DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

University of Padua
Department of Information Engineering
Bachelor's Degree in Automation and Systems Engineering

tactile-to-visual representation learning for vision-language-action models

Supervisor: Pietro Falco
Student: Nicolò Girardello
Student's ID: 2102681
Padua, 20/07/2026

Context and Motivation

- Vision-Language-Action (VLA) models rely almost entirely on vision to perceive the environment
- Visual input can be degraded or absent: occlusions, poor lighting, sensor failure
- Tactile sensing offers a complementary, contact-based signal that is robust to these failure modes
- **Idea:** substitute degraded vision with a tactile signal projected into the same latent space

Background: What is a VLA Model?

- **Vision-Language-Action (VLA)** models are generalist robotic policies that map visual observations and natural-language instructions directly to robot actions
- Built on large pretrained vision-language backbones, exposed to large-scale robot demonstration data
- A vision encoder produces a latent embedding from the camera image; this embedding, together with the language instruction, conditions the action-generation process
- Key examples: RT-2, π_0, **OpenVLA** / **OpenVLA-OFT** (the frozen backbone used in this thesis)

Research Objective

Research Question

Can a lightweight, trainable module f_θ project tactile embeddings $z_\mathcal{T}$ into the frozen visual latent space $z_\mathcal{V}$ of a VLA model, allowing tactile input to substitute for degraded or absent vision **without modifying the underlying policy**?

- Target: a frozen, pretrained VLA policy
- Constraint: no fine-tuning of the policy itself
- Scope of this thesis: feasibility, not yet full deployment

Methodology Overview

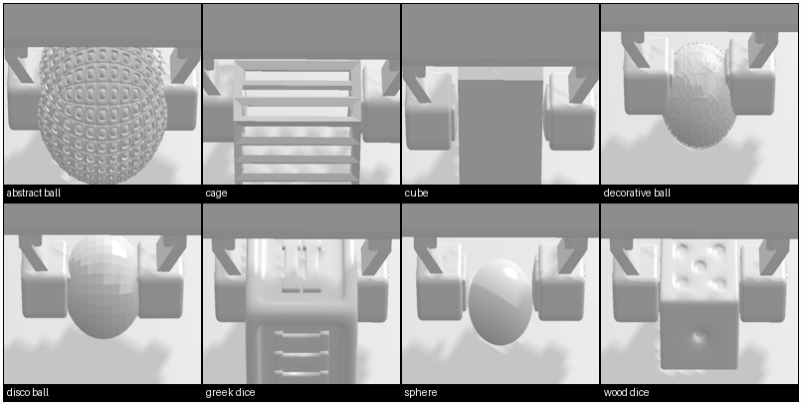
A **feasibility-first** approach, in three stages:

1. **Feasibility analysis** — are tactile and visual embedding spaces compatible enough to justify learning a projection?
2. **Simulation benchmark** — build a physically instrumented, reproducible testbed (DIGIT-equipped gripper in LIBERO/MuJoCo)
3. **Pilot evaluation** — train and evaluate f_θ against a naive tactile-blind baseline

1. Feasibility Analysis

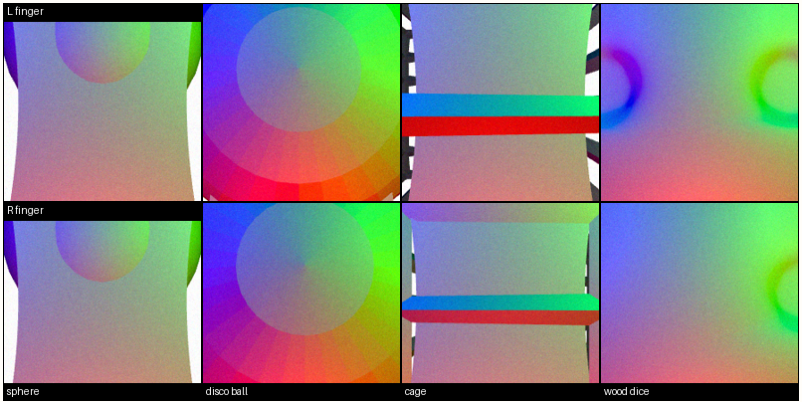
Experimental Setup

- Small, carefully designed exploratory dataset
- **8** object classes, **80** samples total (8 objects $\times$ 10 grasp poses)
- Paired tactile (DIGIT) and visual embeddings extracted for each sample



