Integrated Learning and Planning 16-832

Imitating Task and Motion Planning with Visuomotor Transformers

CORL 23'

- **Soumojit Bhattacharya**

Offline Pretrained TAMP Imitation System

The Core Problem Statement

The Goal

Training generalist robots for complex, real-world manipulation tasks at scale.

The Bottleneck

Deep learning thrives on massive scale, but robotics faces a severe data drought.

The Flaw

Human supervision (teleoperation / kinesthetic teaching) is labor-intensive, costly, and scales poorly — e.g., RT-1 required 1.5 years of data collection.

Key Constraint: RT-1 needed 130,000+ episodes and 1.5 years of human teleoperation to collect training data.

Overview of Related Works

Offline Behavior Cloning

BOTTLENECK

Standard imitation learning approach — bottlenecked by the expense and scalability of human data collection.

Transformers for Robot Control

PRECISION LOSS

Prior work (e.g., RT-1) forces continuous physical space into discrete text-like tokens (bins/voxels). This introduces significant precision loss and limits execution speed.

Task and Motion Planning (TAMP)

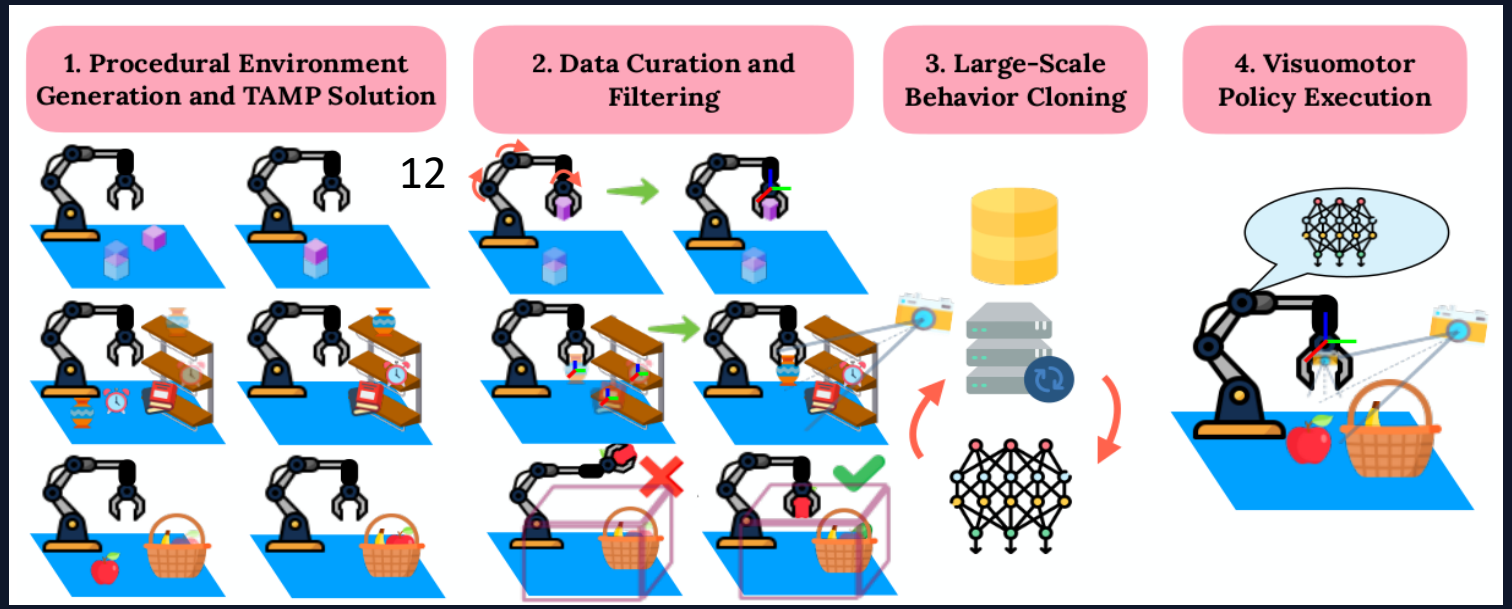
CANNOT DEPLOY

Blends discrete symbolic logic with continuous physics solvers. Powerful in simulation, but requires perfect 3D geometric coordinates — impossible for a real robot in the wild.

