic differences. When two frames have equal temporal distances from an anchor frame, their semantic gaps should be similar, fostering smooth transitions. In contrast, frames that are farther apart should exhibit larger semantic gaps, thus preserving the chronological order.

To formalize the temporal ordering constraints, we define the unique *sorted* set of frame distances from frame i as $\{D_{i,1} < D_{i,2} < \dots < D_{i,M_i}\}$, where each $D_{i,m}, m \in [M_i]$ is obtained by sorting the set $\{d_{i,j} \mid j \in [T] \setminus \{i\}\}$. Additionally, we define the count of frames at each distance level as:

$$n_{i,m} := |\{j \mid d_{i,j} = D_{i,m}, j \in [T] \setminus \{i\}\}|, \quad (4)$$

which denotes the number of frames whose temporal distance from frame i equals $D_{i,m}$. The VLO property is satisfied when the proposed $\mathcal{L}_{\text{VLO}}$ approaches its theoretical lower bound, which is given by:

$$\mathcal{L}^* := \frac{1}{T(T-1)} \sum_{i=1}^T \sum_{m=1}^{M_i} n_{i,m} \log n_{i,m}. \quad (5)$$

This bound characterizes the optimal alignment of VL similarities, ensuring that the learned representations preserve the inherent temporal structure within the video sequence, as guaranteed by the following theorem:

Theorem 1 (Vision-Language Ordering). $\mathcal{L}^*$ is a tight lower bound of $\mathcal{L}_{\text{VLO}}$, i.e., $\mathcal{L}_{\text{VLO}} \geq \mathcal{L}^*$, and for any $\epsilon > 0$, there exists feature embeddings such that $\mathcal{L}_{\text{VLO}} < \mathcal{L}^* + \epsilon$. Furthermore, for any $0 < \delta < 1$, there exist $\epsilon > 0$ such that if $\mathcal{L}_{\text{VLO}} < \mathcal{L}^* + \epsilon$, the learned representations satisfy the VLO property.

Vision-Language Continuity. We establish the following theoretical result to rigorously describe continuity preservation in vision-language representations:

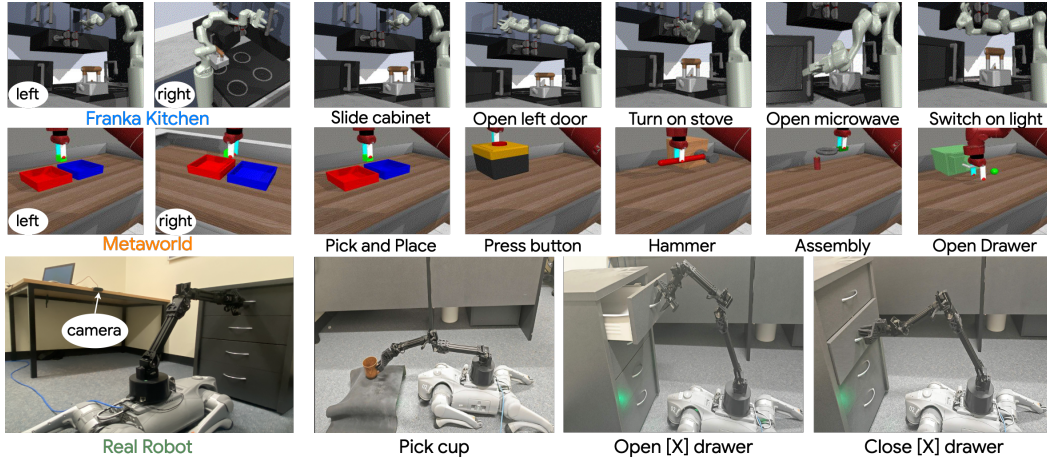


Figure 3: Policy learning environments, including 3 tasks with a real-world Unitree D1 robot arm and 5 tasks each in two simulation environments, *i.e.*, Franka Kitchen and Metaworld.

Theorem 2 (Vision-Language Continuity). *Let $\mathbf{v}_k, \mathbf{v}_l$ be visual representations at arbitrary time steps within a Brownian Bridge-regularized interval $[n(i), n(j)]$, and let $\mathbf{l} \in \mathcal{L}$ be a language embedding. If the VL similarity function $\text{sim}(\cdot)$ is Lipschitz continuous with constant C , then for any $\epsilon > 0$, there exists $\delta > 0$ such that:*

$$\|\mathbf{v}_k - \mathbf{v}_l\|_2 < \delta \quad \Rightarrow \quad |\Re(\mathbf{v}_k, \mathbf{v}_l, \mathbf{l})| < \epsilon.$$

This result follows from two key observations: (i) Brownian Bridge regularization constrains each embedding to remain close to a linear interpolation between anchor frames, with deviations governed by a time-dependent variance; and (ii) under this constraint, the distance between temporally close frames admits an explicit upper bound. Combining this with the Lipschitz continuity of the vision-language similarity function ensures that small changes in frame embeddings lead to proportionally bounded changes in alignment scores.

Building upon the continuity result, we further demonstrate that the semantic alignment score remains stable under small perturbations in language input:

Theorem 3 (Robustness to Language Variations). *Let $\mathbf{l}'$ be a perturbed language embedding such that $\|\mathbf{l} - \mathbf{l}'\| \leq \delta_l$. Then the semantic alignment score $\Re$ satisfies:*

$$|\Re(\mathbf{v}_i, \mathbf{v}_j, \mathbf{l}') - \Re(\mathbf{v}_i, \mathbf{v}_j, \mathbf{l})| \leq 2C\delta_l.$$

This second result guarantees that small shifts in the language representation (*e.g.*, synonym substitution or phrasing variation) lead to bounded changes in the alignment score. Together, Theorems 2 and 3 formalize the local stability of semantic grounding across both time and modality, providing a theoretical basis for continuity-aware vision-language learning.

5 Experiment

In our experiments, we aim to evaluate the effectiveness of ordered and continuous vision-language representations for robotic control. First, we conduct extensive Language-Conditioned Behavior Cloning (LCBC) experiments on both real and simulated robots to validate the importance of ordering and continuity for imitation learning. Second, we assess the utility of the learned representations as reward functions on multiple real-world action videos. The results demonstrate that the ordered and continuous representations enable our method to accurately identify action boundaries and generate dense rewards aligned with the given instructions. Finally, we evaluate the robustness of our method under language perturbations, showcasing its strong generalization capability for application in real-world daily scenarios.

Experimental Setup. Figure 3 shows the experimental environments. For real-world robot evaluation, we deploy the **Unitree D1 robot arm** to perform three challenging manipulation tasks: pick cup, open [X] drawer and close [X] drawer, where [X] is the drawer index specified by the instruction. The pick cup task requires the model to accurately identify the cup handle, while the open/close [X] drawer tasks demand grounding of language instructions to visual observations, enabling the

Table 1: Comparison in simulation environments with varying amounts of demonstrations. Each result reports the success rate over 50 roll-outs, averaged across 2 camera views and 3 random seeds. We also report the relative performance gain in **green** compared to the *strongest* baseline.

Method	FRANKA KITCHEN			METAWORLD		
	5 demos	15 demos	25 demos	5 demos	15 demos	25 demos
CLIP	11.67 $\pm$ 0.95	27.47 $\pm$ 1.01	31.20 $\pm$ 2.62	42.29 $\pm$ 2.65	60.33 $\pm$ 1.32	62.54 $\pm$ 4.36
R3M	28.60 $\pm$ 1.39	42.20 $\pm$ 1.00	51.13 $\pm$ 2.83	46.83 $\pm$ 3.85	56.50 $\pm$ 5.20	60.08 $\pm$ 3.62
LIV	23.40 $\pm$ 0.78	42.73 $\pm$ 1.17	51.93 $\pm$ 0.95	46.95 $\pm$ 2.07	64.33 $\pm$ 3.63	66.67 $\pm$ 1.49
DecisionNCE	25.33 $\pm$ 1.30	43.20 $\pm$ 2.25	50.87 $\pm$ 2.95	44.58 $\pm$ 2.79	59.08 $\pm$ 1.77	69.75 $\pm$ 3.90
AcTOL w/o BB	32.80 $\pm$ 1.23	54.20 $\pm$ 0.85	60.80 $\pm$ 0.87	50.29 $\pm$ 4.05	70.83 $\pm$ 4.21	73.33 $\pm$ 2.83
AcTOL	42.60 $\pm$ 0.53 (+48.95%)	61.80 $\pm$ 2.54 (+43.06%)	64.60 $\pm$ 0.57 (+24.40%)	53.81 $\pm$ 3.89 (+14.61%)	74.13 $\pm$ 1.59 (+15.23%)	81.13 $\pm$ 1.59 (+16.32%)

model to interact with the correct drawer. To isolate manipulation performance, the Unitree Go2 quadruped remains lying down and stationary throughout the evaluation. We use a web camera to capture a third-person view as visual observation. The action space consists of a 6-DoF end-effector displacement vector and gripper state, executed at a control frequency of 20 Hz. For each task, we collect 60 demonstrations via remote control using the Unitree Go app, which is significantly fewer than the 100 trajectories typically used in prior work [26, 23]. For simulation, we choose two widely used simulation environments for evaluation: **Franka Kitchen** [16, 12] and **Metaworld** [43]. For Franka Kitchen, we evaluate five tasks: sliding a cabinet, opening the left door, opening the microwave, turning on the stove, and switching on the light. For Metaworld, we focus on learning five tasks: hammering a nail, pressing a button, picking and placing a block, assembling a ring onto a peg, and opening a drawer. Detailed environment setup can be found at Appendix B.1.

Baselines. Since our model is initialized with **CLIP** [34], a state-of-the-art image-text representation widely applied in various embodied tasks [9, 20, 38, 41], it is a natural choice to include CLIP as a vanilla baseline for comparison. Our primary baselines are **LIV** [26] and **DecisionNCE** [23], as we use the same model architecture and dataset for pre-training. We also compare against **R3M** [30] pre-trained on Ego4D [15], a dataset containing roughly $36\times$ longer videos than EPIC-KITCHEN-100. We also include an ablation variant of AcTOL where the Brownian Bridge loss is removed, referred to as AcTOL w/o BB.

Implementation Details. We initialize our model with the weights of CLIP [34] with ResNet-50 vision backbone and further pre-train it on human action video dataset EPIC-KITCHEN-100 [10, 11]. For hyperparameter selection, we uniformly sample 10 frames of each video per batch. The loss weight λ is 0.1. Other hyperparameters, such as temperature, follow the default value used in CLIP [34]. More details of pre-training and hyperparameter sensitivity can be found in Appendix A.

5.1 Language-Conditioned Behavior Cloning

For LCBC policy learning, we keep the pre-trained vision-language encoders frozen and feed their output representations into a lightweight MLP, which is trained as a policy network.

Simulation results. In simulation, each task is performed from two camera viewpoints (left and right), with varying numbers of demonstrations [5, 15, 25] (*i.e.*, dataset size) for training, and evaluated under three different random seeds. We report the success rate across different environments and dataset sizes, averaged over camera views and seeds. Detailed comparison results for each task can be referred to Appendix B.5. Table 1 presents the comparison results, demonstrating that AcTOL achieves significantly enhanced performance relative to baseline methods across all evaluated datasets and environments. This superiority is particularly pronounced in the complex Franka Kitchen setting, especially under data constraints, where AcTOL with fewer demonstrations (*e.g.*, 5/15) often matches or surpasses other methods using more data (*e.g.*, 15/25), indicating its high data efficiency and robust low-resource generalization capabilities. Furthermore, ablation studies confirm the integral role of the Brownian Bridge (BB) constraint; its removal (AcTOL w/o BB) results in a significant performance decrease, validating its contribution to improving representation quality for effective policy optimization via behavior cloning.

Robustness under visual shifts. To further assess the ability of the model to handle visual distribution shifts, we conduct experiments in the Franka Kitchen environment following the protocol of [5]. Specifically, we compare Ac-

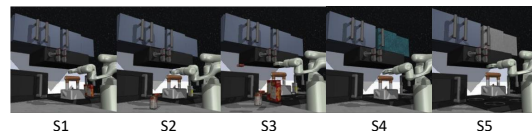


Figure 4: Visual shifts applied in Franka Kitchen.

