OT-Robust3DVLA

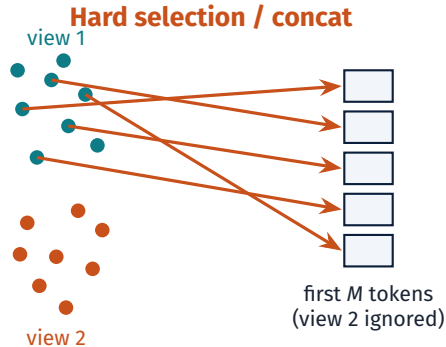
A Wasserstein Barycenter is the Right Inductive Bias for Robust Multi-View 3D Vision-Language-Action Policies

Pawan Kumar
International Institute of Information Technology, Hyderabad, India
optimalstar.org

NeurIPS 2026

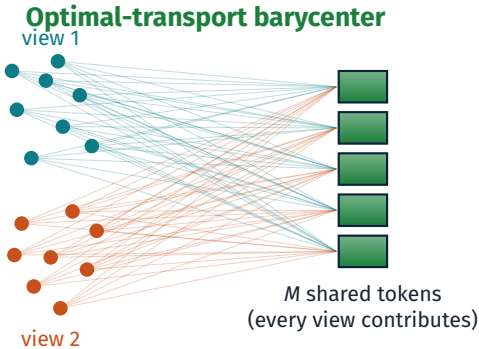
The problem: noisy cameras make robot actions jump

- Robot policies read **several 3D camera views** + a **language command**, then act.
- Real sensors are imperfect: **depth jitter, occlusion, calibration drift, a lost camera**.
- Common fusion **concatenates** tokens or **picks hard samples**.
- A tiny sensor change can swap the picked points
  $\Rightarrow$ **a jump in the action**.



Lose camera 1 and the concat policy has **nothing left** to look at.

Our idea: merge the views by optimal transport

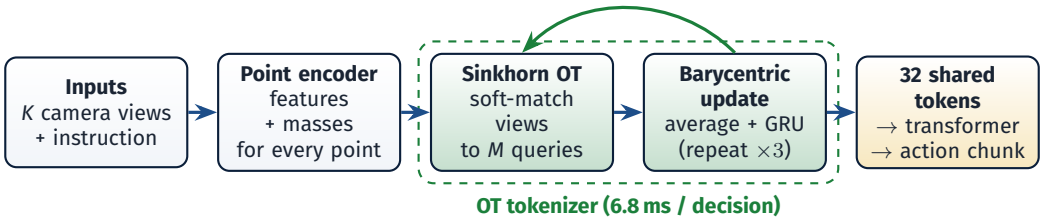


- Each camera view = a cloud of weighted feature points μ_k.
- Find M **shared tokens** β that are "closest on average" to *all* views:

$$\beta^\star = \arg\min_{\beta} \sum_k \alpha_k \, \mathcal{S}_\varepsilon(\beta, \mu_k)$$

Soft couplings move mass **smoothly**: small sensor change $\Rightarrow$ small token change.

How it works, and why it is stable



Theory (under stated assumptions): a sensor error of size ρ grows only linearly,

$$\underbrace{\rho}_{\text{sensor}} \xrightarrow{L_B} \underbrace{L_B\rho + \tau_B}_{\text{tokens}} \xrightarrow{L_A} \underbrace{L_A L_B \rho + L_A \tau_B + \tau_A}_{\text{action distribution}}$$

Keep a task-cost margin $m \geq L_J R_A$ and the expected cost stays within budget for every perturbation in the set.

Key results

46×
more action shift if the OT tokenizer is removed

10.2×
same removal on RLBench (30k updates)

92.1%
worst-case closed-loop success (concat: 24.2%)

49/50
LIBERO-Spatial task successes (FPS: 30/50)

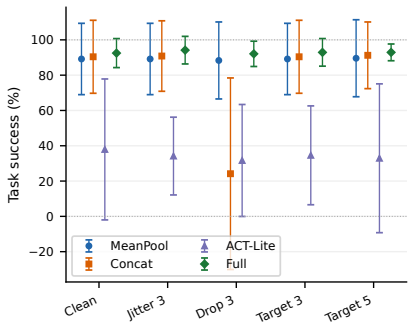
Lowest clean action error on all 4 benchmarks

mean ℓ_2 over 3 seeds, lower is better

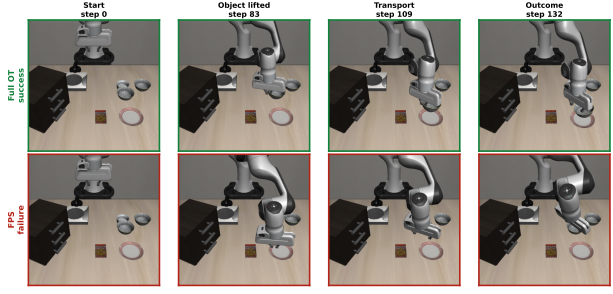
	Synth.	LIBERO	RLBench	CALVIN
MeanPool	0.057	1.243	1.410	2.094
Concat	0.055	2.790	2.192	2.750
ACT-Lite	0.079	2.559	2.235	3.052
Ours	**0.047**	**1.216**	**1.107**	**1.802**

Removing the tokenizer is the **biggest** drop we measured; every other removal changes shift by $< 1.5\times$.

Does the robot still finish the task?



Synthetic pick task, 80 starts × 3 seeds:
concat collapses when a camera is dropped.



LIBERO-Spatial (of 50)	Clean	Jitter	Calib. drift
Ours (OT)	**49**	**42**	**45**
FPS fusion	30	25	17
Concat	0	–	–

Take-home message

- **Fuse camera views by transport, not by picking.**
- Small sensor error $\Rightarrow$ small action change, in theory and in experiments.
- Most accurate *and* most stable fusion we tested, on 4 benchmarks.
- **Scope:** compact policies; plugging into large pretrained VLAs and real-robot safety tests is future work.

Paper

Code

Thank you!

pk.iiith@gmail.com

optimalstar.org

openreview.net/forum?id=tFugZPy5w4

github.com/misterpawan/ot-3dvla-paper