Robotic Perception with a Large Tactile-Vision-Language Model for Physical Property Inference

Zexiang Guo[1*], Hengxiang Chen[1*], Xinheng Mai[1*], Qiusang Qiu[1], Gan Ma[2], Zhanat Kappassov[3], Qiang Li[1†], and Nutan Chen[4]

[1] College of Big Data and Internet, Shenzhen Technology University, China,
[2] Sino-German College of Intelligent Manufacturing, Shenzhen Technology University, China,
[3] Robotics Department, Institute of Smart Systems and Artificial Intelligence (ISSAI), Nazarbayev University, Kazakhstan
[4] Foundation Robotics Labs, Germany

[*]These authors contributed equally to this work. [†]Corresponding author

Abstract. Inferring physical properties can significantly enhance robotic manipulation by enabling robots to handle objects safely and efficiently through adaptive grasping strategies. Previous approaches have typically relied on either tactile or visual data, limiting their ability to fully capture properties. We introduce a novel cross-modal perception framework that integrates visual observations with tactile representations within a multimodal vision-language model. Our physical reasoning framework, which employs a hierarchical feature alignment mechanism and a refined prompting strategy, enables our model to make property-specific predictions that strongly correlate with ground-truth measurements. Evaluated on 35 diverse objects, our approach outperforms existing baselines and demonstrates strong zero-shot generalization.

Keywords: tactile perception, visual-tactile fusion, physical property inference, multimodal integration, robot perception

1 Introduction

Accurate perception of object physical properties is fundamental for robots to perform reliable manipulation in unstructured environments. While humans seamlessly integrate visual and tactile cues to infer material characteristics [1], robotic systems often struggle to achieve comparable performance due to limitations in unimodal sensing. Traditional vision-based approaches, though effective for geometric perception, frequently fail to capture intrinsic material attributes such as hardness, elasticity, and surface roughness [2]. Conversely, tactile sensing provides rich contact information but requires physical contact — a significant drawback when handling delicate or unknown objects [3]. Recent advances

International Conference on Climbing and Walking Robots (CLAWAR), 2025

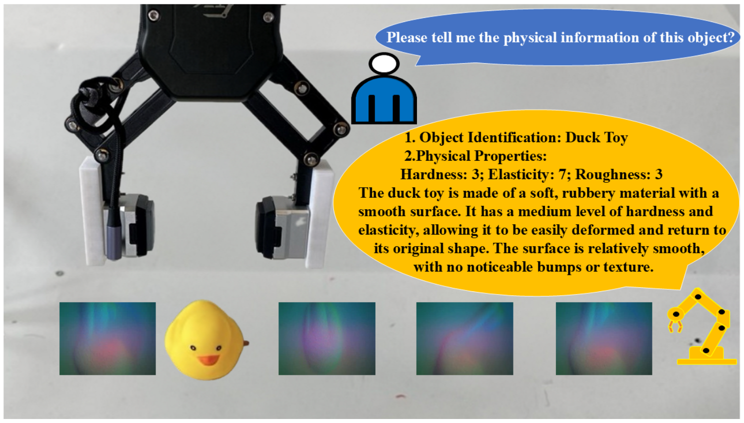
Please tell me the physical information of this object?
1. Object Identification: Duck Toy
2.Physical Properties:
Hardness: 3; Elasticity: 7; Roughness: 3
The duck toy is made of a soft, rubbery material with a smooth surface. It has a medium level of hardness and elasticity, allowing it to be easily deformed and return to its original shape. The surface is relatively smooth, with no noticeable bumps or texture.

Fig. 1: Through visual and tactile image input and human language interaction, our model infers and gives detailed physical properties of the duck toy and gives specific physical property scores as specified by the structured scoring guidelines.

in multimodal learning have shown significant potential for integrating tactile perception with language models to enhance physical reasoning capabilities [4]. However, two fundamental limitations persist: (1) **Sensory constraints in tactile systems:** Current tactile sensors offer insufficient data capture for comprehensive material characterization, particularly when handling objects with complex composite structures; and (2) **Underutilized language model potential:** Existing implementations fail to fully leverage the reasoning capacity of language models through strategic prompting and effective multimodal fusion. To address these challenges, we propose an enhanced multimodal framework with two core innovations that enable physical property inference for robotic grasping tasks. As illustrated in Fig.1, our vision-tactile-language integration empowers the robotic arm to accurately estimate critical material characteristics (e.g., hardness, elasticity, surface roughness), allowing it to grasp the duck toy while preserving its structural integrity. The framework's technical advancements include:

- **Proactive Perception Architecture:** By fusing visual cues with historical tactile information, our model is capable of predicting important physical attributes—such as hardness, elasticity, and roughness—prior to contact.
- **Structured Reasoning Prompts:** A staged reasoning protocol that guides multimodal language models through object recognition, material analysis, and property quantification to enhance inference accuracy.

