Óbudai Egyetem
OBUDA UNIVERSITY

Beyond Pixels: Structured Representations and Robust Robot Manipulation

Presenter: Bálint Károly Farkas

Supervisors:
Prof. Dr. Péter Galambos
Dr. habil. Károly Széll
Doctoral School: Doctoral School of Applied Informatics and Applied Mathematics

Óbudai Egyetem
OBUDA UNIVERSITY

The Problem

✓ Rigid objects — **robot manipulation works well**

✗ Deformable objects — **still extremely hard to manipulate!**

Examples:

- Cables
- Bags
- Textiles
- Tubes in packaging

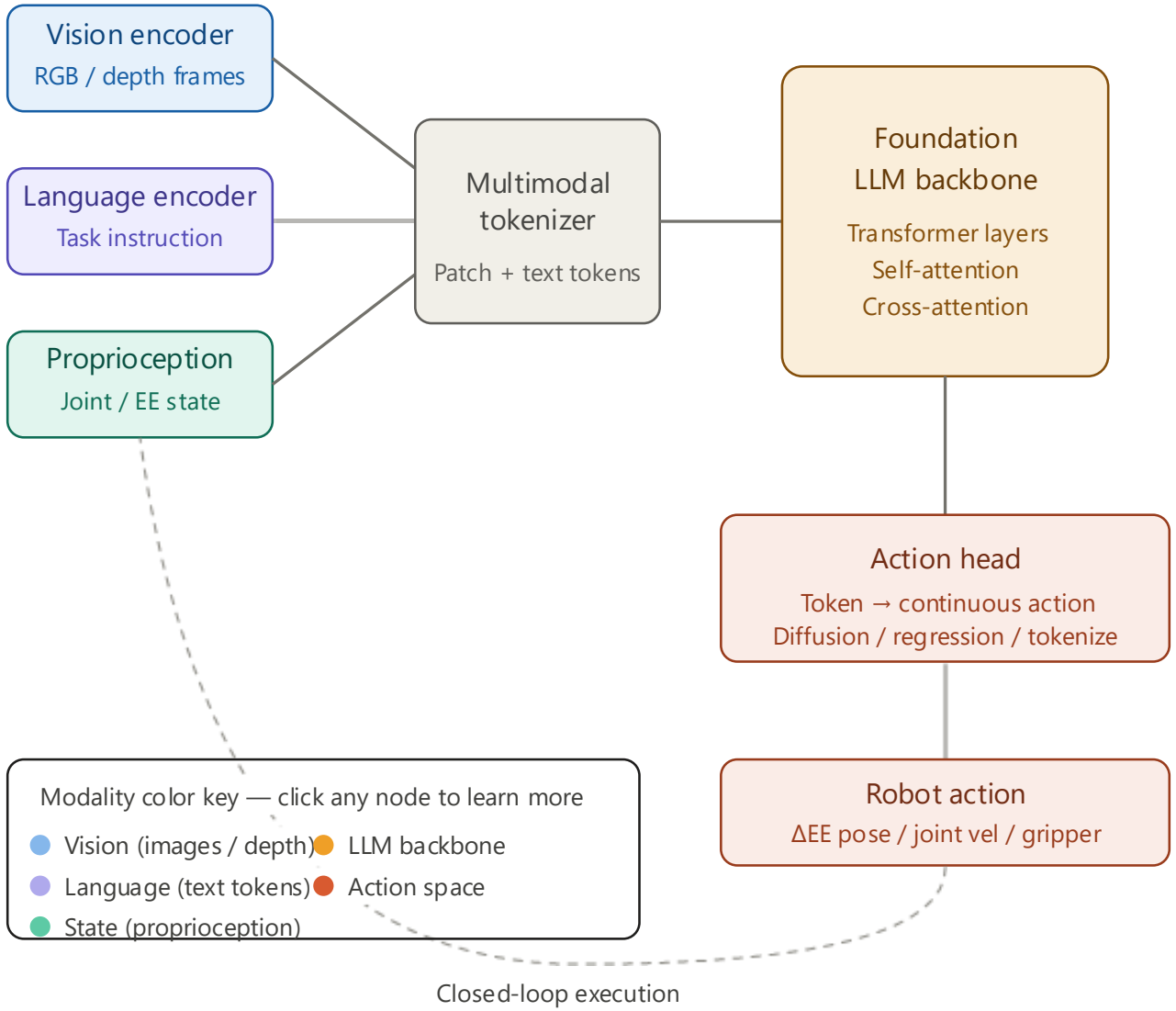
Why is it hard to manipulate?