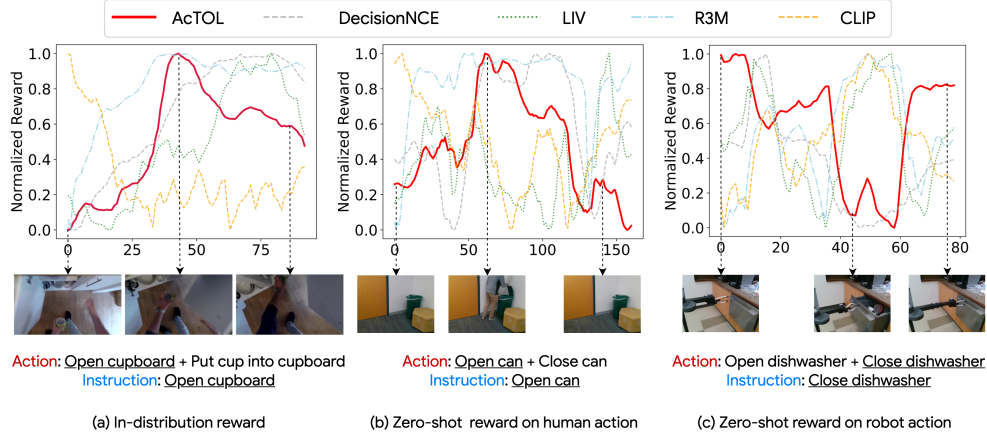


Figure 5: Visualization of the normalized learned reward corresponding to different actions. Our representations effectively help capture the correct temporal order of actions in the instruction. For more results, please refer to Appendix B.6.

TOL with the strongest baseline, DecisionNCE, under visual changes absent from training. These include: (1) object distractors of increasing difficulty: easy (S1), medium (S2), and hard (S3), corresponding to scenes with 1, 3, and 9 additional YCB objects [6], respectively; and (2) background texture variations with marble hinge (S4) and metal slide (S5). All shifts are shown in Figure 4.

Policies are trained with 15 demonstrations per task, and success rates averaged over five tasks are reported in Table 2. While performance drops under visual shifts, which is expected, AcTOL continues to outperform DecisionNCE in all available test conditions. This suggests that the learned representation maintains useful generalization ability even without any specific adaptation for visual domain shift.

Real Robot results. Table 3 shows the real robot comparison results. AcTOL consistently outperforms all baseline models across the three tasks. Among them, the pick cup task yields relatively lower performance, as it requires the model to precisely identify and grasp the cup handle, demanding stronger spatial perception capabilities. For the open/close [X] drawer tasks, AcTOL is able to accurately interpret the drawer number specified in the language instruction, align it with the corresponding location in the visual observation, and execute continuous actions on the correct drawer to complete the task. These results highlight the effectiveness of AcTOL’s learned visual-language representations in real-world manipulation tasks.

Table 2: Success rate comparison across different visual shifts in Franka Kitchen.

Method	No shift	S1	S2	S3	S4	S5
DecisionNCE	43.2	27.2	25.6	4.8	0.0	8.8
AcTOL	61.8	43.2	32.8	9.2	4.4	38.4

Table 3: Performance comparison on Unitree D1 arm. Success rates are reported over 10 trials.

Method	Pick Cup	Open [X] Drawer	Close [X] Drawer
CLIP	0	20	30
R3M	10	40	40
LIV	20	30	50
DecisionNCE	20	40	60
AcTOL	50	80	90

5.2 Language-Conditioned Visual Rewards

By learning semantically smooth visual representations, our model further enables the use of semantic trajectories as effective task rewards. To illustrate this, we first demonstrate the continuity of purely visual representations. In Figure 6, we visualize the learned visual representation trajectories for three tasks, each with ten video clips, using t-SNE. The results show that AcTOL significantly improves the temporal continuity of video feature trajectories while retaining CLIP’s discriminative ability to distinguish

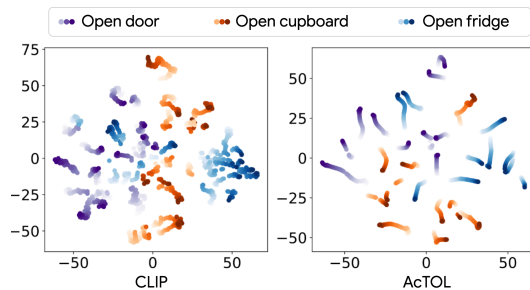


Figure 6: Visual trajectory visualization.

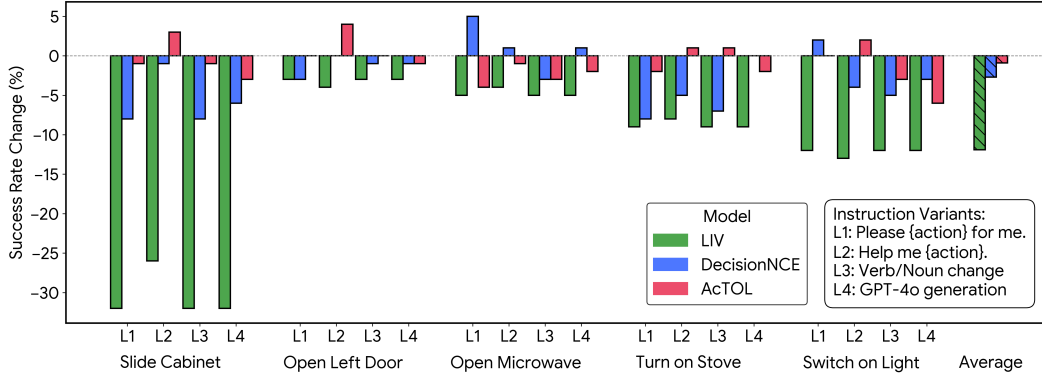


Figure 7: Success rate fluctuation across tasks in Franka Kitchen for different instruction variants.

between actions associated with different instructions. As discussed in Section 3.2, the visual continuity can stabilize learning ordered vision-language alignment. Building on this foundation, we define a dense reward signal based on the semantic alignment between the current visual state and the language goal. Specifically, at each time step i , we define the reward $\cos(\mathbf{v}^i, \mathbf{l})$ as the similarity between the current visual state and the language goal. While prior work [26, 23] focused primarily on single-action video clips, we evaluate reward quality on three clips, each containing two consecutive actions, to assess whether the model can reliably capture fine-grained action semantics. Figure 5 (a) presents an in-distribution evaluation using a video from EPIC-KITCHEN-100. Our model produces a clear reward peak aligned with the completion of the “open cupboard” action, followed by a decline, indicating successful temporal localization of the instructed behavior. In contrast, R3M and DecisionNCE rewards continue increasing beyond the relevant action segment. Figures 5 (b) and (c) show results on real-world videos from [2], where human and robot actors perform opposite actions. Only our method consistently produces symmetric and instruction-aligned reward curves, accurately identifying both action boundaries and semantics.

5.3 Robustness Study under Linguistic Perturbations

In the EPIC-KITCHEN-100 dataset, textual annotations are often concise, such as “open cupboard”. In the default setting of LCBC, we employ similarly structured simple instructions. In this experiment, to validate the robustness of the representations our method learns in real-world scenarios, we introduce several modifications to the language instructions. Specifically, we transform each original instruction into four conversational variants by varying lexical choices (*e.g.*, verbs and nouns) and incorporating ChatGPT-4o [32] generated complex instructions. Details can be found in Appendix B.4. We then evaluate the imitation learning performance conditioned on these modified instructions in the Franka Kitchen environment. For comparison, we select LIV and DecisionNCE, which are also pre-trained on EPIC-KITCHEN-100. As shown in Figure 7, the success rates of LIV and DecisionNCE dropped by 11.9% and 2.7% on average, respectively, while our method maintained a success rate comparable to that before language perturbation. This result demonstrates the robustness of our learned representations, which generalize more effectively to real-world scenarios.

5.4 Mitigating the Human-to-Robot Gap via Fine-Tuning

Although pretraining on human videos provides generalizable knowledge, bridging the human-to-robot domain gap remains a persistent challenge [47, 22, 31]. Since the AcTOL objectives capture the inherent temporal ordering of videos in an embodiment-agnostic manner, we can fine-tune the vision-language encoders with the same objectives used in pretraining during downstream behavior cloning. Notably, we find that fine-tuning with only a small number of robot demonstrations is sufficient to substantially mitigate the domain gap. We take 25 in-domain demonstrations (5 per task) in Franka Kitchen to fine-tune the pre-trained encoders using the AcTOL objectives. Then, as in Sec 5.1, we freeze the fine-tuned encoders and train policy networks on top using behavior cloning. We report the comparisons when using 15 demos for LCBC. As shown in Table 4, the success rate improvement demonstrates that the learned temporal inductive bias can be effectively adapted to the robot domain with limited supervision.

Table 4: Fine-tuning AcTOL encoders efficiently improves the success rate in Franka Kitchen.

Franka Kitchen	Frozen	Finetune
AcTOL	61.8	86.4

6 Related Work

Given the success of large-scale pre-training in the vision and language research communities [4, 24], many studies have attempted to extend this paradigm to the field of robotics. Some work leverage massive robotic trajectory data [8] for pre-training, aiming to establish unified vision-language-action models [48, 7, 21, 3, 13, 40, 33]. However, collecting large amounts of high-quality robot trajectory data is extremely costly and time-consuming. Consequently, many studies have begun to explore the use of large-scale, readily available, out-of-domain human action video data to learn generalizable representations that can be transferred to robotic tasks [37, 27, 35, 30, 19, 26, 28, 42, 44, 23]. Among these, TCN [37], VIP [27], MVP [35], and VC-1 [28] focus solely on studying unimodal visual representations, limiting their performance when understanding language instructions is required. R3M [30] employs language and reward models to shape progressive visual representations, while Voltron [19] and MPI [44] model the transition from the current state to the goal state conditioned on language. However, during training, these approaches freeze the language encoder, using it only to aid in the training of visual representations. As a result, they do not effectively achieve multi-modal representation learning. LIV [26] and DecisionNCE [23] have attempted to leverage CLIP [34] to train embodied multi-modal representations. LIV treats language as the goal of video actions, aligning it with the final frame, while DecisionNCE aligns language with the transition from the initial to final frame. Both rely on a goal-reaching assumption, which can lead to suboptimal results in noisy real-world videos. In contrast, our approach avoids rigid assumptions by enforcing semantic alignment that follows the intrinsic temporal continuity of videos, leading to more robust and generalizable vision-language representations. This property also benefits methods like UVD [46], which rely on pretrained visual features to detect phase changes and decompose long-horizon tasks. Our method more reliably identifies action phases, enabling stronger progress rewards and improving suitability for such goal-conditioned downstream tasks.

7 Conclusion and Limitations

We present Action Temporal Coherence Learning (AcTOL) as a promising vision-language pre-training solution for generalizable embodied agents. By learning action consistency from a large corpus of human action videos, AcTOL theoretically ensures the ordering and continuity of vision-language representations, as well as robustness to language perturbations. Extensive experiments across various environments demonstrate that AcTOL effectively generalizes to complex robotic manipulation tasks. While the temporal ordering of actions provides a strong inductive bias for many goal-directed tasks, it may not align well with tasks that involve ambiguous, repetitive, or cyclic behaviors. In such cases, the assumption of coherent progression might break down, potentially affecting the reliability of the model. Future work could explore adapting AcTOL to handle such repetitive action sequences.

Acknowledgments and Disclosure of Funding

This work was partially supported by ARC DE240100105, DP240101814, DP230101196, DP230101753, BA24006, DE250100363, and ARC Industrial Transformation Research Hubs IH230100013.

References

- [1] Marcin Andrychowicz, Dwight Crow, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, Pieter Abbeel, and Wojciech Zaremba. Hindsight experience replay. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems, (NeurIPS)*, pages 5048–5058, 2017.
- [2] Shikhar Bahl, Abhinav Gupta, and Deepak Pathak. Human-to-robot imitation in the wild. In Kris Hauser, Dylan A. Shell, and Shoudong Huang, editors, *Robotics: Science and Systems XVIII, New York City, NY, USA, June 27 - July 1, 2022*, 2022.