SOLUTION

Introducing OPTIMUS

Offline Pretrained TAMP Imitation System



Intuition 1 - TAMP as a Teacher, Not the Brain

✓ USE AT TRAINING

Why TAMP is the Perfect Data Generator

- Acts as an automated, tireless expert in simulation
- Uses privileged information — exact 3D coordinates, perfect scene geometry
- Generates infinite, geometrically flawless demonstrations autonomously
- Zero human labor required beyond initial environment setup

✗ CANNOT USE AT INFERENCE

Why TAMP Fails on a Real Robot

- Real robots have only raw camera pixels — no magical 3D coordinate feed
- Computationally heavy: ~150 ms per action step
- Too slow to react to dynamic, unstructured environments
- Requires a perfect geometric world model that doesn't exist in the wild

Intuition 1.5 - The TAMP Pipeline Under the Hood

TASK PLANNING

1

Step 1. Task Planning — PDDLStream

Uses PDDLStream to calculate the discrete logical sequence of actions (e.g., move → pick → move → place).

INVERSE KINEMATICS

2

Step 2. Inverse Kinematics (IK) — TRAC-IK

Uses TRAC-IK to calculate the specific joint angles required for the grasp. The authors intentionally restrict IK to find solutions closest to the initial configuration to **make the data more consistent** for learning

MOTION PLANNING

3

Step 3. Motion Planning — BiRRT

Uses BiRRT to calculate the continuous, collision-free path.

SMOOTHING

4

Step 4. Smoothing — Cubic Spline Short-Cutting

Applies cubic spline short-cutting to remove jaggedness, resulting in a smooth, locally time-optimal trajectory for the network to imitate.

Intuition 2 - Filtering the TAMP Dataset

01

The Problem

TAMP relies on random sampling (BiRRT). This produces awkward, jagged, or unnecessarily long physical paths.

02

The Danger

Training on outlier trajectories causes compounding errors — small initial mistakes snowball into catastrophic task failure.

03

The Fix

OPTIMUS aggressively distills the dataset:

- Duration constraint: delete excessively long trajectories
- Containment constraint: delete trajectories that exit the optimal workspace

Intuition 3 - Policy Modeling: Continuous Actions

PRIOR APPROACH (e.g. RT-1)

Forced Discretization

- Continuous 3D space → discrete text-like "buckets"
- Enables use of language model training pipelines
- Causes massive precision loss — position accuracy limited by bin resolution
- Joint-space control requires predicting 7 redundant, nonlinear joint angles

OPTIMUS APPROACH

Continuous Task-Space Control

- Outputs exact, continuous 3D coordinates — no discretization
- Task-space: map 2D pixels → 3D end-effector position
- Mathematically much cleaner than predicting 7 nonlinear joint angles (joint-space)
- Avoids the geometric precision floor imposed by binning

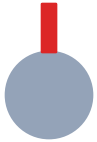
Intuition 4 - Handling Multi-Modal Data

TAMP-generated data is highly multi-modal: the same task may be solved in multiple, equally valid ways.

Ground Truth: Two Valid Modes

Each mode independently yields a successful grasp.

MSE Failure: The Average is Wrong



MSE minimizes squared error by outputting the average of all training modes — a compromised middle path that misses the cup entirely.

Result with MSE loss: only 66% success — a full 20 percentage points below OPTIMUS's GMM-based approach.

MATHEMATICAL SOLUTION

Intuition 5 - Gaussian Mixture Models (GMM)

$$P(a \mid s) = \Sigma_k \; w_k \cdot \mathcal{N}(a ; \mu_k , \Sigma_k)$$

where $k = 1 \ldots 5$ mixture components

w_k

Mixture Weights

Probability each component is responsible for the action. Learned, sum to 1.

μ_k

Gaussian Means

The center of each mode — i.e., the preferred action trajectory for that component.

Σ_k

Covariance Matrices

Uncertainty / spread around each mode. Wide Σ = fuzzy, narrow Σ = confident.