High-dimensional state
Millions of DOF vs. rigid body's 6

Changing topology
Folding, looping, knotting change the structure

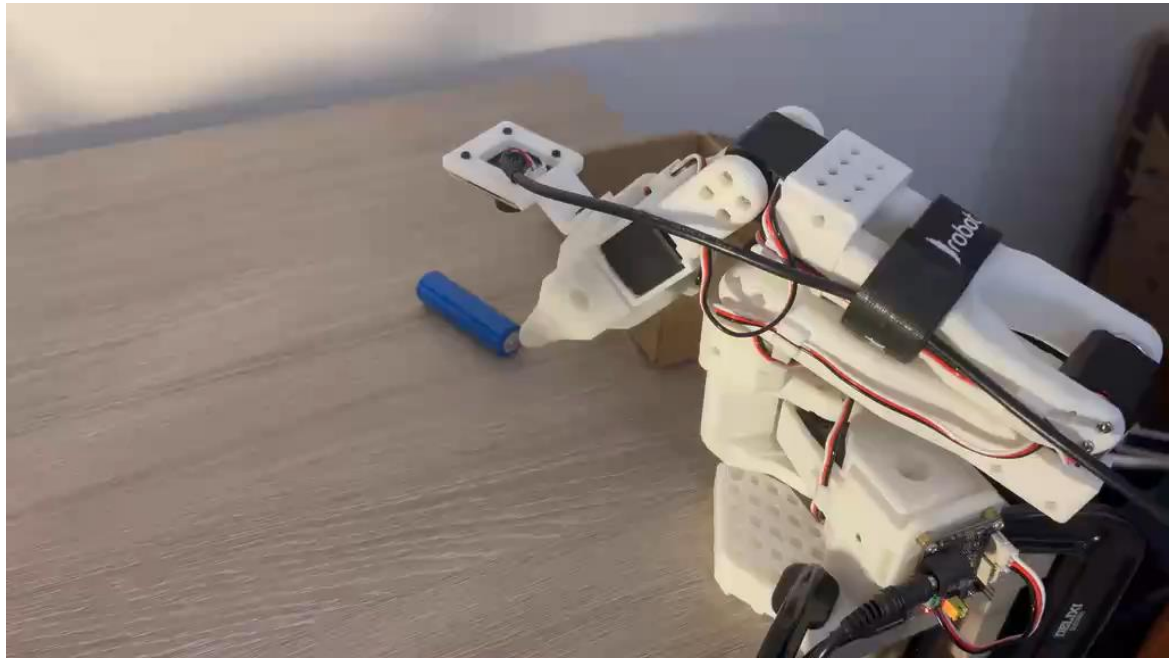
Strong OOD sensitivity
New material or config breaks the policy

