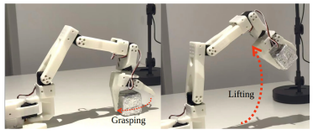
Low-Cost Gating-Based Vision and Proprioception Fusion for Object Property Classification

Jiaming Zhu • Kaitao Meng

The University of Manchester

7th ViTac Workshop
Learning to See and Feel:
Vision–Tactile Synergy for Embodied AI
ICRA 2026 · June 1, 2026 · Vienna, Austria

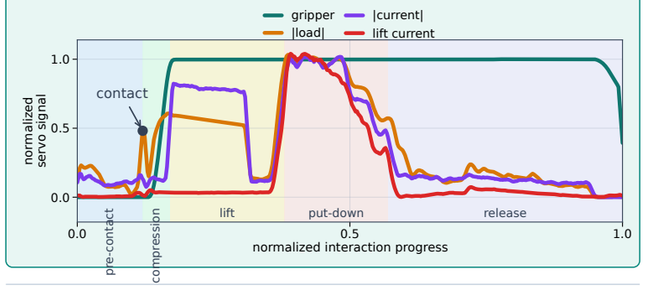
1 Dataset and Problem Setup



Low-cost observations

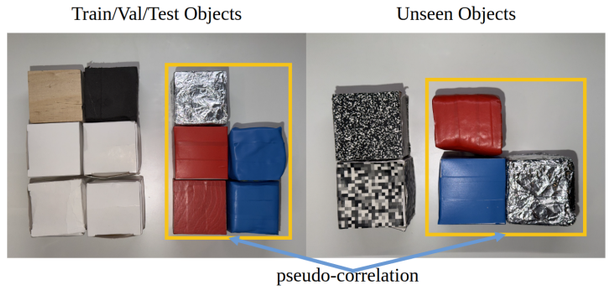
- One pre-grasp RGB image
- Internal servo signals position, load, current, velocity
- Standard grasp-and-lift trial

Real servo traces



Adversarial object split

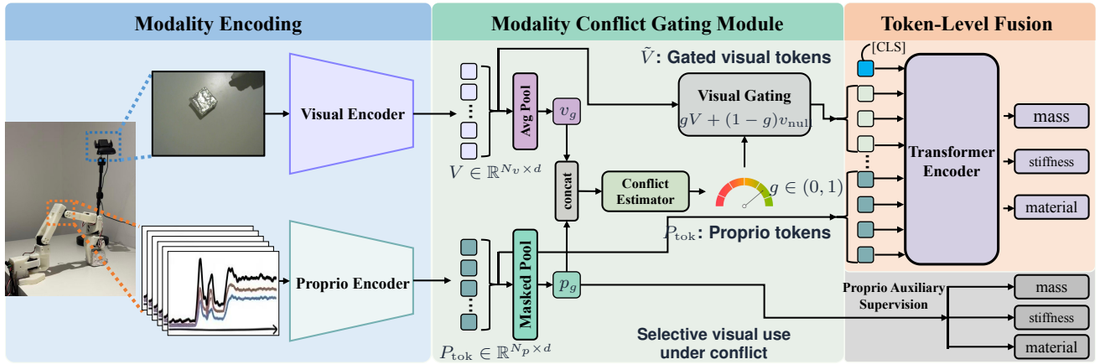
Training objects contain strong appearance–property pseudo-correlations. The unseen split deliberately violates them, exposing visual shortcuts.



Key question

Can a robot classify mass, stiffness, and material when visual appearance no longer predicts physical properties?

2 Gating-Based Visuo-Proprioceptive Fusion



Design principle

A scalar gate is conditioned on both modalities and is designed to reduce reliance on visual tokens when RGB cues conflict with proprioceptive evidence.

1. Encode

RGB image and servo sequence are encoded into visual/proprio tokens. The interface uses one camera frame plus internal motor feedback.

2. Gate

Visual tokens are interpolated with a learned null token: $\tilde{V} = gV + (1-g)v_{\text{null}}$. The gate keeps or suppresses visual evidence.

3. Predict

A shared Transformer predicts mass, stiffness, and material from fused tokens, so the fused code supports all labels.

Model objective

The fused representation solves a three-task physical-property prediction problem. Auxiliary proprioceptive heads keep the internal-signal branch discriminative, while gate regularization discourages premature saturation.

$$L = \sum_{k \in \{\text{mass,stiffness,material}\}} \left[L_{\text{CE}}(\hat{y}^k, y^k) + \lambda_{\text{aux}} L_{\text{CE}}(\hat{y}^k_{\text{aux}}, y^k) \right] + \lambda_{\text{reg}} R_{\text{ent}}(g)$$

Objective components

L_{CE}: main labels for mass, stiffness, and material.

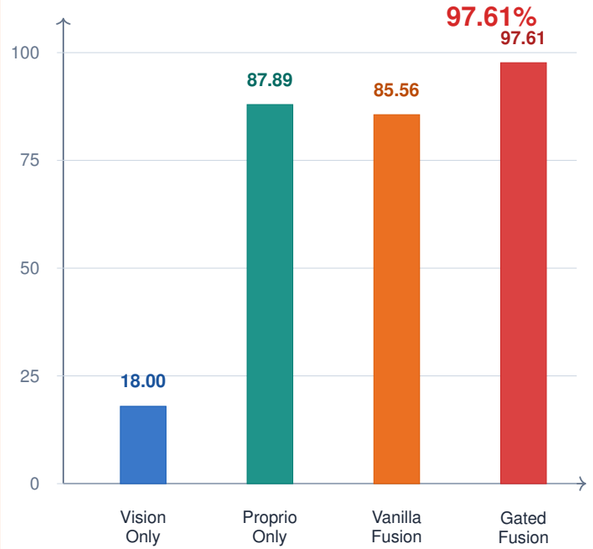
L_{aux}: keeps proprioceptive tokens predictive.

$R_{\text{ent}}(g)$: discourages early gate saturation.

3 Results and Takeaways

Unseen deceptive objects

Average accuracy (%), 5 seeds



Gated Fusion on seen objects **99.71%**

Takeaways

1. **Low-cost sensing** — RGB camera plus internal servo feedback only.
2. **Robustness stress test** — Unseen objects break visual-property shortcuts.
3. **Selective visual use** — Gating fusion improves unseen-object robustness.

Reading the plot

When visual shortcuts break, proprioception carries the stable signal; the gate recovers accuracy by using visual tokens selectively.

Contact: jiaming.zhu-2@student.manchester.ac.uk kaitao.meng@manchester.ac.uk