"Going left is highly probable. Going right is highly probable. The middle is not."

The OPTIMUS Architecture

ARCHITECTURE DIAGRAM

TRAINING OBJECTIVE

The OPTIMUS Architecture

ARCHITECTURE DIAGRAM

TRAINING OBJECTIVE

Training Loss (Negative Log-Likelihood)

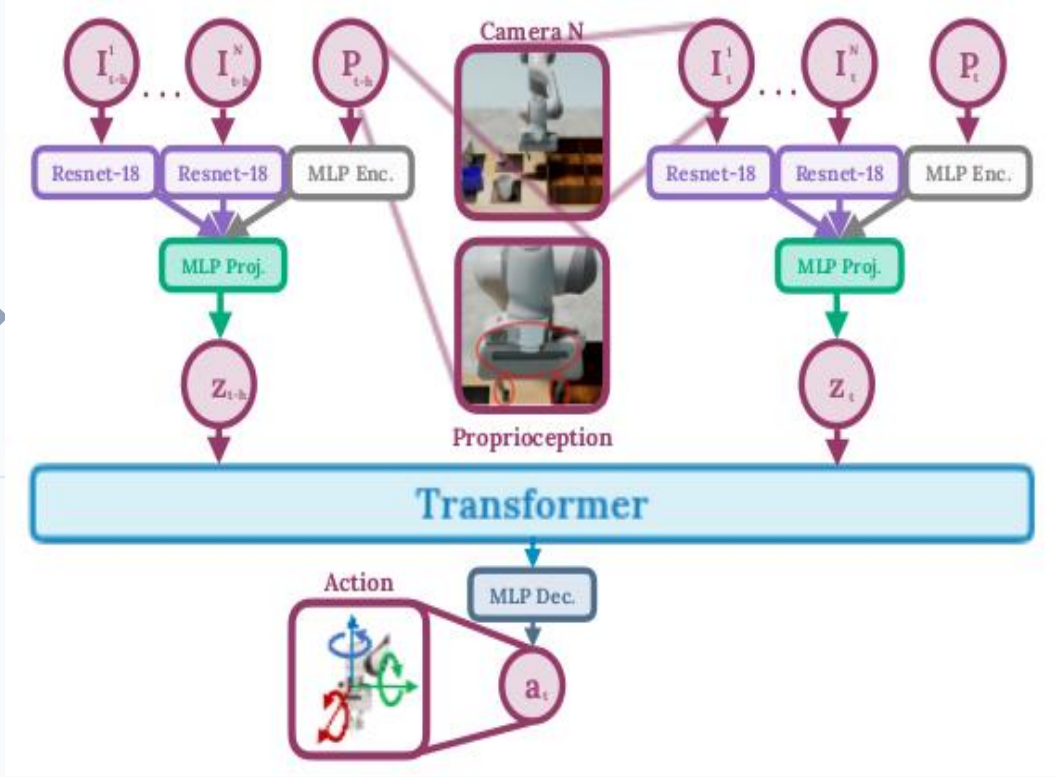
$$L = -\log P(a^* \mid s)$$

Maximize likelihood of expert action a^* given state s
Trained via imitation

The OPTIMUS Architecture

INPUTS

Multi-view cameras
+ proprioception
(joint states)