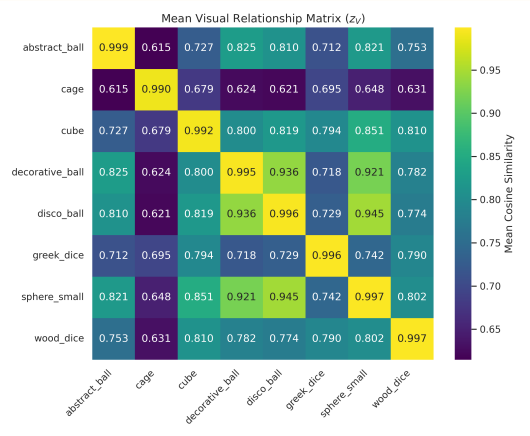
Tactile Signal Characterisation

- Objects vary along two axes: coarse geometry and fine surface texture
- Contact patterns range from near-uniform deformation to sparse, discontinuous contact



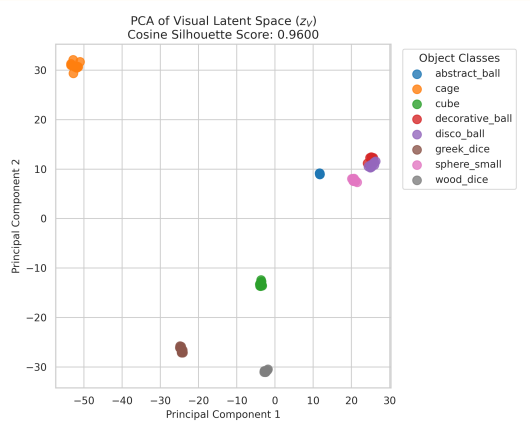
Visual Embedding Structure

- Frozen SigLIP–DINOv2 encoder retains meaningful, object-specific structure
- The open-lattice `cage` object stands out as most dissimilar from the rest



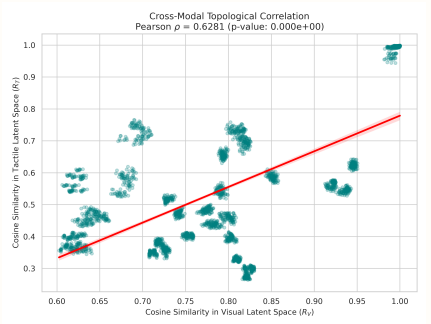
Visual Embedding Structure: PCA View

- First two principal components of $z_{\mathcal{V}}$ across the eight objects
- Tight, well-separated clusters (cosine silhouette score 0.96)

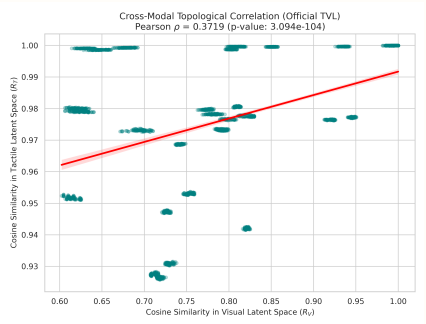


Representational Similarity Analysis (RSA)

- Compares the *relational structure* of two embedding spaces, not raw coordinates
- Tactile similarity (cosine) plotted against visual similarity, for every object pair



(a) Generic ViT-Small encoder



(b) Official TVL encoder

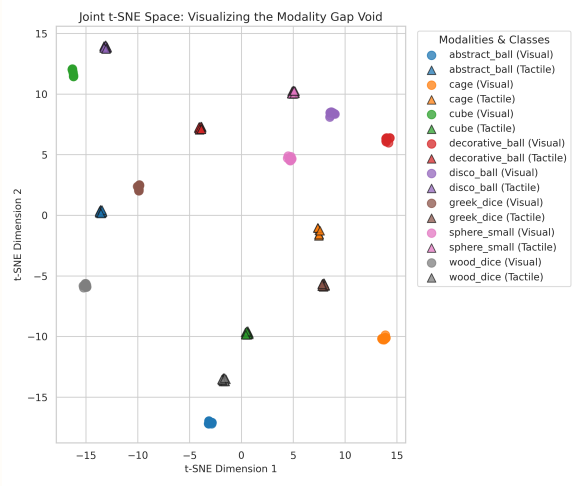
RSA Results: Backbone Comparison

- ViT-Small preserves a visible positive relationship with the visual space
- TVL encoder compresses almost all similarities near 1.0 (anisotropic collapse)