2 Related Work

2.1 Tactile Perception in Robotics

Tactile sensing has become essential for robotic manipulation, with various sensor technologies capturing detailed contact information [5,6]. Vision-based tactile sensors (e.g., GelSight [7], GelSlim [8]) excel at local texture and hardness estimation but are inherently limited in capturing global object properties due to restricted sensing areas. Recent tactile representation learning approaches leveraging deep networks, including tactile-kinematic fusion for shape reconstruction [9] and self-supervised tactile-visual alignment [10], have improved property estimation. Nevertheless, each modality individually faces limitations: tactile sensors suffer from partial observability and difficulties in dynamic property inference, whereas vision alone lacks fine-grained contact information. Previous works have addressed some of these issues by combining visual and tactile sensing [11–13]. However, our approach further enhances this multimodal integration, effectively leveraging visual priors to enrich tactile perception.

2.2 Multimodal Fusion Approaches

The integration of tactile and visual modalities has evolved through several fusion paradigms to address limitations inherent in single-modal perception. Early fusion methods [14], which directly concatenate raw tactile and visual features, face performance degradation due to modality misalignment. Subsequently developed late fusion techniques [15] process each modality independently, yet they are limited in capturing essential cross-modal correlations required for inferring complex physical properties. More advanced hybrid methods adopt intermediate fusion strategies, including contrastive learning for feature alignment [16] and attention mechanisms for adaptive modality weighting [17]. However, these techniques typically rely on extensive paired training datasets, potentially limiting their generalization capabilities when encountering novel objects or properties.Building upon these limitations, our research introduces a novel hierarchical prompting strategy utilizing pre-trained vision-language models as robust knowledge priors. This framework implements property-specific fusion rules, effectively enabling zero-shot generalization through structured physical reasoning.

2.3 Physical Property Reasoning with Large Models

Recent advances in large vision language models have demonstrated their capability for physical property reasoning by leveraging multimodal inputs and commonsense knowledge [18]. For instance, GPT-4V has been shown to infer liquid viscosity by analyzing time-series plots of force/torque sensor data [19], while OCTOPI [20] predicts material properties like hardness and roughness from tactile images through specialized prompting strategies. However, existing methodologies predominantly focus on passive interpretation of sensory data without actively guiding the reasoning processes and rely on generalized multimodal fusion rather than explicitly structured, property-centric prompting. Our research substantially advances this direction by proposing a hierarchical reasoning framework that systematically decomposes physical property inference into sequential, interpretable stages-object recognition, material analysis, and quantitative assessment. Through carefully designed property-specific prompts, our approach actively directs model attention toward relevant sensory cues critical to each property, such as pressure response for hardness, deformation patterns for elasticity, and surface textures for roughness evaluation, thereby improving reasoning accuracy and interpretability.

3 Methodology

In our method, we introduce a multimodal model integrating textual, visual, and tactile data for comprehensive object analysis. As depicted in Fig. 2, the input query is parsed into dedicated modality-specific pathways. Text is tokenized and embedded via a language tokenizer, while visual and tactile images are encoded using ViT-L/14 [21] and projected into a shared embedding space using modality-specific MLP layers. Special markers (`<img_start>`, `<img_end>`, `<tact_start>`, `<tact_end>`) clearly delineate embedding boundaries. These embeddings are concatenated with textual features and fed into a large language model (Vicuna-7B [22]), allowing joint multimodal attention to generate detailed object property descriptions, such as hardness, elasticity, and roughness.