OUTPUT

Continuous 3D
task-space
coordinates

ARCHITECTURE DIAGRAM

Key Architectural Design decisions

SPATIAL SOFTMAX

2D spatial
keypoint coords
Vision + geometry

TEMPORAL HISTORY

Sliding window
of observations
Motion + intent

GMM HEAD

Multi-modal action
distribution
No mode collapse

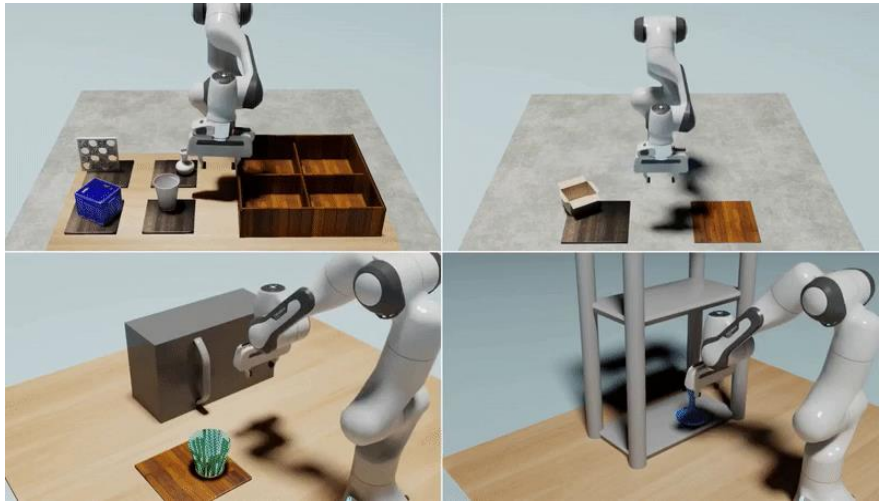
DEEP DIVE: SPATIAL SOFTMAX

Vision encoder O/P passed through a spatial softmax layer to generate 64 "heatmaps"

Take the expected mean for each heatmap to generate 64 "keypoints"

This is passed into the transformer

Results - Tasks



Tasks

Stacking, Pick and Place, Shelf manipulation, Opening a microwave door

Variations

Multi object
Multitask

Dynamic & Zero-Shot scene Adaptation

Alter task plans based on initial state

- Obstacle kept in front of microwave door
- Object to be stacked kept away from base position

Result 1 - Scaling Up Complexity: Comparing with RL

RL Completely Fails

Stack Task (Visual)

- **SAC: 0% & DrQ-v2: 6% & MoDem: 3%**
- **OPTIMUS: 100%**
- Multi-step pick-and-place (4 objects): SAC, DrQ-v2, MoDem all at 0% — OPTIMUS: 90%
- Standard RL completely fails on long-horizon visual tasks due to exploration limits

Long-Horizon Tasks

Stacking & Pick-Place Scaling

- **Stacks 5 blocks: 70% success**
- **Pick-place 4 objects: 60% success**
- RL baselines cannot even begin to solve multi-step tasks requiring 10+ sequential decisions

OPTIMUS achieves 100% on Stack and 90% on PickPlaceFour — RL baselines score 0–6% on these same tasks.

Result 2 - Multi-Task Scaling (19 to 72 Objects)

19-Object Tasks

OPTIMUS vs. BC-MLP / BC-RNN

- PickPlace: **OPTIMUS 85%** vs. BC-MLP 61%
- Shelf: **OPTIMUS 66%** vs. BC-MLP 48%
- Microwave: **OPTIMUS 61%** vs. BC-RNN 41%

72-Object Extreme Scale

Single unified model, all tasks

- PickPlace 75%, Shelf 48%, Microwave 47%
- Prior baselines drop to 29–50% in extreme multi-object scenarios
- **Multi-task unified model: 73% combined success**
- No hard-coded object IDs — scales to unseen objects without retraining

A single OPTIMUS model trained on PickPlace + Shelf + Microwave simultaneously achieves 73% combined success rate.

Result 3 - OPTIMUS vs. Other Architectures

vs. Guided TAMP

Robosuite PickPlace Task

- **OPTIMUS: 90%** vs. Guided TAMP: 88%
- Guided TAMP relies on privileged state information (exact 3D coordinates)
- OPTIMUS achieves this from pixels only — no privileged state access

vs. Transformer-in-Transformer (TiT)

Avg. +16.8% across 5 tasks

- Microwave: **OPTIMUS 86%** vs. TiT 67%
- PickPlaceFour: **OPTIMUS 60%** vs. TiT 45%
- Consistent improvement across all 5 benchmark tasks — +16.8% average margin

OPTIMUS surpasses Guided TAMP (from pixels, no state access) and beats TiT by an average of 16.8% across 5 tasks.