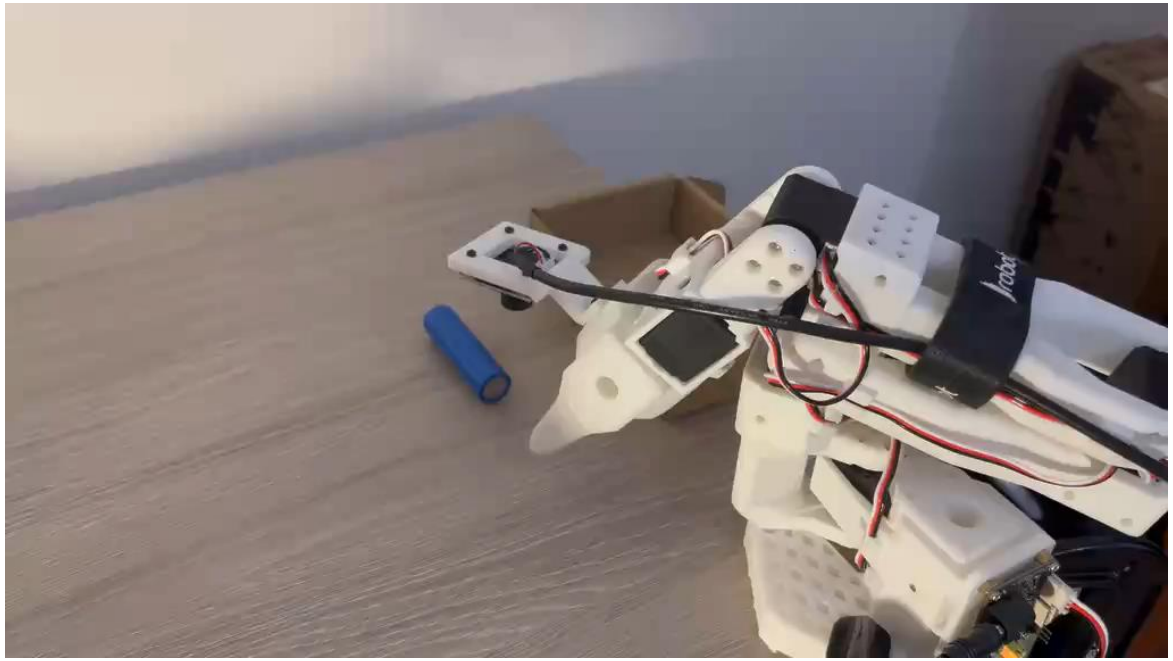
Modern Approach - VLA



Óbudai Egyetem
OBUDA UNIVERSITY

Why robot learning? - SmolVLA on SO-101





SmolVLA & SO-101 with deformable object

- 60% success rate
- The model is not confident

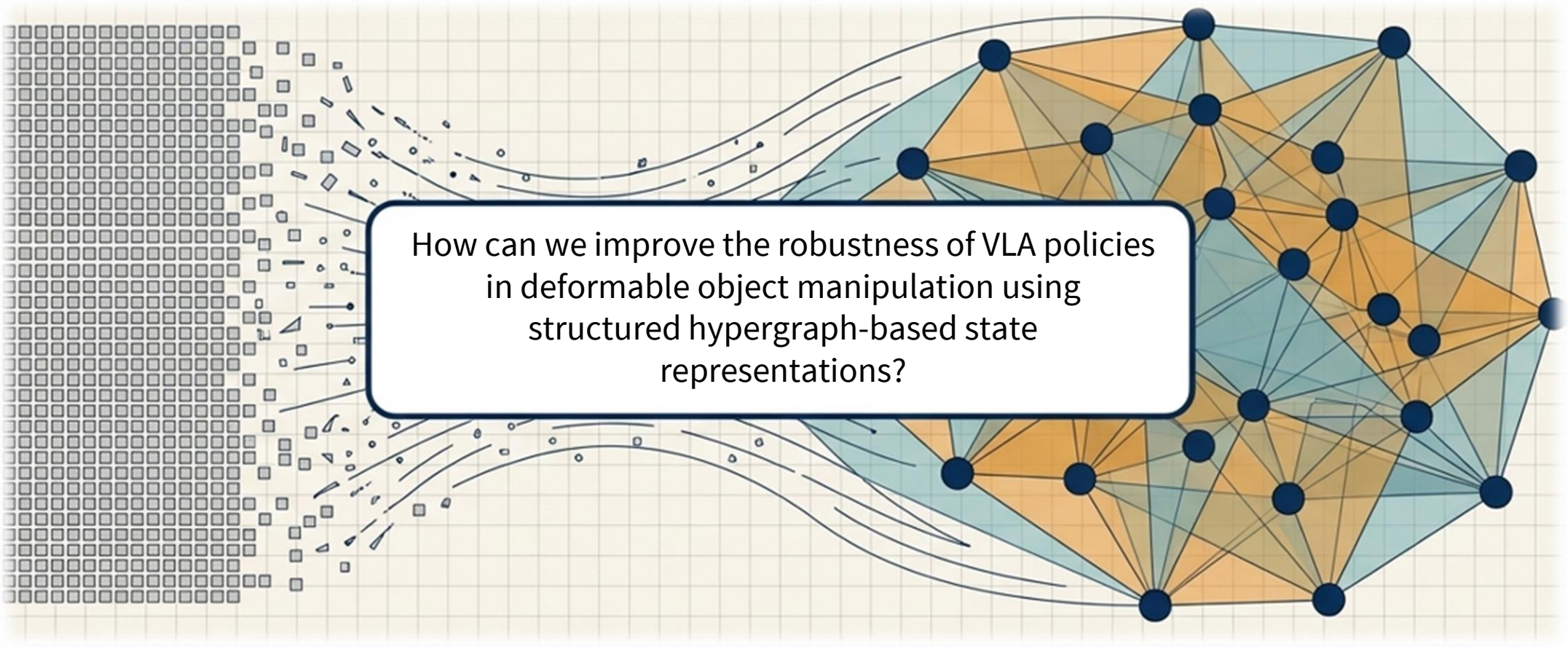
Experiment