- [3] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0 : A vision-language-action flow model for general robot control. *CoRR*, abs/2410.24164, 2024.
- [4] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In Hugo Larochelle, Marc’ Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems, (NeurIPS)*, 2020.
- [5] Kaylee Burns, Zach Witzel, Jubayer Ibn Hamid, Tianhe Yu, Chelsea Finn, and Karol Hausman. What makes pre-trained visual representations successful for robust manipulation?, 2023.
- [6] Berk Calli, Arjun Singh, Aaron Walsman, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M. Dollar. The ycb object and model set: Towards common benchmarks for manipulation research. In *2015 International Conference on Advanced Robotics (ICAR)*, pages 510–517, 2015.
- [7] Chilam Cheang, Guangzeng Chen, Ya Jing, Tao Kong, Hang Li, Yifeng Li, Yuxiao Liu, Hongtao Wu, Jiafeng Xu, Yichu Yang, Hanbo Zhang, and Minzhao Zhu. GR-2: A generative video-language-action model with web-scale knowledge for robot manipulation. *CoRR*, abs/2410.06158, 2024.
- [8] Open X.-Embodiment Collaboration, Abhishek Padalkar, Acorn Pooley, Ajinkya Jain, Alex Bewley, Alexander Herzog, Alex Irpan, Alexander Khazatsky, Anant Raj, Anikait Singh, Anthony Brohan, Antonin Raffin, Ayzaan Wahid, Ben Burgess-Limerick, Beomjoon Kim, Bernhard Schölkopf, Brian Ichter, Cewu Lu, Charles Xu, Chelsea Finn, Chenfeng Xu, Cheng Chi, Chenguang Huang, Christine Chan, Chuer Pan, Chuyuan Fu, Coline Devin, Danny Driess, Deepak Pathak, Dhruv Shah, Dieter Büchler, Dmitry Kalashnikov, Dorsa Sadigh, Edward Johns, Federico Ceola, Fei Xia, Freek Stulp, Gaoyue Zhou, Gaurav S. Sukhatme, Gautam Salhotra, Ge Yan, Giulio Schiavi, Gregory Kahn, Hao Su, Haoshu Fang, Haochen Shi, Heni Ben Amor, Henrik I. Christensen, Hiroki Furuta, Homer Walke, Hongjie Fang, Igor Mordatch, Ilija Radosavovic, and et al. Open x-embodiment: Robotic learning datasets and RT-X models. *CoRR*, abs/2310.08864, 2023.
- [9] Yuchen Cui, Scott Niekum, Abhinav Gupta, Vikash Kumar, and Aravind Rajeswaran. Can foundation models perform zero-shot task specification for robot manipulation? In Roya Firooz, Negar Mehr, Esen Yel, Rika Antonova, Jeannette Bohg, Mac Schwager, and Mykel J. Kochenderfer, editors, *Learning for Dynamics and Control Conference, (L4DC)*, volume 168 of *Proceedings of Machine Learning Research*, pages 893–905. PMLR, 2022.
- [10] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The EPIC-KITCHENS dataset. *CoRR*, abs/1804.02748, 2018.
- [11] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision. *CoRR*, abs/2006.13256, 2020.
- [12] Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4RL: datasets for deep data-driven reinforcement learning. *CoRR*, abs/2004.07219, 2020.
- [13] Dibya Ghosh, Homer Rich Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, Jianlan Luo, You Liang Tan, Lawrence Yunliang Chen, Quan Vuong, Ted Xiao, Pannag R. Sanketi, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An open-source generalist robot policy. In Dana Kulic, Gentiane Venture, Kostas E. Bekris, and Enrique Coronado, editors, *Robotics: Science and Systems XX, Delft, The Netherlands, July 15-19, 2024*, 2024.

- [14] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fründ, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thureau, Ingo Bax, and Roland Memisevic. The "something something" video database for learning and evaluating visual common sense. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 5843–5851. IEEE Computer Society, 2017.
- [15] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abraham Gebreselasie, Cristina González, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jáchym Kolár, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbeláez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard A. Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the world in 3, 000 hours of egocentric video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, (CVPR)*, pages 18973–18990. IEEE, 2022.
- [16] Abhishek Gupta, Vikash Kumar, Corey Lynch, Sergey Levine, and Karol Hausman. Relay policy learning: Solving long-horizon tasks via imitation and reinforcement learning. In Leslie Pack Kaelbling, Danica Kragic, and Komei Sugiura, editors, *3rd Annual Conference on Robot Learning, (CoRL)*, volume 100 of *Proceedings of Machine Learning Research*, pages 1025–1037. PMLR, 2019.
- [17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, (CVPR)*, pages 770–778. IEEE Computer Society, 2016.
- [18] Kai Hu, Jie Shao, Yuan Liu, Bhiksha Raj, Marios Savvides, and Zhiqiang Shen. Contrast and order representations for video self-supervised learning. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 7919–7929. IEEE, 2021.
- [19] Siddharth Karamcheti, Suraj Nair, Annie S. Chen, Thomas Kollar, Chelsea Finn, Dorsa Sadigh, and Percy Liang. Language-driven representation learning for robotics. In Kostas E. Bekris, Kris Hauser, Sylvia L. Herbert, and Jingjin Yu, editors, *Robotics: Science and Systems XIX, (RSS)*, 2023.
- [20] Apoorv Khandelwal, Luca Weihs, Roozbeh Mottaghi, and Aniruddha Kembhavi. Simple but effective: CLIP embeddings for embodied AI. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, (CVPR)*, pages 14809–14818. IEEE, 2022.
- [21] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Paul Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. Openvla: An open-source vision-language-action model. *CoRR*, abs/2406.09246, 2024.
- [22] Guangrun Li, Yaoyu Lyu, Zhuoyang Liu, Chengkai Hou, Jieyu Zhang, and Shanghang Zhang. H2r: A human-to-robot data augmentation for robot pre-training from videos, 2025.
- [23] Jianxiong Li, Jinliang Zheng, Yinan Zheng, Liyuan Mao, Xiao Hu, Sijie Cheng, Haoyi Niu, Jihao Liu, Yu Liu, Jingjing Liu, Ya-Qin Zhang, and Xianyu Zhan. Decisionncc: Embodied multimodal representations via implicit preference learning. In *Forty-first International Conference on Machine Learning, (ICML)*. OpenReview.net, 2024.

- [24] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems, (NeurIPS)*, 2023.
- [25] Yang Liu, Weixing Chen, Yongjie Bai, Guanbin Li, Wen Gao, and Liang Lin. Aligning cyber space with physical world: A comprehensive survey on embodied AI. *CoRR*, abs/2407.06886, 2024.
- [26] Yecheng Jason Ma, Vikash Kumar, Amy Zhang, Osbert Bastani, and Dinesh Jayaraman. LIV: language-image representations and rewards for robotic control. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pages 23301–23320. PMLR, 2023.
- [27] Yecheng Jason Ma, Shagun Sodhani, Dinesh Jayaraman, Osbert Bastani, Vikash Kumar, and Amy Zhang. VIP: towards universal visual reward and representation via value-implicit pre-training. In *The Eleventh International Conference on Learning Representations, (ICLR)*. OpenReview.net, 2023.
- [28] Arjun Majumdar, Karmesh Yadav, Sergio Arnaud, Yecheng Jason Ma, Claire Chen, Sneha Silwal, Aryan Jain, Vincent-Pierre Berges, Tingfan Wu, Jay Vakil, Pieter Abbeel, Jitendra Malik, Dhruv Batra, Yixin Lin, Oleksandr Maksymets, Aravind Rajeswaran, and Franziska Meier. Where are we in the search for an artificial visual cortex for embodied intelligence? In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, (NeurIPS)*, 2023.
- [29] Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhai Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. Embodiedgpt: Vision-language pre-training via embodied chain of thought. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.
- [30] Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3M: A universal visual representation for robot manipulation. In Karen Liu, Dana Kulic, and Jeffrey Ichnowski, editors, *Conference on Robot Learning, (CoRL)*, volume 205 of *Proceedings of Machine Learning Research*, pages 892–909. PMLR, 2022.
- [31] Nghia Nguyen, Minh Nhat Vu, Tung D Ta, Baoru Huang, Thieu Vo, Ngan Le, and Anh Nguyen. Robotic-clip: Fine-tuning clip on action data for robotic applications. In *ICRA*, 2025.
- [32] OpenAI. Gpt-4o technical report, 2024.
- [33] Delin Qu, Haoming Song, Qizhi Chen, Yuanqi Yao, Xinyi Ye, Yan Ding, Zhigang Wang, JiaYuan Gu, Bin Zhao, Dong Wang, and Xuelong Li. Spatialvla: Exploring spatial representations for visual-language-action model. *CoRR*, abs/2501.15830, 2025.
- [34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021.
- [35] Ilija Radosavovic, Tete Xiao, Stephen James, Pieter Abbeel, Jitendra Malik, and Trevor Darrell. Real-world robot learning with masked visual pre-training. In Karen Liu, Dana Kulic, and Jeffrey Ichnowski, editors, *Conference on Robot Learning, (CoRL)*, volume 205 of *Proceedings of Machine Learning Research*, pages 416–426. PMLR, 2022.
- [36] Daniel Revuz and Marc Yor. *Continuous martingales and Brownian motion*, volume 293. Springer Science & Business Media, 2013.

- [37] Pierre Sermanet, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, and Sergey Levine. Time-contrastive networks: Self-supervised learning from video. In *2018 IEEE International Conference on Robotics and Automation, (ICRA)*, pages 1134–1141. IEEE, 2018.
- [38] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Cliport: What and where pathways for robotic manipulation. In Aleksandra Faust, David Hsu, and Gerhard Neumann, editors, *Conference on Robot Learning, (CoRL)*, volume 164 of *Proceedings of Machine Learning Research*, pages 894–906. PMLR, 2021.
- [39] Nina Shvetsova, Felix Petersen, Anna Kukleva, Bernt Schiele, and Hilde Kuehne. Learning by sorting: Self-supervised learning with group ordering constraints. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 16407–16417. IEEE, 2023.
- [40] Andrew Szot, Bogdan Mazouze, Harsh Agrawal, R. Devon Hjelm, Zsolt Kira, and Alexander Toshev. Grounding multimodal large language models in actions. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024.
- [41] Allison C. Tam, Neil C. Rabinowitz, Andrew K. Lampinen, Nicholas A. Roy, Stephanie C. Y. Chan, DJ Strouse, Jane Wang, Andrea Banino, and Felix Hill. Semantic exploration from language abstractions and pretrained representations. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, (NeurIPS)*, 2022.
- [42] Seonghyeon Ye, Joel Jang, Byeongguk Jeon, Se June Joo, Jianwei Yang, Baolin Peng, Ajay Mandlekar, Reuben Tan, Yu-Wei Chao, Bill Yuchen Lin, Lars Liden, Kimin Lee, Jianfeng Gao, Luke Zettlemoyer, Dieter Fox, and Minjoon Seo. Latent action pretraining from videos. *CoRR*, abs/2410.11758, 2024.
- [43] Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In Leslie Pack Kaelbling, Danica Kragic, and Komei Sugiura, editors, *3rd Annual Conference on Robot Learning, (CoRL)*, volume 100 of *Proceedings of Machine Learning Research*, pages 1094–1100. PMLR, 2019.
- [44] Jia Zeng, Qingwen Bu, Bangjun Wang, Wenke Xia, Li Chen, Hao Dong, Haoming Song, Dong Wang, Di Hu, Ping Luo, Heming Cui, Bin Zhao, Xuelong Li, Yu Qiao, and Hongyang Li. Learning manipulation by predicting interaction. *CoRR*, abs/2406.00439, 2024.
- [45] Kaiwen Zha, Peng Cao, Jeany Son, Yuzhe Yang, and Dina Katabi. Rank-n-contrast: Learning continuous representations for regression. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, (NeurIPS)*, LA, USA, December 10 - 16, 2023, 2023.
- [46] Zichen Zhang, Yunshuang Li, Osbert Bastani, Abhishek Gupta, Dinesh Jayaraman, Yecheng Jason Ma, and Luca Weihs. Universal visual decomposer: Long-horizon manipulation made easy. In *IEEE International Conference on Robotics and Automation, ICRA 2024, Yokohama, Japan, May 13-17, 2024*, pages 6973–6980. IEEE, 2024.
- [47] Jiaming Zhou, Teli Ma, Kun-Yu Lin, Ronghe Qiu, Zifan Wang, and Junwei Liang. Mitigating the human-robot domain discrepancy in visual pre-training for robotic manipulation. *arXiv preprint arXiv:2406.14235*, 2024.
- [48] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, Huong T. Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag R. Sanketi, Grecia Salazar, Michael S. Ryoo, Krista Reymann, Kanishka Rao, Karl Pertsch, Igor Mordatch, Henryk Michalewski, Yao