Intuition 6 - Comparision to Guided TAMP

Prior Imitation Planners
(e.g. Guided TAMP)

- **Rigid symbolic vocabulary: Action = "pick_up_apple"**
- Requires every object to be pre-registered in the planner codebook
- New object not in the code → system crash
- Scales poorly: adding 72 objects requires 72 hardcoded symbols

OPTIMUS
(pixel-based geometry)

- **No symbolic vocabulary — operates purely on pixels and spatial geometry**
- New geometry → No need to "register"
- Scales to 72 distinct objects with no architectural modification

Guided TAMP	OPTIMUS
88	**90**

This geometric generalization is the key architectural enabler for scaling to diverse manipulation tasks.

Result 4 - The Student Surpasses the Teacher

Success Rate

Student beats teacher on quality

- **OPTIMUS: 87% avg. success rate**
- **TAMP supervisor: 52% avg. success rate**
- OPTIMUS trains only on flawless trajectories — TAMP's occasional physics glitches (slipping grasps, IK failures) are filtered out

Inference Speed

Student beats teacher on speed

- **30M model: 21ms per action**
- **100M model: 31ms per action**
- TAMP: ~150ms per action — 5x to 7.5x slower than the trained policy

OPTIMUS: 87% vs. TAMP 52% success, and 5–7.5x faster at inference (21–31ms vs. 150ms per action).

Result 5 - Generalization to Unseen Receptacle Sizes

Shelf Receptacle

80% success on held-out shelves

- **OPTIMUS: 80% success rate**
- **Best baseline (BeT): 31%**
- OPTIMUS learns to generalize motions to randomized shelf sizes using visual input from TAMP supervision

Microwave Receptacle

70% success on held-out microwaves

- **OPTIMUS: 70% success rate**
- **Best baseline (BeT): 59%**
- OPTIMUS adapts motions to unseen microwave sizes via scene-conditioned generalization from TAMP supervision

OPTIMUS achieves 80% (ShelfReceptacle) and 70% (MicrowaveReceptacle), outperforming the best baseline of 31% and 59%.

Result 6 - Key Takeaways from Ablations

Control Space & Data Quality

Task space, filtration, gripper stalls

- Task-space control: **100%** vs. Joint-space: 86%
- Dual duration + workspace filters: **+10% boost** over unfiltered data
- Reducing gripper stall: 25 steps → 5 steps jumps success **78% → 100%**
- Discrete gripper actions add further **+4% boost**

Loss Function Comparison

GMM handles multi-modal demos

- **GMM output: 86% success**
- MSE loss: 66% success (−20%)
- Gaussian Log-Likelihood: 70% (−16%)
- GMM correctly captures multi-modal demonstrations (e.g., pick from left vs. right)

Every design choice contributes: Task-space control, data filtration, gripper stall reduction, and GMM loss all contribute measurably.

Key Takeaway - Wrist Camera Utilization

🎥 Fixed External Camera

- Mounted far from the robot — wide field of view
- As the arm moves into the workspace, the robot's own body blocks the camera
- Critical lost at the moment of grasp
- Sufficient for simple, large-target tasks (e.g., stacking large blocks)

🤖 Wrist (End-Effector) Camera

- Mounted on the gripper — dynamic, unblocked close-up view
- Critical for fine-grained tasks
- **Necessary: removes occlusion-induced state uncertainty at the policy level**

Ablation result: wrist camera adds little for block-stacking; adds ~18% success for microwave tasks requiring fine spatial precision.

LIMITATIONS & FUTURE WORK

Looking Forward

01

TAMP Dependency

OPTIMUS is fundamentally bounded by what TAMP can solve. Tasks requiring dynamic or contact-rich manipulation — pouring liquids, aggressive sliding — remain out of scope.

02

Multi-Task Ceiling

Success rates degrade as task count and object diversity reach extremes. At 72-object Microwave tasks, success drops to 47% — indicating a remaining distribution gap.

03

Real-World Deployment Path

Future: use external pose estimation to generate real-world TAMP data in a lab setting, distill it into OPTIMUS, then deploy purely on standard RGB cameras in the wild.

OPTIMUS demonstrates that algorithmic planning and neural perception can be systematically fused — without human demonstrations.

The END