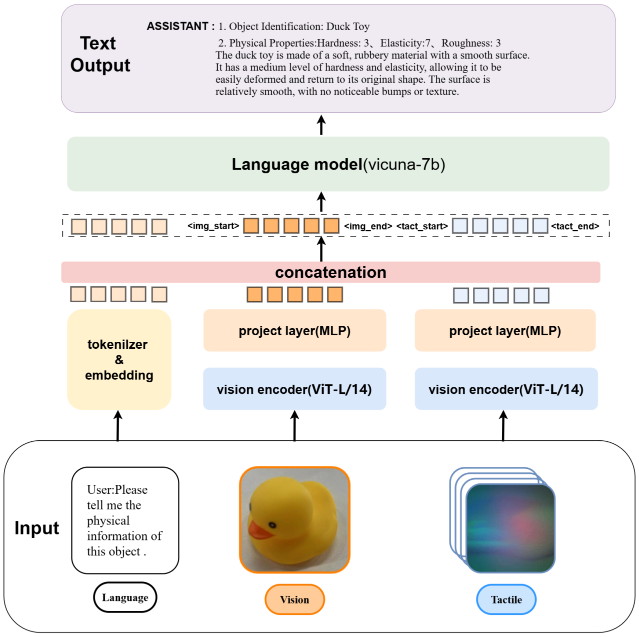
Text Output
ASSISTANT : 1. Object Identification: Duck Toy
2. Physical Properties:Hardness: 3、Elasticity:7、Roughness: 3
The duck toy is made of a soft, rubbery material with a smooth surface. It has a medium level of hardness and elasticity, allowing it to be easily deformed and return to its original shape. The surface is relatively smooth, with no noticeable bumps or texture.
Language model(vicuna-7b)
<img_start>
<img_end>
<tact_start>
<tact_end>
concatenation
tokenilzer & embedding
project layer(MLP)
project layer(MLP)
vision encoder(ViT-L/14)
vision encoder(ViT-L/14)
Input
User:Please tell me the physical information of this object .
Language
Vision
Tactile

Fig. 2: The architecture of a multimodal large model. After embedding and tokenizing the object image and tactile image alongside the text, the resulting vectors are concatenated and input into the large language model.

3.1 Vision Processing

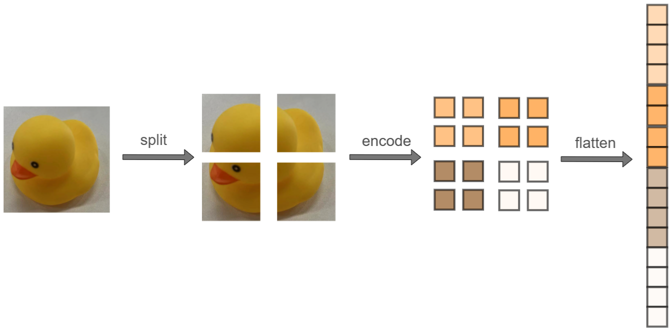

Fig. 3: Vision processing pipeline. The image is partitioned into multiple regions by the segmentation module, after which the encoder extracts a feature matrix that is flattened into a one-dimensional vector and fed into the LLM(see Fig.2).

We employ CLIP [21] to process visual information, leveraging a visual encoder (ViT-L/14) trained to learn shared representations between images and text. As shown in Fig.2 , the overall multimodal architecture incorporates image embeddings alongside textual inputs into the LLM. To align dimensionality and semantics with the LLM's native embedding space, we adopt the pre-trained linear transformation layer from LLaVA [23], which projects the penultimate output of CLIP into the language model's word embedding space.

In addition to extracting features from the CLIP visual encoder, we insert two specialized boundary tokens, `<img_start>` and `<img_end>`, around image-derived embeddings. These tokens (initialized by semantic averaging of descriptive phrases and then frozen during fine-tuning) explicitly separate visual content from text inputs, thus assisting the model in distinguishing modalities. As depicted in Fig.3, we further segment the image into multiple regions, extract a feature matrix from each region, and flatten it into a one-dimensional representation to feed into the LLM.

3.2 Tactile Processing

To incorporate tactile perception and enhance physical reasoning, we adopt the OCTOPI framework. This framework employs a CLIP-based [21] tactile encoder that processes tactile data and fuses it with the LLM, enabling a deeper understanding of object properties. Specifically, the tactile encoder extracts features from a sequence of tactile images, encoding both spatial and temporal information. As shown in Fig.4, we add positional encodings to these sequential features to preserve the order and timing of tactile interactions. By training on physics-based datasets with annotated tactile videos and physical property labels, the model acquires rich, tactile-aware representations that improve performance in tasks such as object property prediction and scenario reasoning.