Lu, Sergey Levine, Lisa Lee, Tsang-Wei Edward Lee, Isabel Leal, Yuheng Kuang, Dmitry Kalashnikov, Ryan Julian, Nikhil J. Joshi, Alex Irpan, Brian Ichter, Jasmine Hsu, Alexander Herzog, Karol Hausman, Keerthana Gopalakrishnan, Chuyuan Fu, Pete Florence, Chelsea Finn, Kumar Avinava Dubey, Danny Driess, Tianli Ding, Krzysztof Marcin Choromanski, Xi Chen, Yevgen Chebotar, Justice Carbajal, Noah Brown, Anthony Brohan, Montserrat Gonzalez Arenas, and Kehang Han. RT-2: vision-language-action models transfer web knowledge to robotic control. In Jie Tan, Marc Toussaint, and Kourosh Darvish, editors, *Conference on Robot Learning, (CoRL)*, volume 229 of *Proceedings of Machine Learning Research*, pages 2165–2183. PMLR, 2023.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: The abstract and introduction consistently articulate the problem, the proposed AcTOL solution, its key contributions regarding ordered and continuous representation learning, and its demonstrated benefits and scope in vision-language pre-training for embodied agents.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: Limitations have been discussed in Section 7 and Appendix D.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification: Full set of assumptions and a complete (and correct) proof is provided in Appendix C.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: Implementation details have been provided in Appendix B.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The dataset used for pre-training is EPIC-KITCHENS-100 which is publicly available, and the code is included in supplemental materials.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Experimental setting/details are discussed in Section 5 and Appendix B.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Errors are provided for all language-conditioned behavior cloning experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Experiments compute resources are discussed in Appendix B.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: Authors have read NeurIPS Code of Ethics and confirm to preserve anonymity.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Broader impacts are discussed in Appendix D.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: This paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The EPIC-KITCHENS-100 dataset used in this paper has been explicitly mentioned and properly respected.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset’s creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: This paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: This paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

Appendix

A Pre-training Details

Following [26, 23], we use a modified ResNet-50 [17] from CLIP [34] for the vision encoder and a CLIP transformer for the language encoder. We initialize our model with CLIP and train them on EPIC-KITCHEN-100 [10, 11]. The training hyperparameters used during the pre-training are listed in Table 5. The training was conducted on two NVIDIA A800 GPUs taking approximately 30 hours. For hyperparameter sensitivity, we report the model performance under varying numbers of sampled frames and different values of the loss weight λ . As shown in Figure 8, increasing the number of sampled frames leads to higher success rates, likely because it better preserves the temporal ordering and continuity in the video sequence. The model shows low sensitivity to λ , as we observe that $\mathcal{L}_{BB}$ converges much faster than $\mathcal{L}_{VLO}$ due to its unimodal nature. As a result, $\mathcal{L}_{BB}$ primarily serves as a constraint during training rather than a dominant optimization objective.

Table 5: Hyper-parameters for pre-training.

Config	Value
Training epochs	1000
Optimizer	Adam
Learning rate	1×10^{-5}
Batch size	128
Frames per video	10
Loss weight λ	0.1
Weight decay	0.001
Momentum (β_1, β_2)	0.9, 0.999
Augmentation	RandomCropResize

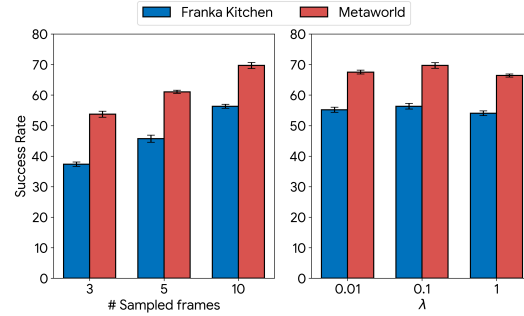


Figure 8: Hyper-parameters sensitivity.

B Evaluation Details

B.1 Simulation Environment

We follow [30] for the specific simulation environment setup and code details.

Franka Kitchen. The Franka Kitchen environment [16, 12] is based on the 9 degrees of freedom Franka robot. The Franka robot is placed in a kitchen environment containing several common household items: a microwave, a kettle, an overhead light, cabinets, and an oven. Following [30], the Franka Kitchen environments used in this paper are modified from their original design. Specifically, we introduce additional randomization to the scene by randomly altering the kitchen’s position between episodes. This modification makes the tasks significantly more challenging in terms of both perception and control.

Metaworld. The Metaworld environment [43] is an open-source simulated benchmark for robot learning. In our settings, the target object position is randomized between episodes in all tasks.

We present the specific default language instructions for each tasks in Table 6.

B.2 Real Robot Environment

Our real robot environment is a real-world office scene where the Unitree D1 robot arm can interact with a cup and a drawer. The pick cup task requires the robot to accurately identify the handle of the cup, while the open/close [X] drawer task requires the robot to understand the drawer index specified in the language instruction and align it with the visual observation. As shown in Figure 9, we use the Unitree Go app interface to remotely control the robotic arm for action data collection. Visual observations are collected using a third-person perspective web camera in a same frequency (20Hz) with action. During control, the whole system, including AcTOL and the policy MLP, runs on a GeForce GTX 880M GPU.

Table 6: Language Instructions for tasks in Franka Kitchen and Metaworld.

Environment ID	Language Instruction
kitchen_micro_open-v3	open microwave
kitchen_sdoor_open-v3	slide cabinet
kitchen_ldoor_open-v3	open left door
kitchen_knob1_on-v3	turn on stove
kitchen_light_on-v3	switch on light
hammer-v2-goal-observable	hammer nail
button-press-topdown-v2-goal-observable	press button
bin-picking-v2-goal-observable	pick and place the block between bins
assembly-v2-goal-observable	assemble the ring onto peg
drawer-open-v2-goal-observable	open drawer



Figure 9: Action space of Unitree D1 arm and the remote control interface on Unitree Go app.

B.3 Language-Conditioned Behavior Cloning Hyperparameters

We present the LCBC imitation learning hyperparameters in Table 7. For each distinct task in simulation, we run an evaluation episode every 1,000 gradient steps by running 50 roll-outs and computing their average success rate. Over a total of 10,000 gradient steps, we conduct this evaluation 10 times. The highest success rate among these 10 evaluations is reported as the final result. To ensure robustness, we average the results across two different camera viewpoints and three independent random seeds. In total, we run: 9 (tasks) * 2 (views) * 3 (demosizes) * 3 (seeds) * 6 (models) = 972 (episodes), each episode takes approximately 2 hours on our workstation with a 24-core CPU, resulting in a total of roughly 1,944 hours for the simulated LCBC experiments. For each task on the real robot, we use the final checkpoint and perform 10 evaluation runs with a fixed random seed, due to the cost of real-world policy evaluation.

Table 7: Hyper-parameters for LCBC.

	Franka Kitchen	Metaworld	Real robot
MLP achitecture	[256,256]	[256,256]	[256,256]
Non-linear activation	ReLU	ReLU	ReLU
Optimizer	Adam	Adam	Adam
Gradient Steps	10K	10K	50K
Learning rate	1×10^{-3}	1×10^{-3}	1×10^{-3}
Batch size	32	32	32
Horizon	50	100	100
Proprioception	9	4	No

B.4 Linguistic Perturbation Results

To assess the robustness of ActOL under language perturbations, we perform extensive experiments across four instruction variants. Instructions 1 and 2 transform the original action into more con-

Table 8: Success rate fluctuation across tasks in Franka Kitchen for different instruction variants.

Task	Instruction	LIV	DecisionNCE	AcTOL
Slide Cabinet	1. Please slide cabinet for me.	-32	-8	-1
	2. Help me slide cabinet.	-26	-1	3
	3. Push open the right cupboard door.	-32	-8	-1
	4. Mind pushing open the right cupboard cabinet door? I need to grab the cups inside.	-32	-6	-3
	Average	-30.5 ± 2.6	-5.8 ± 2.9	-0.5 ± 2.2
Open Left Door	1. Please open left door for me.	-3	-3	0
	2. Help me open left door.	-4	0	4
	3. Pull open the left cabinet door.	-3	-1	0
	4. Can you pull open the left cabinet door? I need to grab something inside.	-3	-1	-1
	Average	-3.3 ± 0.4	-1.3 ± 1.1	0.8 ± 1.9
Open Microwave	1. Please open microwave for me.	-5	5	-4
	2. Help me open microwave.	-4	1	-1
	3. Pop open the microwave oven door.	-5	-3	-3
	4. Would you mind helping me pop open the microwave oven door so I can heat up my lunch?	-5	1	-2
	Average	-4.8 ± 0.4	1.0 ± 2.8	-2.5 ± 1.1
Turn on Stove	1. Please turn on stove for me.	-9	-8	-2
	2. Help me turn on stove.	-8	-5	1
	3. Rotate the control knob to activate the stove.	-9	-7	1
	4. Let us rotate the control knob to activate the stove for cooking dinner.	-9	0	-2
	Average	-8.8 ± 0.4	-5.0 ± 3.1	-0.5 ± 1.5
Switch on Light	1. Please switch on light for me.	-12	2	0
	2. Help me switch on light.	-13	-4	2
	3. Flip the light switch.	-12	-5	-3
	4. Could you reach over and flip the light switch to brighten the kitchen area?	-12	-3	-6
	Average	-12.3 ± 0.4	-2.5 ± 2.7	-1.8 ± 3.0
Average		-11.9 ± 0.5	-2.7 ± 1.2	-0.9 ± 1.7

versational forms. Instruction 3 introduces vocabulary diversity by varying the verbs and nouns used. Instruction 4 further extends Instruction 3 by incorporating linguistically complex expressions generated using ChatGPT-4o. We present the comparison results obtained from experiments in the Franka Kitchen environment, with a data size of 5. As shown in Table 8, AcTOL outperforms the baselines in most instruction perturbation scenarios, thereby validating its robustness.

B.5 Language-Conditioned Behavior Cloning Results

In Table 9- 14, we report detailed Language-Conditioned Behavior Cloning results for different task and dataset size. The results demonstrate that our method achieves significant improvements across different simulation environments, varying dataset sizes, and diverse robotic manipulation tasks.

B.6 Language-Conditioned Visual Reward Results

As shown in Figure 10, we present more visualizations of Language-Conditioned Visual Reward on real-world robot manipulation videos from [2]. In Figure 10(a), the robot performs two consecutive and opposing actions. Our method effectively identifies the action boundaries and generates the correct reward sequence, increasing first and then decreasing, in alignment with the given instructions. In Figures 10(b)-(d), where the robot performs a single action, the robot initially moves slowly as it

Table 9: LCBC results when dataset size= 5 on Franka Kitchen.