Initial prompt – 60% success rate:
„Pick up the glasses cloth and place it on the shelf.”

Improved prompt – 70% success rate:
„[TASK=cloth_shelf] Pick up the glasses cloth and place it on the shelf.”

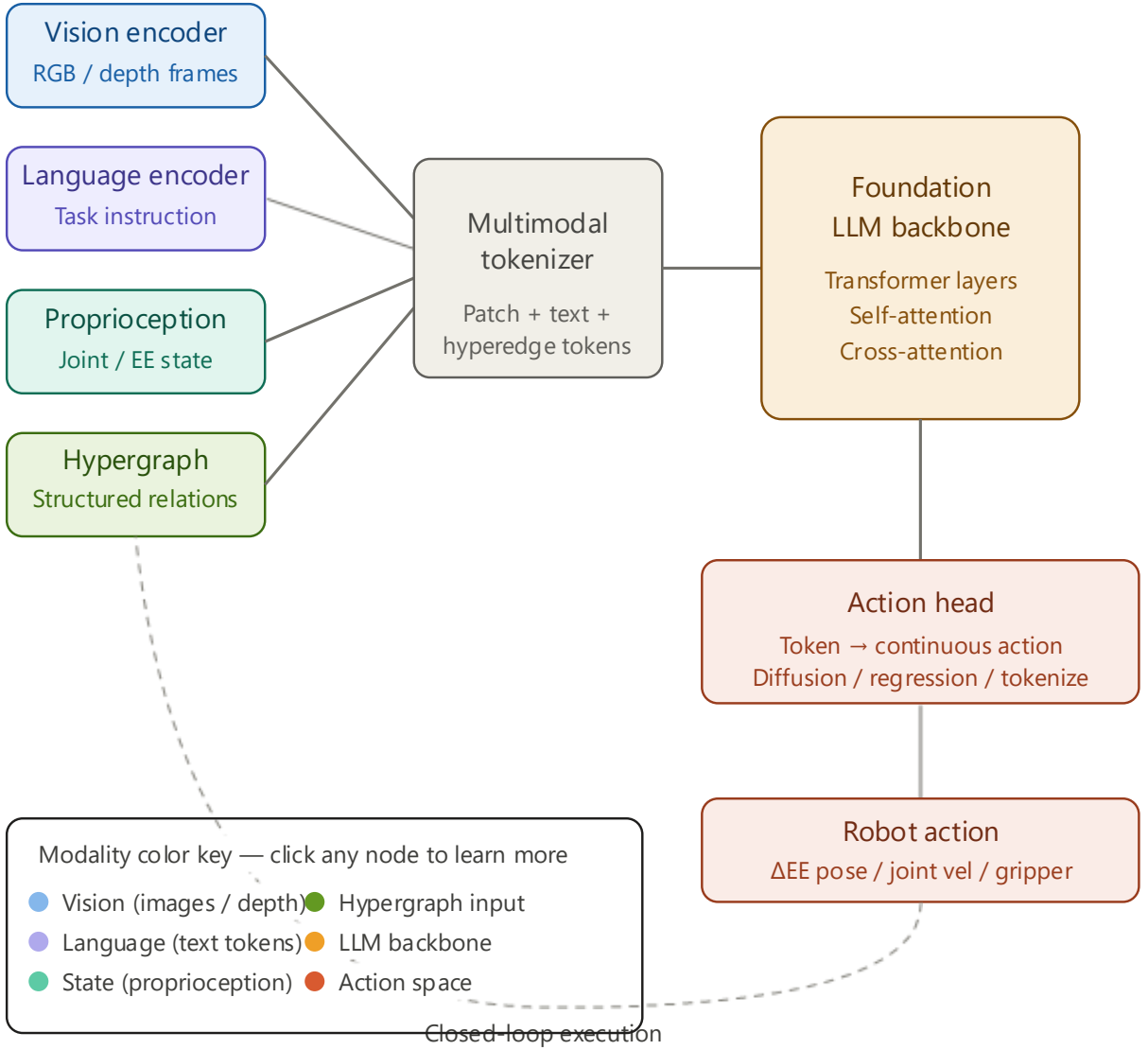
Better success rate with structured input