Metric	$z_{\mathcal{V}}$ (target)	$z_{\mathcal{T}}$ (ViT-Small)	$z_{\mathcal{T}}$ (TVL)
Cosine silhouette score	0.9600	0.9749	0.9769
Mean inter-class cosine sim.	0.7634	0.4867	0.9730
RSA correlation (Pearson ρ)	—	**0.628**	0.372

Joint Modality Visualisation

- Shared t-SNE projection of visual and tactile embeddings
- Same-object markers do not overlap (a modality gap exists) but preserve consistent relative layout



Backbone Decision

Decision

Adopt **ViT-Small** (ImageNet-pretrained) as the tactile backbone for all subsequent experiments.

Alternative explanation considered: a sim-to-real domain gap may explain why the TVL encoder, trained on real tactile data, underperforms on TACTO-simulated frames.

- Stated explicitly as an open hypothesis, not a settled fact
- Revisited later as future work

2. Simulation Benchmark

Benchmark Architecture

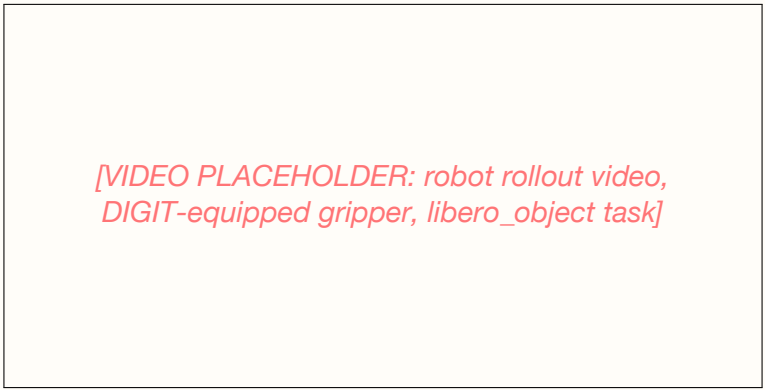
- **LIBERO** (`libero_object` suite) — standardized manipulation task suite
- **robosuite / MuJoCo** — physics simulation backend
- **TACTO** — tactile rendering pipeline (originally validated in a separate PyBullet-based setup)
- **Contribution:** bridging TACTO into a physically different simulator (MuJoCo) and integrating it with LIBERO

DIGIT-Equipped Gripper Integration

- Custom Franka Emika Panda gripper modified to host **two DIGIT** tactile sensors, one per fingertip
- Rigidly attached via fixed joints: no additional degrees of freedom
- Ported from a standalone URDF/PyBullet description to a MJCF model in robosuite's gripper registry
- Enables synchronized tactile + visual data collection during policy rollouts

Simulated Rollouts

- Frozen OpenVLA-OFT policy executed on `libero_object` tasks
- Rollouts recorded with `agentview` visual stream



Key Result: Mechanical Confound

- Success rate on `libero_object` drops sharply after gripper modification

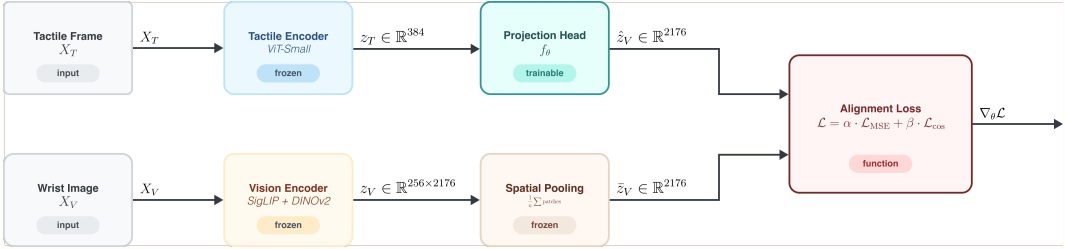
Gripper configuration	Success rate
Standard Panda gripper	98.2%
DIGIT-equipped gripper	56.8%

Cause: identified as a mechanical/kinematic confound from the sensor's physical footprint, independent of the tactile-visual alignment question.