Method	Slide Cabinet	Open Left Door	Open Microwave	Turn On Stove	Switch On Light	Average
CLIP	38.7 $\pm$ 5.1	2.0 $\pm$ 1.0	3.0 $\pm$ 0.0	7.0 $\pm$ 2.6	7.7 $\pm$ 1.5	11.7 $\pm$ 0.9
R3M	68.7 $\pm$ 0.6	18.3 $\pm$ 4.0	7.7 $\pm$ 3.2	19.3 $\pm$ 7.6	29.0 $\pm$ 6.1	28.6 $\pm$ 1.4
LIV	55.0 $\pm$ 1.0	6.0 $\pm$ 2.9	7.0 $\pm$ 0.6	13.0 $\pm$ 0.6	22.0 $\pm$ 2.6	20.6 $\pm$ 0.7
DecisionNCE	59.3 $\pm$ 6.8	9.7 $\pm$ 1.5	7.0 $\pm$ 2.0	26.3 $\pm$ 4.5	24.3 $\pm$ 2.5	25.3 $\pm$ 1.3
AcTOL w/o BB	71.5 $\pm$ 3.5	11.5 $\pm$ 0.7	10.5 $\pm$ 0.7	23.5 $\pm$ 6.4	47.0 $\pm$ 4.2	32.8 $\pm$ 2.8
AcTOL	85.5 $\pm$ 0.7	20.0 $\pm$ 2.1	18.3 $\pm$ 4.9	24.7 $\pm$ 4.9	62.3 $\pm$ 2.8	42.6 $\pm$ 0.3

Table 10: LCBC results when dataset size= 15 on Franka Kitchen.

Method	Slide Cabinet	Open Left Door	Open Microwave	Turn On Stove	Switch On Light	Average
CLIP	71.0 $\pm$ 3.6	8.0 $\pm$ 2.0	15.7 $\pm$ 2.1	14.7 $\pm$ 0.6	28.0 $\pm$ 1.0	27.5 $\pm$ 1.0
R3M	81.0 $\pm$ 1.0	31.0 $\pm$ 1.7	22.0 $\pm$ 2.6	19.3 $\pm$ 4.7	57.7 $\pm$ 3.8	42.2 $\pm$ 1.0
LIV	85.0 $\pm$ 5.6	19.0 $\pm$ 3.0	28.3 $\pm$ 2.9	29.7 $\pm$ 3.5	51.7 $\pm$ 2.3	42.7 $\pm$ 1.2
DecisionNCE	92.0 $\pm$ 6.6	18.7 $\pm$ 4.5	27.0 $\pm$ 4.0	33.3 $\pm$ 3.5	45.0 $\pm$ 7.5	43.2 $\pm$ 2.3
AcTOL w/o BB	84.5 $\pm$ 3.5	29.5 $\pm$ 0.7	29.5 $\pm$ 2.1	54.0 $\pm$ 2.8	73.5 $\pm$ 2.1	54.2 $\pm$ 0.8
AcTOL	99.5 $\pm$ 0.7	37.5 $\pm$ 5.6	37.0 $\pm$ 4.2	53.5 $\pm$ 3.5	81.5 $\pm$ 2.1	61.8 $\pm$ 2.5

Table 11: LCBC results when dataset size= 25 on Franka Kitchen.

Method	Slide Cabinet	Open Left Door	Open Microwave	Turn On Stove	Switch On Light	Average
CLIP	66.3 $\pm$ 7.5	8.7 $\pm$ 1.2	18.7 $\pm$ 1.5	23.7 $\pm$ 3.1	38.7 $\pm$ 2.3	31.2 $\pm$ 2.6
R3M	84.7 $\pm$ 6.8	35.3 $\pm$ 4.0	40.0 $\pm$ 1.0	34.0 $\pm$ 5.3	61.7 $\pm$ 10.7	51.1 $\pm$ 2.8
LIV	91.7 $\pm$ 5.9	26.0 $\pm$ 2.6	35.0 $\pm$ 4.6	45.3 $\pm$ 0.6	61.7 $\pm$ 3.2	51.9 $\pm$ 0.9
DecisionNCE	91.7 $\pm$ 1.5	27.0 $\pm$ 10.4	37.0 $\pm$ 1.7	47.3 $\pm$ 1.2	51.3 $\pm$ 4.0	50.9 $\pm$ 2.9
AcTOL w/o BB	92.0 $\pm$ 2.4	37.0 $\pm$ 5.4	40.0 $\pm$ 2.4	57.0 $\pm$ 1.5	78.0 $\pm$ 6.2	60.8 $\pm$ 1.3
AcTOL	100.0 $\pm$ 0.0	37.0 $\pm$ 7.1	42.5 $\pm$ 2.1	62.5 $\pm$ 2.1	81.0 $\pm$ 4.2	64.6 $\pm$ 0.6

Table 12: LCBC results when dataset size= 5 on Metaworld.

Method	Assembly	Pick bin	Press button	Hammer	Open drawer	Average
CLIP	48.3 $\pm$ 5.7	35.3 $\pm$ 2.3	34.3 $\pm$ 4.9	51.2 $\pm$ 2.8	91.0 $\pm$ 1.0	52.0 $\pm$ 2.7
R3M	63.5 $\pm$ 5.6	33.3 $\pm$ 5.1	27.3 $\pm$ 5.1	63.2 $\pm$ 7.1	92.3 $\pm$ 0.6	55.9 $\pm$ 3.9
LIV	61.8 $\pm$ 6.5	32.3 $\pm$ 9.0	32.7 $\pm$ 3.5	61.0 $\pm$ 6.1	100.0 $\pm$ 0.0	57.7 $\pm$ 2.1
DecisionNCE	54.0 $\pm$ 3.6	31.0 $\pm$ 3.6	27.7 $\pm$ 5.5	65.7 $\pm$ 3.8	100.0 $\pm$ 0.0	55.7 $\pm$ 2.8
AcTOL w/o BB	66.8 $\pm$ 1.4	39.0 $\pm$ 16.8	20.7 $\pm$ 1.5	74.7 $\pm$ 1.5	100.0 $\pm$ 0.0	60.2 $\pm$ 5.1
AcTOL	62.8 $\pm$ 6.0	41.0 $\pm$ 6.3	42.0 $\pm$ 4.5	69.5 $\pm$ 0.7	100.0 $\pm$ 0.0	63.1 $\pm$ 3.9

Table 13: LCBC results when dataset size= 15 on Metaworld.

Method	Assembly	Pick bin	Press button	Hammer	Open drawer	Average
CLIP	73.0 $\pm$ 7.8	40.3 $\pm$ 5.5	52.0 $\pm$ 7.9	76.0 $\pm$ 5.0	96.7 $\pm$ 0.6	67.6 $\pm$ 1.5
R3M	80.7 $\pm$ 7.6	17.0 $\pm$ 12.3	45.0 $\pm$ 4.6	83.3 $\pm$ 4.5	94.0 $\pm$ 1.0	64.0 $\pm$ 5.2
LIV	84.3 $\pm$ 2.5	37.0 $\pm$ 8.7	54.7 $\pm$ 3.8	81.3 $\pm$ 5.9	100.0 $\pm$ 0.0	71.4 $\pm$ 3.6
DecisionNCE	73.3 $\pm$ 10.8	36.7 $\pm$ 5.0	43.3 $\pm$ 2.1	83.0 $\pm$ 6.0	100.0 $\pm$ 0.0	67.3 $\pm$ 1.8
AcTOL w/o BB	94.0 $\pm$ 3.0	50.3 $\pm$ 18.6	48.3 $\pm$ 1.5	90.7 $\pm$ 1.2	100.0 $\pm$ 0.0	76.7 $\pm$ 5.3
AcTOL	82.5 $\pm$ 0.7	64.5 $\pm$ 3.2	65.5 $\pm$ 3.9	84.0 $\pm$ 2.1	100.0 $\pm$ 0.0	79.3 $\pm$ 1.6

Table 14: LCBC results when dataset size= 25 on Metaworld.

Method	Assembly	Pick bin	Press button	Hammer	Open drawer	Average
CLIP	69.3 $\pm$ 5.7	36.0 $\pm$ 11.8	66.0 $\pm$ 2.5	78.8 $\pm$ 4.9	99.3 $\pm$ 0.6	69.9 $\pm$ 4.4
R3M	87.7 $\pm$ 2.4	14.7 $\pm$ 11.6	48.3 $\pm$ 2.1	89.7 $\pm$ 3.5	100.0 $\pm$ 0.0	68.1 $\pm$ 3.6
LIV	87.3 $\pm$ 5.5	23.7 $\pm$ 6.8	66.0 $\pm$ 6.8	89.7 $\pm$ 2.5	100.0 $\pm$ 0.0	73.3 $\pm$ 1.5
DecisionNCE	85.7 $\pm$ 4.9	47.0 $\pm$ 12.8	58.0 $\pm$ 7.8	88.3 $\pm$ 6.7	100.0 $\pm$ 0.0	75.8 $\pm$ 3.9
AcTOL w/o BB	93.7 $\pm$ 0.6	51.7 $\pm$ 11.9	55.0 $\pm$ 3.5	93.0 $\pm$ 1.0	100.0 $\pm$ 0.0	78.7 $\pm$ 3.5
AcTOL	93.5 $\pm$ 3.4	66.0 $\pm$ 2.8	76.5 $\pm$ 4.9	88.5 $\pm$ 3.9	100.0 $\pm$ 0.0	84.9 $\pm$ 1.6

searches for the target. Correspondingly, the reward grows gradually. Once the robot interacts with the object and completes the task, our method captures the distinct semantic changes in the action, leading to a rapid reward increase. In Figures 10(e)-(f), we test two complex actions and instructions

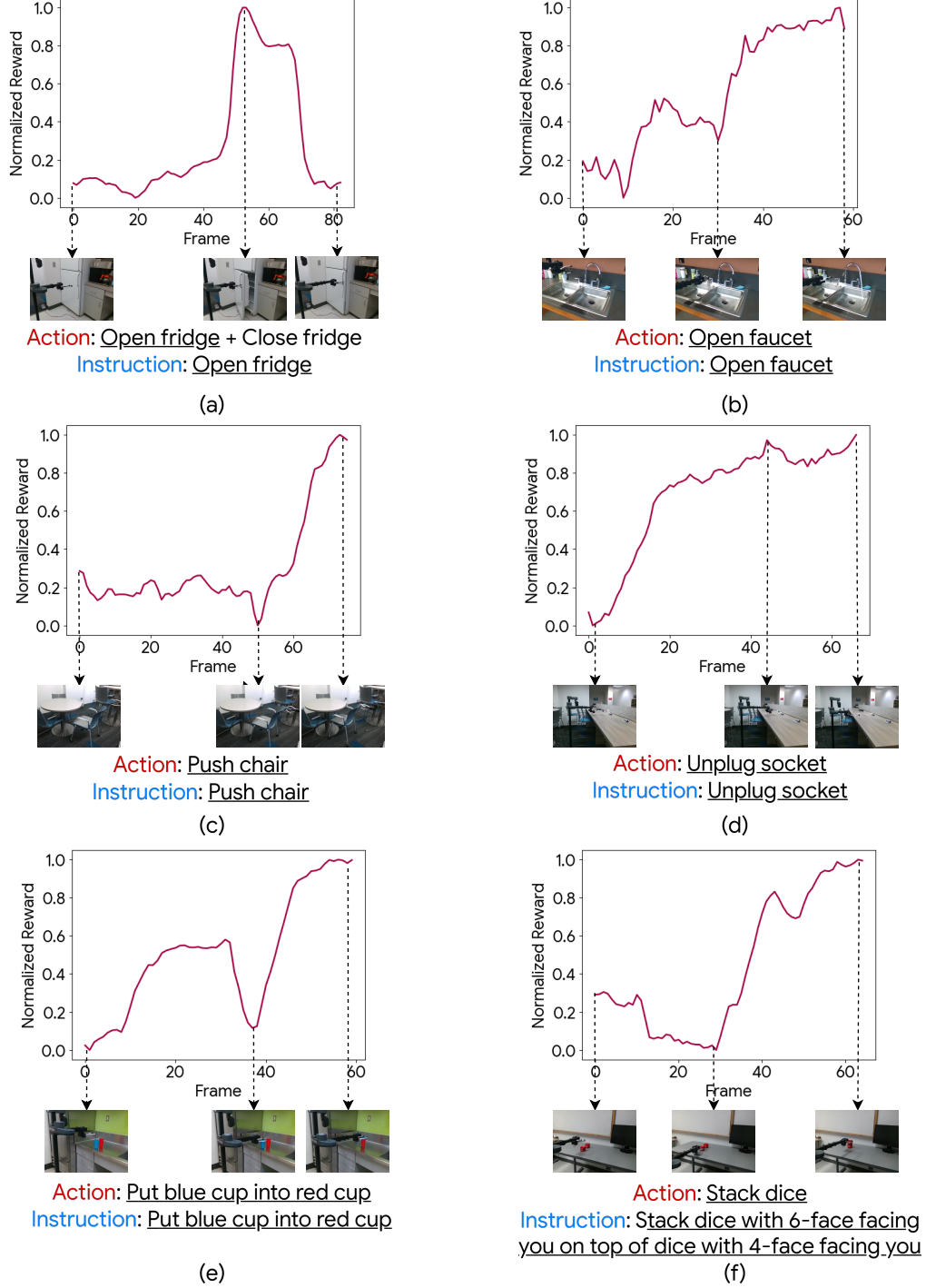


Figure 10: Reward plots for exemplar robot action videos.

to explore the limits of our method. In Figure 10(e), the model is required to accurately distinguish between the blue and red cups to complete the task. In Figure 10(f), the model needs to differentiate the orientation and face values of two dice. These scenarios impose high demands on the model's visual and semantic understanding. Our method successfully produces the correct rewards in both tasks, showcasing its potential for application in real-world, complex scenarios.

