OPAL: Encoding Causal Understanding of Physical Systems for Robot Learning

Daniel Tcheurekdjian, Avery Stevens, Joshua Klasmeier, Christopher McCann, Tom Cooney and Tyler Fenstermaker

Apiary Systems and Aleph Innovations

Introduction

Abstract

Robotic control in unstructured environments remains a significant challenge in artificial intelligence. Recent advances in vision-language models have enabled more flexible multi-modal understanding, but generating coherent action sequences for physical tasks remains difficult.

Previous approaches including Octo, OpenVLA, and $\pi 0$ have made progress in this domain, but struggle with maintaining long-horizon coherence and physical consistency.

We introduce OPAL, a transformer-based architecture that addresses these limitations through a principled approach to action generation based on topological quantum field theory.

Our key insight is that complex action sequences exhibit topological structure analogous to string-net models in condensed matter physics, where local constraints determine global behavior. By incorporating these constraints into our flow matching framework, we achieve more coherent and physically plausible action sequences.

The theoretical guarantees provided by our topological approach result in more coherent long-horizon action sequences without requiring task-specific fine-tuning.

Problem

Let $o_t = [I_t^1, ..., I_t^n, \ell_t, q_t]$ denote a multimodal observation at time t, where I_t^i represents the i-th RGB image (typically from base, left wrist, and right wrist cameras), ℓ_t is a language command, and q_t encodes proprioceptive state information.

Our objective is to model the conditional distribution $p(A_t|o_t)$, where $A_t = [a_t, a_{t+1}, ..., a_{t+H-1}]$ represents a sequence of future actions over horizon $H = 50$. We structure this sequence hierarchically as:

$$A_t = [P_t^1, P_t^2, ..., P_t^K] \tag{1}$$
$$\tag{2}$$

Where each primitive $P_t^k = [a_t^{k,1}, a_t^{k,2}, ..., a_t^{k,m}]$ contains m detailed actions, maintaining $H = K \cdot m = 50$.

Similarly to $\pi 0$ OPAL employs a PaliGemma vision-language backbone with Gemma transformer variants (2B/300M parameters). Visual inputs are encoded using SigLIP ("So400m/14"), with the architecture defined as:

$$\text{Vis}(I_t^i) = \text{SigLIP}(I_t^i) \in \mathbb{R}^{n_v \times d} \tag{3}$$

Language commands ℓ_t are processed through token embeddings:

$$\text{Lang}(\ell_t) = \text{Embedder}(\ell_t) \in \mathbb{R}^{n_\ell \times d} \tag{4}$$

Proprioceptive state vectors are projected to match embedding dimensions:

$$\text{State}(q_t) = \text{Linear}(q_t) \in \mathbb{R}^d \tag{5}$$

Topological Action Modeling

Flow Matching Extension

We extend the flow matching framework first implemented by *Zhilinsky et al.* by introducing topological constraints. Given a trajectory from noise to data distribution:

$$q(A_t^\tau | A_t) = \mathcal{N}(A_t^\tau; \tau A_t, (1 - \tau^2) I)$$

We train a vector field $v_\theta(A_t^\tau, o_t)$ to match the optimal transport direction $u(A_t^\tau | A_t)$:

$$L_\tau(\theta) = \mathbb{E}_{p(A_t|o_t), q(A_t^\tau|A_t)} \| v_\theta(A_t^\tau, o_t) - u(A_t^\tau | A_t) \|_T^2$$

Where $\| \cdot \|_T^2$ is a norm that respects the topological structure of the action space, encoding invariances present in the task domain. This is implemented as:

$$\|v\|_T^2 = v^T M_{\text{topo}} v$$

The matrix M_{topo} encodes the topological constraints of the action space.