3. Architecture and Pilot Evaluation

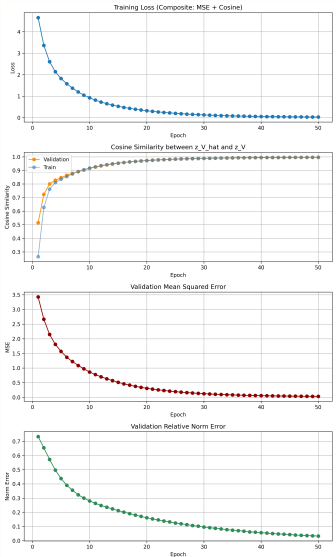
Projection Module f_θ

- Lightweight, trainable module mapping tactile embeddings $z_\mathcal{T}$ into the frozen visual latent space: $f_\theta(z_\mathcal{T}) \approx z_\mathcal{V}$
- Visual encoder and downstream policy remain **entirely frozen**
- Only f_θ (and a small channel adapter) is trained



Training Behaviour

- f_θ trained to regress frozen visual target embeddings from tactile input
- Verified to train in a numerically well-behaved way (stable loss curves, no divergence)

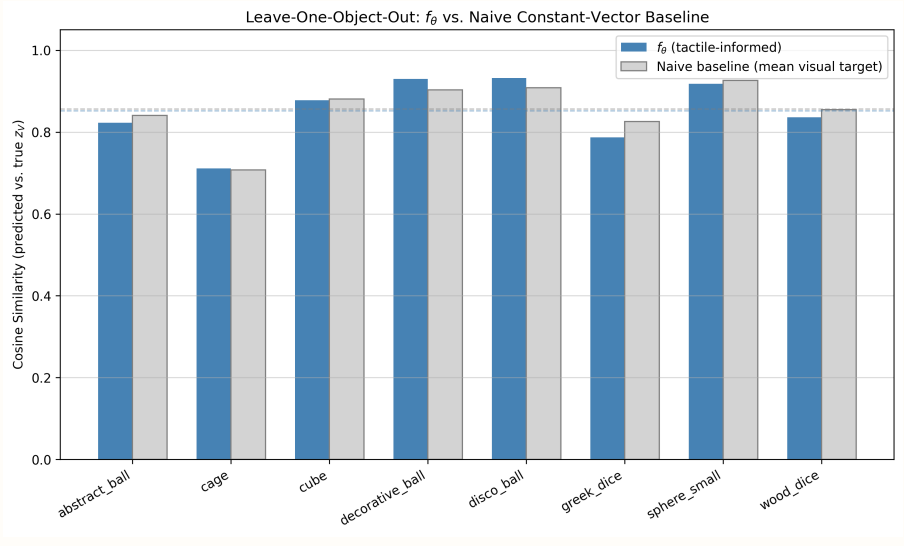


Evaluation Protocol

- **Leave-one-object-out** cross-validation
- Compared against a **naive, tactile-blind baseline** (constant-vector predictor)
- This baseline comparison is **absent from every closely related prior work** reviewed in this thesis

A deliberately rigorous test: does the learned projection add anything beyond a trivial baseline?

Pilot Results: Leave-One-Object-Out



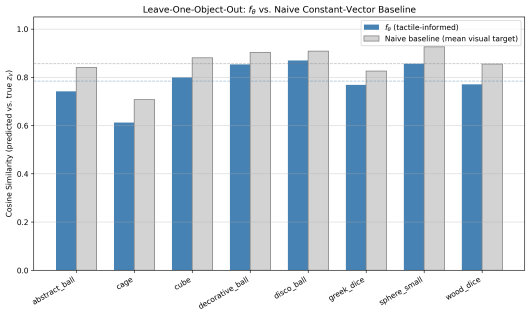
Pilot Results: Numbers

Held-out object	f_θ CosSim	Baseline CosSim	Margin
abstract_ball	0.8228	0.8406	-0.0178
cage	0.7113	0.7072	+0.0040
cube	0.8777	0.8806	-0.0029
decorative_ball	0.9302	0.9032	+0.0270
disco_ball	0.9322	0.9086	+0.0236
greek_dice	0.7868	0.8260	-0.0392
sphere	0.9181	0.9264	-0.0083
wood_dice	0.8360	0.8550	-0.0190
Mean	**0.8519**	**0.8559**	**-0.0041**

Mean cross-object cosine similarity is, on average, statistically indistinguishable from the naive baseline.