C Proofs

C.1 Proofs of Theorem 1

For the proof of Theorem 1, we closely follow the approaches presented in [45] and adapted to our triplet case. We prove the theorem in three steps:

- (1) $\mathcal{L}^* := \frac{1}{T(T-1)} \sum_{i=1}^T \sum_{m=1}^{M_i} n_{i,m} \log n_{i,m}$ is a lower bound of $\mathcal{L}_{\text{VLO}}$, i.e., $\mathcal{L}_{\text{VLO}} > \mathcal{L}^*$.
 - (2) $\mathcal{L}^*$ is tight, i.e., for any $\epsilon > 0$, there exists representations such that $\mathcal{L}_{\text{VLO}} < \mathcal{L}^* + \epsilon$.
 - (3) For any $0 < \delta < 1$, there exist $\epsilon > 0$, such that if $\mathcal{L}_{\text{VLO}} < \mathcal{L}^* + \epsilon$, then the learned representations satisfy VLO property.
- (1) Recall that $\mathcal{L}_{\text{VLO}} = \frac{1}{T} \sum_{i=1}^T \frac{1}{T-1} \sum_{j=1, j \neq i}^T -\log \frac{\exp(\mathfrak{R}_{i,j,l})}{\sum_{\mathbf{v}^k \in \mathcal{N}_{i,j}} \exp(\mathfrak{R}_{i,k,l})}$, where $\mathcal{N}_{i,j} = \{\mathbf{v}_k | k \neq i, d_{i,j} < d_{i,k}\}$, we rewrite it as

$$\begin{aligned}
\mathcal{L}_{\text{VLO}} &= -\frac{1}{T(T-1)} \sum_{i=1}^T \sum_{j \in [T] \setminus \{i\}} \log \frac{\exp(\mathfrak{R}_{i,j,l})}{\sum_{k \in [T] \setminus \{i\}, d_{i,k} \geq d_{i,j}} \exp(\mathfrak{R}_{i,k,l})} \\
&= -\frac{1}{T(T-1)} \sum_{i=1}^T \sum_{m=1}^{M_i} \sum_{j \in [T] \setminus \{i\}, d_{i,j} = D_{i,m}} \log \frac{\exp(\mathfrak{R}_{i,j,l})}{\sum_{k \in [T] \setminus \{i\}, d_{i,k} \geq D_{i,m}} \exp(\mathfrak{R}_{i,k,l})} \\
&= -\frac{1}{T(T-1)} \sum_{i=1}^T \sum_{m=1}^{M_i} \sum_{j \in [T] \setminus \{i\}, d_{i,j} = D_{i,m}} \log \frac{1}{\sum_{k \in [T] \setminus \{i\}, d_{i,k} \geq D_{i,m}} \exp(\mathfrak{R}_{i,k,l} - \mathfrak{R}_{i,j,l})} \\
&= -\frac{1}{T(T-1)} \sum_{i=1}^T \sum_{m=1}^{M_i} \sum_{j \in [T] \setminus \{i\}, d_{i,j} = D_{i,m}} \log \frac{1}{\sum_{k \in [T] \setminus \{i\}, d_{i,k} = D_{i,m}} \exp(\mathfrak{R}_{i,k,l} - \mathfrak{R}_{i,j,l})} \\
&\quad - \frac{1}{T(T-1)} \sum_{i=1}^T \sum_{m=1}^{M_i} \sum_{j \in [T] \setminus \{i\}, d_{i,j} = D_{i,m}} \log \frac{\sum_{k \in [T] \setminus \{i\}, d_{i,k} = D_{i,m}} \exp(\mathfrak{R}_{i,k,l} - \mathfrak{R}_{i,j,l})}{\sum_{k \in [T] \setminus \{i\}, d_{i,k} \geq D_{i,m}} \exp(\mathfrak{R}_{i,k,l} - \mathfrak{R}_{i,j,l})} \\
&= -\frac{1}{T(T-1)} \sum_{i=1}^T \sum_{m=1}^{M_i} \sum_{j \in [T] \setminus \{i\}, d_{i,j} = D_{i,m}} \log \frac{\exp(\mathfrak{R}_{i,j,l})}{\sum_{k \in [T] \setminus \{i\}, d_{i,k} = D_{i,m}} \exp(\mathfrak{R}_{i,k,l})} \\
&\quad + \frac{1}{T(T-1)} \sum_{i=1}^T \sum_{m=1}^{M_i} \sum_{j \in [T] \setminus \{i\}, d_{i,j} = D_{i,m}} \log \left(1 + \frac{\sum_{k \in [T] \setminus \{i\}, d_{i,k} > D_{i,m}} \exp(\mathfrak{R}_{i,k,l} - \mathfrak{R}_{i,j,l})}{\sum_{k \in [T] \setminus \{i\}, d_{i,k} = D_{i,m}} \exp(\mathfrak{R}_{i,k,l} - \mathfrak{R}_{i,j,l})} \right) \\
&> -\frac{1}{T(T-1)} \sum_{i=1}^T \sum_{m=1}^{M_i} \sum_{j \in [T] \setminus \{i\}, d_{i,j} = D_{i,m}} \log \frac{\exp(\mathfrak{R}_{i,j,l})}{\sum_{k \in [T] \setminus \{i\}, d_{i,k} = D_{i,m}} \exp(\mathfrak{R}_{i,k,l})}.
\end{aligned} \tag{6}$$

$\forall i \in [T], m \in [M_i]$, from Jensen's Inequality we have

$$\begin{aligned}
& - \sum_{j \in [T] \setminus \{i\}, d_{i,j} = D_{i,m}} \log \frac{\exp(\mathfrak{R}_{i,j,l})}{\sum_{k \in [T] \setminus \{i\}, d_{i,k} = D_{i,m}} \exp(\mathfrak{R}_{i,k,l})} \\
& \geq -n_{i,m} \log \left(\frac{1}{n_{i,m}} \sum_{j \in [T] \setminus \{i\}, d_{i,j} = D_{i,m}} \frac{\exp(\mathfrak{R}_{i,j,l})}{\sum_{k \in [T] \setminus \{i\}, d_{i,k} = D_{i,m}} \exp(\mathfrak{R}_{i,k,l})} \right) = n_{i,m} \log n_{i,m}.
\end{aligned} \tag{7}$$

Thus, by plugging Eq. (7) into Eq. (6), we have

$$\mathcal{L}_{\text{VLO}} > \frac{1}{T(T-1)} \sum_{i=1}^T \sum_{m=1}^{M_i} n_{i,m} \log n_{i,m} = L^* \quad (8)$$

(2) We will show for $\forall \epsilon > 0$, there is a set of representations where

$$\begin{cases} \mathfrak{R}_{i,j,l} > \mathfrak{R}_{i,k,l} + \gamma \text{ if } d_{i,j} < d_{i,k} \\ \mathfrak{R}_{i,j,l} = \mathfrak{R}_{i,k,l} \text{ if } d_{i,j} = d_{i,k} \end{cases}$$

and $\gamma := \log \frac{T}{\min_{i \in [T], m \in [M_i]} n_{i,m} \epsilon}$, $\forall i \in [T], j, k \in [T] \setminus \{i\}$, such that $\mathcal{L}_{\text{VLO}} < L^* + \epsilon$. For such a set of representations, $\forall i \in [T], m \in [M_i], j \in \{[T] \setminus \{i\} \mid d_{i,j} = D_{i,m}\}$,

$$-\log \frac{\exp(\mathfrak{R}_{i,j,l})}{\sum_{k \in [T] \setminus \{i\}, d_{i,k} = D_{i,m}} \exp(\mathfrak{R}_{i,k,l})} = \log n_{i,m} \quad (9)$$

since $\mathfrak{R}_{i,k,l} = \mathfrak{R}_{i,j,l}$ for all k such that $d_{i,k} = D_{i,m} = d_{i,j}$, and

$$\begin{aligned} & \log \left(1 + \frac{\sum_{k \in [T] \setminus \{i\}, d_{i,k} > D_{i,m}} \exp(\mathfrak{R}_{i,k,l} - \mathfrak{R}_{i,j,l})}{\sum_{k \in [T] \setminus \{i\}, d_{i,k} = D_{i,m}} \exp(\mathfrak{R}_{i,k,l} - \mathfrak{R}_{i,j,l})} \right) \\ & < \log \left(1 + \frac{T \exp(-\gamma)}{n_{i,m}} \right) < \frac{T \exp(-\gamma)}{n_{i,m}} \leq \epsilon. \end{aligned} \quad (10)$$

As $\mathfrak{R}_{i,k,l} - \mathfrak{R}_{i,j,l} < -\gamma$ for all k such that $d_{i,k} > D_{i,m} = d_{i,j}$ and $\mathfrak{R}_{i,k,l} - \mathfrak{R}_{i,j,l} = 0$ for all k such that $d_{i,k} = D_{i,m} = d_{i,j}$. From Eq. (6) we have

$$\begin{aligned} \mathcal{L}_{\text{VLO}} = & -\frac{1}{T(T-1)} \sum_{i=1}^T \sum_{m=1}^{M_i} \sum_{\substack{j \in [T] \setminus \{i\} \\ d_{i,j} = D_{i,m}}} \log \frac{\exp(\mathfrak{R}_{i,j,l})}{\sum_{\substack{k \in [T] \setminus \{i\} \\ d_{i,k} = D_{i,m}}} \exp(\mathfrak{R}_{i,k,l})} \\ & + \frac{1}{T(T-1)} \sum_{i=1}^T \sum_{m=1}^{M_i} \sum_{\substack{j \in [T] \setminus \{i\} \\ d_{i,j} = D_{i,m}}} \log \left(1 + \frac{\sum_{\substack{k \in [T] \setminus \{i\} \\ d_{i,k} > D_{i,m}}} \exp(\mathfrak{R}_{i,k,l} - \mathfrak{R}_{i,j,l})}{\sum_{\substack{k \in [T] \setminus \{i\} \\ d_{i,k} = D_{i,m}}} \exp(\mathfrak{R}_{i,k,l} - \mathfrak{R}_{i,j,l})} \right) \end{aligned} \quad (11)$$

By plugging Eq. (9) and Eq. (10) into Eq. (11) we have

$$\mathcal{L}_{\text{VLO}} < \frac{1}{T(T-1)} \sum_{i=1}^T \sum_{m=1}^{M_i} n_{i,m} \log n_{i,m} + \epsilon = L^* + \epsilon \quad (12)$$