Topological Attention

We define our attention mechanism as a modified self-attention operation with topology-preserving constraints:

$$\text{Att}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}} \cdot M_{\text{topo}} \right) V$$

Where M_{topo} is a topological mask derived from fusion categories:

$$M_{\text{topo}}(i, j) = \sum_k F_k^{ij} \cdot \delta(C(i, j, k))$$

Here, F_k^{ij} represents the fusion coefficients between action tokens i and j into channel k, and $C(i, j, k)$ enforces consistency conditions.
The attention mechanism operates at three distinct levels:

1. **Local fusion rules** govern interactions within primitive blocks:

$$\sum_c N_c^{ab} = 1 \quad \forall a, b \in P_t^k$$

This ensures that local action sequences maintain physical consistency.

2. **Non-local fusion channels** enable long-range dependencies with topological protection:

$$\text{Inv}(P_t^i \otimes P_t^j) = \text{Inv}(P_t^i) \cdot \text{Inv}(P_t^j) \cdot \Omega(i, j)$$

The coupling term $\Omega(i, j)$ is learned during training but constrained to satisfy braiding relations.

3. **Invariant subspaces** in the attention mechanism correspond to anyonic excitations:

$$\Pi_a = \sum_\alpha |\psi_a^\alpha\rangle\langle\psi_a^\alpha|$$

These projections ensure that the attention mechanism respects the topological sectors of the action space.

Neural Architecture and Outputs

Results

We evaluated OPAL against baseline approaches (Octo, OpenVLA, and pi0) on 10 complex robotic manipulation tasks. The table below presents the Average Task Progress (ATP) across all benchmark tasks:

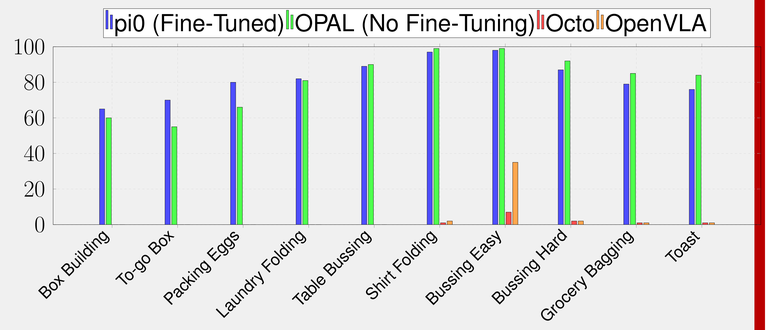

Figure: Performance comparison of vision-language-action models across 10 robotic manipulation tasks. OPAL achieves comparable performance to fine-tuned pi0 models without requiring task-specific optimization, while both significantly outperform previous approaches.

Conclusion

OPAL represents a significant advancement in vision-language-action architectures for robotic control. By introducing topological attention we achieve:

- More coherent and physically plausible action sequences
- Strong zero-shot performance without task-specific fine-tuning
- Enhanced computational efficiency through hierarchical representation and improved integration techniques
- Superior robustness to perturbations in input conditions

Future work will explore further connections between topological quantum field theory and robot learning, including more advanced categorical structures and multi-agent systems with non-trivial braiding statistics.

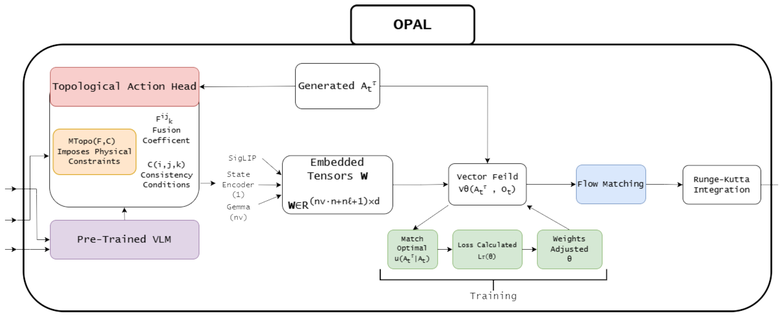

Figure: The OPAL Transformer