Key Idea



Pixels → Structured relational representation

Key Idea


Vision encoder
RGB / depth frames
Language encoder
Task instruction
Proprioception
Joint / EE state
Hypergraph
Structured relations
Multimodal tokenizer
Patch + text + hyperedge tokens
Foundation LLM backbone
Transformer layers
Self-attention
Cross-attention
Action head
Token → continuous action
Diffusion / regression / tokenize
Robot action
ΔEE pose / joint vel / gripper
Modality color key — click any node to learn more
Vision (images / depth)
Language (text tokens)
State (proprioception)
Hypergraph input
LLM backbone
Action space
Closed-loop execution

Industrial Relevance

What industry already has

- Operator workstation with multiple sensor modalities
- Step-by-step animations for assembly processes
- Structured process descriptions

What is missing?

✗ This information does not reach the robot policy
✗ VLA sees only raw pixels — not the structure

→ **Hypergraph as a bridge between existing industrial knowledge and robot policies**

Industrial Use Cases

Cable & Tube Handling

- Deformable, topology-changing objects
- Contact & grasp relations as hyperedges
- Robust to initial config variation

Textile & Garment Assembly

- Folding, stacking, draping tasks
- Fold-line & affordance hyperedges
- Material & size OOD robustness

Operator Workstation

- Animations → hypergraph input
- Multi-modal sensors at station
- Existing knowledge reused — no new data collection

Common thread: structured relational knowledge already exists — hypergraph makes it robot-readable

Óbudai Egyetem
OBUDA UNIVERSITY

Future Work

1

Graph / Hypergraph Construction

Keypoint detection
Contact / affordance relations as hyperedges
Sim + real validation

2

VLA Integration

Hyperedge tokens fed into multimodal tokenizer
Frozen backbone
Baselines: B0 pixel-only, B1 graph, B2 hypergraph

3

Generalization Evaluation

OOD axes: material, size / topology, initial config, distractors
Metric: success rate + IID vs OOD gap

Goal: show that structured relational input improves OOD robustness of frozen VLA policies

Thank You for Your Attention!

„PROJECT NO. 2025-2.1.2-EKÖP-KDP-2025-00003 HAS BEEN IMPLEMENTED WITH THE SUPPORT PROVIDED BY THE MINISTRY OF CULTURE AND INNOVATION OF HUNGARY FROM THE NATIONAL RESEARCH, DEVELOPMENT AND INNOVATION FUND, FINANCED UNDER THE 2025-2.1.2 UNIVERSITY RESEARCH SCHOLARSHIP PROGRAM - COOPERATIVE DOCTORAL PROGRAM FUNDING SCHEME.”