(3) We will show $\forall 0 < \delta < 1$, there is a

$$\epsilon = \frac{1}{T(T-1)} \min \left(\min_{i \in [T], m \in [M_i]} \log \left(1 + \frac{1}{n_{i,m} \exp(\delta + \frac{1}{\delta})} \right), 2 \log \frac{1 + \exp(\delta)}{2} - \delta \right) > 0,$$

such that when $\mathcal{L}_{\text{VLO}} < L^* + \epsilon$, the representations satisfy VLO property. We first show that $|\mathfrak{R}_{i,j,l} - \mathfrak{R}_{i,k,l}| < \delta$ if $d_{i,j} = d_{i,k}$, $i \in [T], j, k \in [T] \setminus \{i\}$ when $\mathcal{L}_{\text{VLO}} < L^* + \epsilon$. From Eq. (6) we have

$$\mathcal{L}_{\text{VLO}} > -\frac{1}{T(T-1)} \sum_{i=1}^T \sum_{m=1}^{M_i} \sum_{j \in [T] \setminus \{i\}, d_{i,j}=D_{i,m}} \log \frac{\exp(\mathfrak{R}_{i,j,l})}{\sum_{k \in [T] \setminus \{i\}, d_{i,k}=D_{i,m}} \exp(\mathfrak{R}_{i,k,l})} \quad (13)$$

Let $p_{i,m} := \arg \min_{j \in [T] \setminus \{i\}, d_{i,j}=D_{i,m}} \mathfrak{R}_{i,j,l}$, $q_{i,m} := \arg \max_{j \in [T] \setminus \{i\}, d_{i,j}=D_{i,m}} \mathfrak{R}_{i,j,l}$, $\zeta_{i,m} := \mathfrak{R}_{i,p_{i,m},l}$, $\eta_{i,m} := s_{i,q_{i,m},l} - s_{i,p_{i,m},l}$, $\forall i \in [T], m \in [M_i]$, by splitting out the maximum term and the minimum term we have

$$\begin{aligned} \mathcal{L}_{\text{VLO}} &> -\frac{1}{T(T-1)} \sum_{i=1}^T \sum_{m=1}^{M_i} \left\{ \log \frac{\exp(\zeta_{i,m})}{\sum_{k \in [T] \setminus \{i\}, d_{i,k}=D_{i,m}} \exp(\mathfrak{R}_{i,k,l})} \right. \\ &\quad \left. + \log \frac{\exp(\zeta_{i,m} + \eta_{i,m})}{\sum_{k \in [T] \setminus \{i\}, d_{i,k}=D_{i,m}} \exp(\mathfrak{R}_{i,k,l})} + \log \frac{\exp\left(\sum_{j \in [T] \setminus \{i, p_{i,m}, q_{i,m}\}, d_{i,j}=D_{i,m}} \mathfrak{R}_{i,j,l}\right)}{\left(\sum_{k \in [T] \setminus \{i\}, d_{i,k}=D_{i,m}} \exp(\mathfrak{R}_{i,k,l})\right)^{n_{i,m}-2}} \right\}. \end{aligned} \quad (14)$$

Let $\theta_{i,m} := \frac{1}{n_{i,m}-2} \sum_{j \in [T] \setminus \{i, p_{i,m}, q_{i,m}\}, d_{i,j}=D_{i,m}} \exp(\mathfrak{R}_{i,j,l} - \zeta_{i,m})$, we have

$$-\log \frac{\exp(\zeta_{i,m})}{\sum_{k \in [T] \setminus \{i\}, d_{i,k}=D_{i,m}} \exp(\mathfrak{R}_{i,k,l})} = \log(1 + \exp(\eta_{i,m}) + (n_{i,m} - 2)\theta_{i,m}) \quad (15)$$

and

$$-\log \frac{\exp(\zeta_{i,m} + \eta_{i,m})}{\sum_{k \in [T] \setminus \{i\}, d_{i,k}=D_{i,m}} \exp(\mathfrak{R}_{i,k,l})} = \log(1 + \exp(\eta_{i,m}) + (n_{i,m} - 2)\theta_{i,m}) - \eta_{i,m} \quad (16)$$

Then, from Jensen's inequality, we know

$$\exp\left(\sum_{\substack{j \in [T] \setminus \{i, p_{i,m}, q_{i,m}\} \\ d_{i,j}=D_{i,m}}} \mathfrak{R}_{i,j,l}\right) \leq \left(\frac{1}{n_{i,m}-2} \sum_{\substack{j \in [T] \setminus \{i, p_{i,m}, q_{i,m}\} \\ d_{i,j}=D_{i,m}}} \exp(\mathfrak{R}_{i,j,l})\right)^{n_{i,m}-2} \quad (17)$$

thus

$$\begin{aligned} &-\log \frac{\exp\left(\sum_{\substack{j \in [T] \setminus \{i, p_{i,m}, q_{i,m}\} \\ d_{i,j}=D_{i,m}}} \mathfrak{R}_{i,j,l}\right)}{\left(\sum_{\substack{k \in [T] \setminus \{i\} \\ d_{i,k}=D_{i,m}}} \exp(\mathfrak{R}_{i,k,l})\right)^{n_{i,m}-2}} \\ &\geq (n_{i,m} - 2) \log(1 + \exp(\eta_{i,m}) + (n_{i,m} - 2)\theta_{i,m}) - (n_{i,m} - 2) \log(\theta_{i,m}) \end{aligned} \quad (18)$$

By plugging Eq. (15), Eq. (16) and Eq. (18) into Eq. (14), we have

$$\begin{aligned} \mathcal{L}_{\text{VLO}} &> \frac{1}{T(T-1)} \sum_{i=1}^T \sum_{m=1}^{M_i} \left(n_{i,m} \log(1 + \exp(\eta_{i,m}) + (n_{i,m} - 2)\theta_{i,m}) \right. \\ &\quad \left. - \eta_{i,m} - (n_{i,m} - 2) \log(\theta_{i,m}) \right) \end{aligned} \quad (19)$$

Let $h(\theta) := n_{i,m} \log(1 + \exp(\eta_{i,m}) + (n_{i,m} - 2)\theta) - \eta_{i,m} - (n_{i,m} - 2) \log(\theta)$. From derivative analysis we know $h(\theta)$ decreases monotonically when $\theta \in \left[1, \frac{1+\exp(\eta_{i,m})}{2}\right]$ and increases monotonically when $\theta \in \left[\frac{1+\exp(\eta_{i,m})}{2}, \exp(\eta_{i,m})\right]$, thus

$$h(\theta) \geq h\left(\frac{1 + \exp(\eta_{i,m})}{2}\right) = n_{i,m} \log n_{i,m} + 2 \log \frac{1 + \exp(\eta_{i,m})}{2} - \eta_{i,m}. \quad (20)$$

By plugging Eq. (20) into Eq. (19), we have

$$\begin{aligned} \mathcal{L}_{\text{VLO}} &> \frac{1}{T(T-1)} \sum_{i=1}^T \sum_{m=1}^{M_i} \left(n_{i,m} \log n_{i,m} + 2 \log \frac{1 + \exp(\eta_{i,m})}{2} - \eta_{i,m} \right) \\ &= L^* + \frac{1}{T(T-1)} \sum_{i=1}^T \sum_{m=1}^{M_i} \left(2 \log \frac{1 + \exp(\eta_{i,m})}{2} - \eta_{i,m} \right) \end{aligned} \quad (21)$$

Then, since $\eta_{i,m} \geq 0$, we have $2 \log \frac{1+\exp(\eta_{i,m})}{2} - \eta_{i,m} \geq 0$. Thus, $\forall i \in [T], m \in [M_i]$,

$$\mathcal{L}_{\text{VLO}} > L^* + \frac{1}{T(T-1)} \left(2 \log \frac{1 + \exp(\eta_{i,m})}{2} - \eta_{i,m} \right) \quad (22)$$

If $\mathcal{L}_{\text{VLO}} < L^* + \epsilon \leq L^* + \frac{1}{T(T-1)} \left(2 \log \frac{1+\exp(\delta)}{2} - \delta \right)$, then

$$2 \log \frac{1 + \exp(\eta_{i,m})}{2} - \eta_{i,m} < 2 \log \frac{1 + \exp(\delta)}{2} - \delta \quad (23)$$

Since $y(x) = 2 \log \frac{1+\exp(x)}{2} - x$ increases monotonically when $x > 0$, we have $\eta_{i,m} < \delta$. Hence $\forall i \in [T], j, k \in [T] \setminus \{i\}$, if $d_{i,j} = d_{i,k} = D_{i,m}$, $|\mathfrak{R}_{i,j,l} - \mathfrak{R}_{i,k,l}| \leq \eta_{i,m} < \delta$. Next, we show $\mathfrak{R}_{i,j,l} > \mathfrak{R}_{i,k,l} + \delta$ if $d_{i,j} < d_{i,k}$ when $\mathcal{L}_{\text{VLO}} < L^* + \epsilon$. From Eq. (6) we have

$$\begin{aligned} \mathcal{L}_{\text{VLO}} &= -\frac{1}{T(T-1)} \sum_{i=1}^T \sum_{m=1}^{M_i} \sum_{\substack{j \in [T] \setminus \{i\} \\ d_{i,j} = D_{i,m}}} \log \frac{\exp(\mathfrak{R}_{i,j,l})}{\sum_{\substack{k \in [T] \setminus \{i\} \\ d_{i,k} = D_{i,m}}} \exp(\mathfrak{R}_{i,k,l})} \\ &\quad + \frac{1}{T(T-1)} \sum_{i=1}^T \sum_{m=1}^{M_i} \sum_{\substack{j \in [T] \setminus \{i\} \\ d_{i,j} = D_{i,m}}} \log \left(1 + \frac{\sum_{\substack{k \in [T] \setminus \{i\} \\ d_{i,k} = D_{i,m}}} \exp(\mathfrak{R}_{i,k,l} - \mathfrak{R}_{i,j,l})}{\sum_{\substack{k \in [T] \setminus \{i\} \\ d_{i,k} > D_{i,m}}} \exp(\mathfrak{R}_{i,k,l} - \mathfrak{R}_{i,j,l})} \right) \end{aligned} \quad (24)$$

and combining it with Eq. (7) we have

$$\begin{aligned} \mathcal{L}_{\text{VLO}} &\geq L^* + \frac{1}{T(T-1)} \sum_{i=1}^T \sum_{m=1}^{M_i} \sum_{\substack{j \in [T] \setminus \{i\} \\ d_{i,j} = D_{i,m}}} \log \left(1 + \frac{\sum_{\substack{k \in [T] \setminus \{i\} \\ d_{i,k} > D_{i,m}}} \exp(\mathfrak{R}_{i,k,l} - \mathfrak{R}_{i,j,l})}{\sum_{\substack{k \in [T] \setminus \{i\} \\ d_{i,k} = D_{i,m}}} \exp(\mathfrak{R}_{i,k,l} - \mathfrak{R}_{i,j,l})} \right) \\ &> L^* + \frac{1}{T(T-1)} \log \left(1 + \frac{\exp(\mathfrak{R}_{i,k,l} - \mathfrak{R}_{i,j,l})}{\sum_{\substack{h \in [T] \setminus \{i\} \\ d_{i,h} = d_{i,j}}} \exp(\mathfrak{R}_{i,h,l} - \mathfrak{R}_{i,j,l})} \right) \end{aligned} \quad (25)$$

$\forall i \in [T], j \in [T] \setminus \{i\}, k \in \{k \in [T] \setminus \{i\} \mid d_{i,j} < d_{i,k}\}$. When $\mathcal{L}_{\text{VLO}} < L^* + \epsilon$, we already have $|\mathfrak{R}_{i,h,l} - \mathfrak{R}_{i,j,l}| < \delta, \forall d_{i,h} = d_{i,j}$, which derives $\mathfrak{R}_{i,h,l} - \mathfrak{R}_{i,j,l} < \delta$ and thus $\exp(\mathfrak{R}_{i,h,l} - \mathfrak{R}_{i,j,l}) < \exp(\delta)$. By putting this into Eq. (24), we have $\forall i \in [T], j \in [T] \setminus \{i\}, k \in \{k \in [T] \setminus \{i\} \mid d_{i,j} < d_{i,k}\}$,

$$\mathcal{L}_{\text{VLO}} > L^* + \frac{1}{T(T-1)} \log \left(1 + \frac{\exp(\mathfrak{R}_{i,k,l} - \mathfrak{R}_{i,j,l})}{n_{i,r_{i,j}} \exp(\delta)} \right) \quad (26)$$

where $r_{i,j} \in [M_i]$ is the index such that $D_{i,r_{i,j}} = d_{i,j}$.