Follow-Up: Capacity Reduction

- Additional experiment: reduce the capacity of f_θ ($d_h{=}1024 \rightarrow d_h{=}256$, $15\times$ fewer parameters)
- Goal: distinguish between two explanations for the inconclusive result
  - Insufficient training data (80 samples)
  - Insufficient/mismatched representational capacity



Capacity Reduction: Numbers

Held-out object	f_θ CosSim ($d_h{=}256$)	Baseline	Margin
abstract_ball	0.7420	0.8406	-0.0986
cage	0.6124	0.7072	-0.0948
cube	0.8000	0.8806	-0.0806
decorative_ball	0.8532	0.9032	-0.0500
disco_ball	0.8695	0.9086	-0.0391
greek_dice	0.7685	0.8260	-0.0575
sphere	0.8562	0.9264	-0.0702
wood_dice	0.7704	0.8550	-0.0846
Mean	**0.7840**	**0.8559**	**-0.0719**

Reducing capacity **widened** the gap on every fold: this rules out the simplest capacity-collapse explanation.

Comparison with Prior Work

	TVL	VLA-Touch	Tactile-VLA	This work
Frozen VLA backbone	n/a	✓	partial	✓
Optical tactile images	✓	×	×	✓
Explicit latent substitution	×	×	×	✓
Degrades to base VLA	—	—	—	✓
Robotics downstream task	×	✓	✓	✓
Cross-object validation reported	×	×	×	✓

4. Discussion

Interpreting the Results

- Tactile and visual embeddings occupy **compatible enough** regions of representational space in aggregate (RSA evidence)
- This motivates pursuing a learned projection f_θ
- However, at the data scale available, this compatibility has **not translated** into demonstrated generalization to unseen objects

The Mechanical Confound

- Two **independent** sources of performance degradation identified:
  1. Cross-modal alignment quality (the core research question)
  2. Physical footprint of the DIGIT sensors on the gripper
- Current experimental design does not yet separate these two effects
- Necessary distinction for correctly attributing future results

Open Ambiguity

Unresolved Question

Is the inconclusive generalization result caused by:

- **Insufficient data scale** (80 samples, 8 object classes), or
- **Insufficient representational capacity** of the current tactile backbone/projection design?

The capacity-reduction follow-up narrows, but does not fully resolve, this ambiguity.

Answering the Research Question

Qualified Answer

Tactile and visual embeddings are compatible enough, in aggregate, to justify pursuing a learned projection f_θ — but this has **not yet been shown to generalize** beyond a trivial baseline at the current data scale.

Neither finding invalidates the central hypothesis; both narrow the conditions under which it remains open to confirmation.

Conclusion and Future Work

Summary of Contributions

1. **Feasibility analysis grounded in data**, not assumption (RSA-based backbone decision)
2. **Fully reproducible, physically instrumented benchmark** (DIGIT + LIBERO/robosuite/MuJoCo/TACTO)
3. **Rigorous pilot evaluation** of the core hypothesis, including a baseline comparison missing from prior work
4. **Identification of the remaining open questions**, separating alignment quality from mechanical confound

Limitations

1. **Dataset scale** — 8 object classes, 80 samples; sufficient for a pilot, not for generalization claims
2. **Simulator- and backbone-specificity** — conclusions tied to TACTO-simulated DIGIT frames and ViT-Small; untested on real hardware
3. **Unresolved mechanical confound** — gripper redesign effect not yet isolated from projection effect

Future Work (1/2)

- **Scaling up data collection** within the validated LIBERO/TACTO-MuJoCo pipeline — the most direct lever to disambiguate data scale vs. representational capacity
- **Revisiting the tactile backbone**: fine-tuned TVL encoder on TACTO frames, domain-randomized rendering, or jointly-trained tactile encoder

Future Work (2/2)

- **Isolating the gripper confound** — fine-tune the policy's action head on DIGIT-gripper kinematics alone, without tactile input, to build a corrected baseline
- **Visual degradation protocol** — systematically occlude/degrade `agentview` input to test the substitution mechanism directly
- **Policy-level evaluation** of the full substitution mechanism, once cross-object generalization exceeds baseline at larger scale

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

University of Padua
Department of Information Engineering
Bachelor's Degree in Automation and Systems Engineering