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

Jiaming Zhu[1] and Kaitao Meng[1]

***Abstract*— We propose a low-cost gating-based visuo-proprioceptive fusion framework that combines a pre-grasp RGB image with internal servo signals, including position, load, current, and velocity, collected during a standard grasp-and-lift interaction. To rigorously evaluate multimodal robustness, we introduce an adversarial dataset containing visually deceptive objects where appearance-based pseudo-correlations are intentionally disrupted. Our architecture dynamically suppresses unreliable visual hypotheses under modality conflict while using auxiliary proprioceptive supervision to maintain an independently discriminative proprioceptive branch. On visually deceptive objects, our gated fusion achieves 97.61% accuracy, effectively mitigating visual biases and outperforming both proprioception-only (87.89%) and vanilla fusion (85.56%) baselines. These results show that low-cost proprioception provides reliable physical grounding, while visual information should be leveraged selectively rather than trusted uniformly.**

I. INTRODUCTION

By combining vision and touch, humans can perform dexterous object manipulation while dynamically adapting to object characteristics. Estimating inherent object properties is crucial during contact-rich manipulation of diverse objects, especially fragile or deformable ones [1], [2]. Recent vision-based methods exhibit strong semantic reasoning capabilities, and can utilize common knowledge to infer properties [3], [4]. However, appearance is often an unreliable proxy for underlying physical attributes: objects that look similar may differ substantially in weight, compliance, or material composition. While tactile sensing can provide more direct and reliable physical evidence, many existing tactile or force-torque solutions depend on costly and fragile hardware, limiting their scalability and practical deployment in low-cost robotic systems [5], [6], [7].

This paper proposes a low-cost gating-based vision and proprioception fusion approach. The method utilizes only internal robotic arm servo signals and a single pre-interaction top-down photograph to rapidly classify the mass, stiffness, and material of objects during interaction. To rigorously evaluate robustness under appearance–physics mismatch, we construct a dataset of visually deceptive objects in which spurious correlations between appearance cues and physical properties are intentionally introduced or broken. This setting tests whether a model genuinely reasons about physical attributes or merely relies on superficial visual regularities. We collected over 800 samples via a standardized automated procedure. Each sample contains a pre-interaction global photograph and a segment of multi-channel internal servo signals (position, load, current, and velocity) recorded during the interaction.

Our model (Fig. 1) encodes visual and proprioceptive observations into modality-specific representations and fuses them through a gating mechanism that is jointly conditioned on both streams. The key idea is to dynamically suppress unreliable visual evidence when the two modalities conflict, thereby improving robustness on visually ambiguous objects. The fused representation is then used for multi-task prediction of mass, stiffness, and material. Our contributions are twofold: First, we introduce a low-cost gating-based visuo-proprioceptive fusion architecture classifying object properties using only a single RGB photograph and internal servo signals. Second, we provide a dataset containing visual traps, consisting of photographs and internal servo signals.

II. DATASET AND METHOD

A. Adversarial Dataset Design

To rigorously evaluate multimodal robustness, we construct a custom dataset of 16 specially crafted objects (Fig. 2) featuring visual traps: training-set visual features (e.g., color, texture) strongly correlate with physical properties (mass, stiffness, material), whereas the unseen test set deliberately breaks these spurious correlations. For example, an object visually identical to a heavy training object might actually possess a low mass. This adversarial split ensures models relying solely on visual shortcuts fail, necessitating genuine physical grounding through proprioception. Our data are collected via a standardized automatic grasp-and-lift procedure, capturing a pre-interaction RGB image and multi-channel internal servo signals (position, load, current, and velocity).

B. Gated Token-Level Fusion

The visual branch employs a frozen ResNet encoder [8] extracting visual tokens $V \in \mathbb{R}^{N_v \times d}$, whereas the proprioceptive branch uses a temporal encoder mapping raw servos signals into proprioceptive tokens $P_{\text{tok}} \in \mathbb{R}^{N_p \times d}$.

To prevent misleading visual cues from dominating the predictions, we introduce a sample-wise gating module that aggregates the modalities into compact summary vectors (v_{g} and p_{g}) by applying global average pooling over the visual and non-padded proprioceptive token sequences. A multi-layer perceptron with Sigmoid activation maps the concatenated summary $[v_{\text{g}}; p_{\text{g}}]$ to a scalar gate $g \in (0,1)$. Subsequently, we gate the visual tokens through interpolation with a learned null token:

$$\tilde{V} = gV + (1-g)v_{\text{null}}, \tag{1}$$

[1]Jiaming Zhu and Kaitao Meng are with the Department of Electrical & Electronic Engineering, University of Manchester, Manchester, UK.
Correspondence: kaitao.meng@manchester.ac.uk.