Further, given $\mathcal{L}_{\text{VLO}} < L^* + \epsilon < L^* + \frac{1}{T(T-1)} \log \left(1 + \frac{1}{n_{i,r_{i,j}} \exp(\delta + \frac{1}{\delta})} \right)$, we have

$$\log \left(1 + \frac{\exp(\mathfrak{R}_{i,k,l} - \mathfrak{R}_{i,j,l})}{n_{i,r_{i,j}} \exp(\delta)} \right) < \log \left(1 + \frac{1}{n_{i,r_{i,j}} \exp(\delta + \frac{1}{\delta})} \right) \quad (27)$$

which derives $\mathfrak{R}_{i,j,l} > \mathfrak{R}_{i,k,l} + \frac{1}{\delta}, \forall i \in [T], j \in [T] \setminus \{i\}, k \in \{[T] \setminus \{i\} \mid d_{i,j} < d_{i,k}\}$. Finally, $\forall i \in [T], j, k \in [T] \setminus \{i\}, \mathfrak{R}_{i,j,l} < \mathfrak{R}_{i,k,l} - \frac{1}{\delta}$ if $d_{i,j} > d_{i,k}$ directly follows from $\mathfrak{R}_{i,j,l} > \mathfrak{R}_{i,k,l} + \frac{1}{\delta}$ if $d_{i,j} < d_{i,k}$. $\square$

C.2 Proofs of Theorem 2

Setup and Assumptions. To provide the vision-language continuity, we first assume that the frame embeddings $\{\mathbf{v}_t\}$, where $t \in [1, T]$ are regularized under a Brownian Bridge process $\mathbf{B}(t)$ as discussed in Section 3.2, where the transition density for any intermediate time $t \in [n(i), n(j)]$ within a sampled interval is given as:

$$\mathbf{B}(t) \sim \mathcal{N}(\mathbb{E}[\mathbf{B}(t)], \text{Var}[\mathbf{B}(t)]), \quad (28)$$

with:

$$\mathbb{E}[\mathbf{B}(t)] = \mathbf{v}_i + \frac{t - n(i)}{n(j) - n(i)}(\mathbf{v}_j - \mathbf{v}_i), \quad \text{Var}[\mathbf{B}(t)] = \frac{(t - n(i))(n(j) - t)}{n(j) - n(i)}. \quad (29)$$

All time steps $t \in [1, T]$ are covered by at least one sampled interval, ensuring the entire video sequence satisfies the Brownian Bridge regularization. Now, let $\mathbf{v}_k, \mathbf{v}_l \in \mathbb{R}^d$ be arbitrary embeddings, *not necessarily* the endpoints $\mathbf{v}_i$ and $\mathbf{v}_j$ of a sampled interval. These embeddings fall within the union $\mathfrak{U}$ of all sampled local intervals. Without loss of generality, here we can identify the interval $[n(i), n(j)] \in \mathfrak{U}$ from the union containing $\mathbf{v}_k$ and $\mathbf{v}_l$.

Bounding Local Continuity. Recall that semantic alignment score $\mathfrak{R}(\mathbf{v}_k, \mathbf{v}_l, \mathbf{l})$ is defined as:

$$\mathfrak{R}(\mathbf{v}_k, \mathbf{v}_l, \mathbf{l}) = -\|\text{sim}(\mathbf{v}_k, \mathbf{l}) - \text{sim}(\mathbf{v}_l, \mathbf{l})\|_2,$$

where $\text{sim}(\cdot)$ is Lipschitz continuous with constant $C > 0$ when embeddings are normalized as unit vectors. By the Lipschitz continuity of $\text{sim}(\cdot)$, we have:

$$\|\text{sim}(\mathbf{v}_k, \mathbf{l}) - \text{sim}(\mathbf{v}_l, \mathbf{l})\|_2 \leq C \cdot \|\mathbf{v}_k - \mathbf{v}_l\|_2.$$

To ensure the continuity of $\mathfrak{R}$, we must bound $\|\mathbf{v}_k - \mathbf{v}_l\|_2$. Under the Brownian Bridge regularization, the embeddings are aligned with the mean trajectory $\mathbb{E}[\mathbf{B}(t)]$, and deviations are constrained by the variance $\text{Var}[\mathbf{B}(t)]$. Specifically:

$$\|\mathbf{v}_t - \mathbb{E}[\mathbf{B}(t)]\|_2^2 \leq \lambda \cdot \text{Var}[\mathbf{B}(t)],$$

where $\lambda > 0$ depends on the strength of the Brownian Bridge loss $\mathcal{L}_{\text{BB}}$. Below we omit λ for simplicity. Substituting the variance:

$$\text{Var}[\mathbf{B}(t)] = \frac{(t - n(i))(n(j) - t)}{n(j) - n(i)}.$$

Bounding Pairwise Distance. The total pairwise distance between $\mathbf{v}_k$ and $\mathbf{v}_l$ can be expressed as:

$$\|\mathbf{v}_k - \mathbf{v}_l\|_2 \leq \|\mathbb{E}[\mathbf{B}(k)] - \mathbb{E}[\mathbf{B}(l)]\|_2 + \sqrt{\text{Var}[\mathbf{B}(k)]} + \sqrt{\text{Var}[\mathbf{B}(l)]}.$$

Since the mean trajectory $\mathbb{E}[\mathbf{B}(t)]$ is linear within the interval $[n(i), n(j)]$, we have:

$$\|\mathbb{E}[\mathbf{B}(k)] - \mathbb{E}[\mathbf{B}(l)]\|_2 \leq \frac{|k - l|}{n(j) - n(i)} \|\mathbf{v}_j - \mathbf{v}_i\|_2.$$

Combining these bounds, now we can rewrite into the following inequality:

$$\|\mathbf{v}_k - \mathbf{v}_l\|_2 \leq \frac{|k - l|}{n(j) - n(i)} \|\mathbf{v}_j - \mathbf{v}_i\|_2 + \sqrt{\frac{(k - n(i))(n(j) - k)}{n(j) - n(i)}} + \sqrt{\frac{(l - n(i))(n(j) - l)}{n(j) - n(i)}}.$$

For the variance terms, the Brownian Bridge process achieves its maximum variance at the midpoint $t = \frac{n(i) + n(j)}{2}$. This gives us,

$$\text{Var}[\mathbf{B}(t_{\max})] = \frac{n(j) - n(i)}{4}, \quad \|\mathbf{v}_k - \mathbf{v}_l\|_2 \leq 2 \frac{|k - l|}{n(j) - n(i)} + \sqrt{n(j) - n(i)}.$$

Bounding Semantic Alignment Score. Finally, by substituting this bound into the Lipschitz continuity of sim, we obtain,

$$|\mathfrak{R}(\mathbf{v}_k, \mathbf{v}_l, \mathbf{l})| \leq C \cdot \left(\frac{2|k - l|}{n(j) - n(i)} + \sqrt{n(j) - n(i)} \right).$$

To ensure $|\mathfrak{R}(\mathbf{v}_k, \mathbf{v}_l, \mathbf{l})| < \epsilon$, we require:

$$C \cdot \left(2 \frac{|k - l|}{n(j) - n(i)} + \sqrt{n(j) - n(i)} \right) < \epsilon.$$

Here, we consider these two terms respectively:

$$2C \frac{|k - l|}{n(j) - n(i)} < \frac{\epsilon}{2}, \quad C \sqrt{n(j) - n(i)} < \frac{\epsilon}{2},$$

which gives:

$$|k - l| < \delta_1 = \frac{\epsilon \cdot (n(j) - n(i))}{4C}, \quad n(j) - n(i) < \left(\frac{\epsilon}{2C} \right)^2.$$

Combining these conditions, we choose:

$$\delta = \min \left(\frac{\epsilon \cdot (n(j) - n(i))}{4C}, \frac{\epsilon^2}{4C^2} \right).$$

Final Conclusion. For any given $\epsilon > 0$, setting $\delta = \min \left(\frac{\epsilon \cdot (n(j) - n(i))}{4C}, \frac{\epsilon^2}{4C^2} \right)$ ensures:

$$\|\mathbf{v}_k - \mathbf{v}_l\|_2 < \delta \quad \Rightarrow \quad |\mathfrak{R}(\mathbf{v}_k, \mathbf{v}_l, \mathbf{l})| < \epsilon.$$

□

C.3 Proofs of Theorem 3

From the definition of the semantic alignment score, we have:

$$\mathfrak{R}(\mathbf{v}_i, \mathbf{v}_j, \mathbf{l}) = -|\text{sim}(\mathbf{v}_i, \mathbf{l}) - \text{sim}(\mathbf{v}_j, \mathbf{l})|, \quad \mathfrak{R}(\mathbf{v}_i, \mathbf{v}_j, \mathbf{l}') = -|\text{sim}(\mathbf{v}_i, \mathbf{l}') - \text{sim}(\mathbf{v}_j, \mathbf{l}')|.$$

The difference in scores can be bounded using the reverse triangle inequality:

$$|\mathfrak{R}(\mathbf{v}_i, \mathbf{v}_j, \mathbf{l}') - \mathfrak{R}(\mathbf{v}_i, \mathbf{v}_j, \mathbf{l})| \leq |(\text{sim}(\mathbf{v}_i, \mathbf{l}') - \text{sim}(\mathbf{v}_j, \mathbf{l}')) - (\text{sim}(\mathbf{v}_i, \mathbf{l}) - \text{sim}(\mathbf{v}_j, \mathbf{l}))|.$$

Simplifying the inequalities above, it gives us:

$$|\mathfrak{R}(\mathbf{v}_i, \mathbf{v}_j, \mathbf{l}') - \mathfrak{R}(\mathbf{v}_i, \mathbf{v}_j, \mathbf{l})| \leq |\text{sim}(\mathbf{v}_i, \mathbf{l}') - \text{sim}(\mathbf{v}_i, \mathbf{l})| + |\text{sim}(\mathbf{v}_j, \mathbf{l}') - \text{sim}(\mathbf{v}_j, \mathbf{l})|.$$

By the Lipschitz continuity of sim, we have: for some constant $C > 0$,

$$|\text{sim}(\mathbf{v}_i, \mathbf{l}') - \text{sim}(\mathbf{v}_i, \mathbf{l})| \leq C \|\mathbf{l}' - \mathbf{l}\|_2, \quad |\text{sim}(\mathbf{v}_j, \mathbf{l}') - \text{sim}(\mathbf{v}_j, \mathbf{l})| \leq C \|\mathbf{l}' - \mathbf{l}\|_2.$$

Substituting these bounds and considering $\|\mathbf{l}' - \mathbf{l}\|_2 \leq \delta_l$

$$|\mathfrak{R}(\mathbf{v}_i, \mathbf{v}_j, \mathbf{l}') - \mathfrak{R}(\mathbf{v}_i, \mathbf{v}_j, \mathbf{l})| \leq 2C \|\mathbf{l}' - \mathbf{l}\|_2 \leq 2C \delta_l. \quad (30)$$

□

D Broader Impacts

We introduce Action Temporal Coherence Learning (AcTOL), a vision-language pretraining framework aimed at improving the generalization capabilities of embodied agents in a variety of manipulation tasks. By learning from large-scale human action videos, AcTOL helps agents acquire temporally consistent representations aligned with natural language, which can support more flexible and data-efficient robotic learning. However, some potential risks should be acknowledged. If AcTOL is trained on video data that contains societal biases or stereotypes, those patterns may be reflected in the model’s behavior. For instance, if certain groups or actions are underrepresented or portrayed inaccurately, the resulting agents could behave in ways that are inappropriate or unreliable in diverse real-world settings. While these challenges are common across many data-driven systems in robotics and vision-language learning, we believe future work should explore strategies such as dataset auditing, fairness-aware training, and improved transparency to support more responsible and robust deployment.