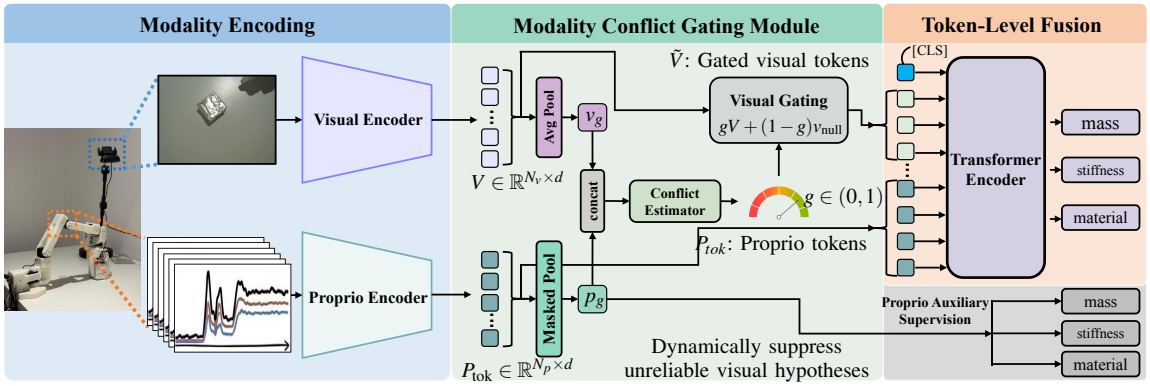

Fig. 1. Overview of the proposed architecture.

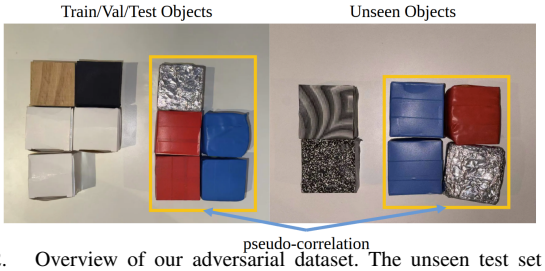

Fig. 2. Overview of our adversarial dataset. The unseen test set deliberately breaks training-set visual-physical pseudo-correlations to evaluate multimodal robustness.

where v_{null} broadcasts over the visual sequence. The gated visual tokens $\tilde{V}$, proprioceptive tokens P_{tok}, and a classification token ([CLS]) [9] are concatenated and processed by a shared Transformer encoder [10] to predict the mass, stiffness, and material.

To ensure the proprioceptive branch learns independent physical representations, we attach three proprioception-only auxiliary heads predicting identical targets from p_{g}. The overall loss combines the main classification, auxiliary proprioceptive, and gate entropy regularization losses:

$$L = \sum_{k \in \{m,s,u\}} \left[L_{\text{CE}}(\hat{y}^k, y^k) + \lambda_{\text{aux}} L_{\text{CE}}(\hat{y}^k_{\text{aux}}, y^k) \right] + \lambda_{\text{reg}} R_{\text{ent}}(g). \tag{2}$$

Minimizing the negative Bernoulli entropy $R_{\text{ent}}(g)$ penalizes premature gate saturation, promoting early cross-modal exploration.

III. Experiments and Discussion

A. Protocol

We evaluate on a unified task predicting three mass classes, four stiffness classes, and five material classes, reporting the average accuracy across these tasks over five random seeds. We compare our Gated Fusion model against three baselines. The Vision-only baseline uses a ResNet-18 on pre-grasp images followed by a Transformer encoder; the Proprio-only baseline applies 1D-CNNs on servo signals before the Transformer encoder; and the Vanilla Fusion baseline concatenates visual and proprioceptive tokens for the shared Transformer without gating. Such feature concatenation serves as a representative proxy for standard early-fusion approaches widely adopted in multimodal robot learning [11], [12]. In our evaluation, seen objects refer to those present during training, whereas unseen objects are exclusively drawn from the out-of-distribution (OOD) test set featuring visually deceptive appearances. Given our adversarial split design, high unseen-object accuracy indicates effective resistance to deliberate visual bias rather than mere generalization.

TABLE I

Main results and task-wise unseen-object accuracy (%). Values are mean ± std across five seeds.

Overall Accuracy & Gate Score			
Method	Seen-object	Unseen-object	Gate
Vision-only	95.39 ± 0.73	18.00 ± 6.16	–
Proprio-only	95.29 ± 1.93	87.89 ± 1.62	–
Vanilla Fusion	99.31 ± 0.73	85.56 ± 8.39	–
Ours (Gated Fusion)	**99.71 ± 0.59**	**97.61 ± 3.68**	0.589
Task-wise Unseen-object Accuracy			
Method	Mass	Stiffness	Material
Vision-only	17.17 ± 6.55	17.83 ± 6.84	19.00 ± 5.15
Proprio-only	**100.00 ± 0.00**	81.00 ± 2.76	82.67 ± 2.20
Vanilla Fusion	87.67 ± 7.91	84.50 ± 9.61	84.50 ± 8.04
Ours (Gated Fusion)	**100.00 ± 0.00**	**95.17 ± 7.61**	**97.67 ± 3.43**

B. Main Unseen-Object Results

Table I reveals that vision-only and vanilla fusion models degrade significantly on out-of-distribution objects. In practical robotics, deceptive appearances frequently contradict actual physical properties, breaking the spurious correlations exploited by standard architectures. Our Gated Fusion mitigates these modality conflicts by dynamically suppressing unreliable visual hypotheses, achieving 97.61% accuracy on unseen objects while maintaining near-ceiling performance on seen data. Task-wise analysis further shows that while proprioception alone perfectly predicts unseen mass, it fails in stiffness and material estimation; conversely, Gated Fusion consistently excels across all tasks.

IV. Conclusion

This paper presents a low-cost, gating-based visuo-proprioceptive fusion approach for robust physical property estimation. By dynamically suppressing unreliable visual cues during modality conflicts, our architecture effectively reduces reliance on misleading visual cues and maintains high accuracy on visually deceptive objects.

REFERENCES

[1] A. Dutta, E. Burdet, and M. Kaboli, "Predictive Visuo-Tactile Interactive Perception Framework for Object Properties Inference," *IEEE Transactions on Robotics*, vol. 41, pp. 1386–1403, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10847911

[2] J. Zhu, A. Cherubini, C. Dune, D. Navarro-Alarcon, F. Alambeigi, D. Berenson, F. Ficuciello, K. Harada, J. Kober, X. Li, J. Pan, W. Yuan, and M. Gienger, "Challenges and Outlook in Robotic Manipulation of Deformable Objects," *IEEE Robotics & Automation Magazine*, vol. 29, no. 3, pp. 67–77, Sep. 2022. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9721534

[3] X. Xu, W. Ge, D. Qiu, Z. Chen, D. Yan, Z. Liu, H. Zhao, H. Zhao, S. Zhang, J. Liang, and Y.-C. Chen, "GaussianProperty: Integrating Physical Properties to 3D Gaussians with LMMs."

[4] Z. Guo, H. Chen, X. Mai, Q. Qiu, G. Ma, Z. Kappassov, Q. Li, and N. Chen, "Robotic Perception with a Large Tactile-Vision-Language Model for Physical Property Inference," in *AI Enabled Robotic Loco-Manipulation*, Q. Li, M. Xie, M. O. Tokhi, and M. F. Silva, Eds. Cham: Springer Nature Switzerland, 2025, pp. 146–157.

[5] J. Lin, R. Calandra, and S. Levine, "Learning to Identify Object Instances by Touch: Tactile Recognition via Multimodal Matching," in *2019 International Conference on Robotics and Automation (ICRA)*, May 2019, pp. 3644–3650, iSSN: 2577-087X. [Online]. Available: https://ieeexplore.ieee.org/document/8793885

[6] S. Gano, A. George, and A. B. Farimani, "Low-Fidelity Visuo-Tactile Pre-Training Improves Vision-Only Manipulation Performance," in *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct. 2025, pp. 15 983–15 989, iSSN: 2153-0866. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/11246929

[7] A. Maldonado, H. Alvarez, and M. Beetz, "Improving robot manipulation through fingertip perception," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Oct. 2012, pp. 2947–2954, iSSN: 2153-0866. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/6385560

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 770–778, iSSN: 1063-6919. [Online]. Available: https://ieeexplore.ieee.org/document/7780459

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423/

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[11] M. A. Lee, Y. Zhu, K. Srinivasan, P. Shah, S. Savarese, L. Fei-Fei, A. Garg, and J. Bohg, "Making Sense of Vision and Touch: Self-Supervised Learning of Multimodal Representations for Contact-Rich Tasks," in *2019 International Conference on Robotics and Automation (ICRA)*, May 2019, pp. 8943–8950, iSSN: 2577-087X. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8793485

[12] A. Nagrani, S. Yang, A. Arnab, A. Jansen, C. Schmid, and C. Sun, "Attention Bottlenecks for Multimodal Fusion," in *Advances in Neural Information Processing Systems*, vol. 34. Curran Associates, Inc., 2021, pp. 14 200–14 213. [Online]. Available: https://proceedings.neurips.cc/paper/2021/hash/76ba9f564ebbc35b1014ac498fafadd0-Abstract.html