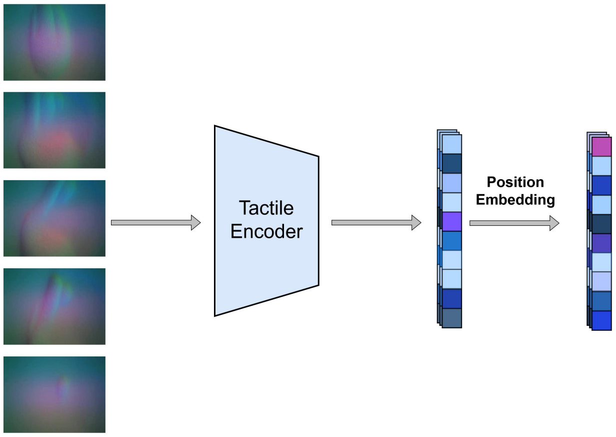

Fig. 4: A sequence of tactile images is first processed by the tactile encoder to extract feature representations. The extracted features are then transformed into a structured feature vector, followed by the addition of positional embeddings to encode temporal dependencies.

3.3 Multimodal Fusion through Feature Concatenation

After we obtain the projected object image feature vector (F_o), the projected tactile image feature vector (F_t), and the linguistic feature vector (F_l) from the LLM's embedding space, we concatenate them channel-wise into a unified representation:

$$F_{\text{concat}} = [\, F_o \; ; \; F_t \; ; \; F_l \,].$$

This fused vector F_{concat} retains distinguishing features from each modality while enabling cross-modal interaction. It then serves as the input to downstream modules for tasks such as multimodal reasoning, classification, or object recognition, thereby capturing both the physical and semantic attributes of the target object.

3.4 Refined Prompting Strategy for Physical Property Scoring

We designed a structured prompt to enable comprehensive physical property analysis using multimodal (visual and tactile) data. The prompt clearly defines

Table 1: Physical Property Rating Scales

Property	Score Range	Characterization	Example Materials
Hardness	1–2	Extremely soft	Cotton, sponge
	3–4	Soft	Rubber ball, soft plastic toy
	5–6	Medium	Plastic container, shoe sole
	7–8	Hard	Wood, ceramic plate
	9–10	Extremely hard	Metal, diamond
Elasticity	1–2	Minimal elasticity	Clay, dry sponge, wooden ruler
	3–4	Low elasticity	Rubber eraser, hard plastic, book cover
	5–6	Medium elasticity	Foam ball, silicone, thick rubber mat
	7–8	High elasticity	Rubber band, bouncy ball, yoga mat
	9–10	Maximum elasticity	Trampoline surface, latex sheet, inflated balloon
Roughness	1–2	Extremely smooth	Glass, polished marble
	3–4	Smooth	Plastic surface, ceramic mug
	5–6	Medium texture	Paper, leather, cardboard
	7–8	Rough	Sandpaper, concrete, bark of a tree
	9–10	Extremely rough	Gravel, coarse fabric, pumice stone

the analysis goal, emphasizing material-aware reasoning and avoiding generic responses. It guides the model through two phases: visual-based object identification (color, shape, texture) and combined material-tactile property evaluation. A 10-point Likert scale quantifies three essential properties (Table 1), enhancing nuanced differentiation. Outputs include justified object identification and property scores with material rationales. Constraints ensure balanced score usage and material-focused reasoning.

When the user inputs a request along with an image, the request is encoded by the text encoder together with the prompt. Simultaneously, the object image and the tactile image are processed by their respective encoders. To facilitate proper identification and integration of different modalities, special tokens `<img_start>`, `<img_end>`, `<tact_start>`, and `<tact_end>` are used to mark the boundaries of visual and tactile features. This design enhances multimodal integration by associating visual and tactile information to provide a more comprehensive object representation. By unifying feature alignment, it improves cross-modal compatibility, while deep semantic understanding optimizes adaptation to complex scenarios.

4 Experiments

4.1 Hardware

To evaluate our cross-modal perception framework, we conducted comprehensive experiments using a robotic system equipped with a GelSight Mini tactile sensor for high-resolution contact data acquisition and a RealSense D410 camera for visual perception. We selected 35 common household objects (Fig.5) spanning diverse materials (plastic, metal, wood, rubber, etc.) and geometric properties.

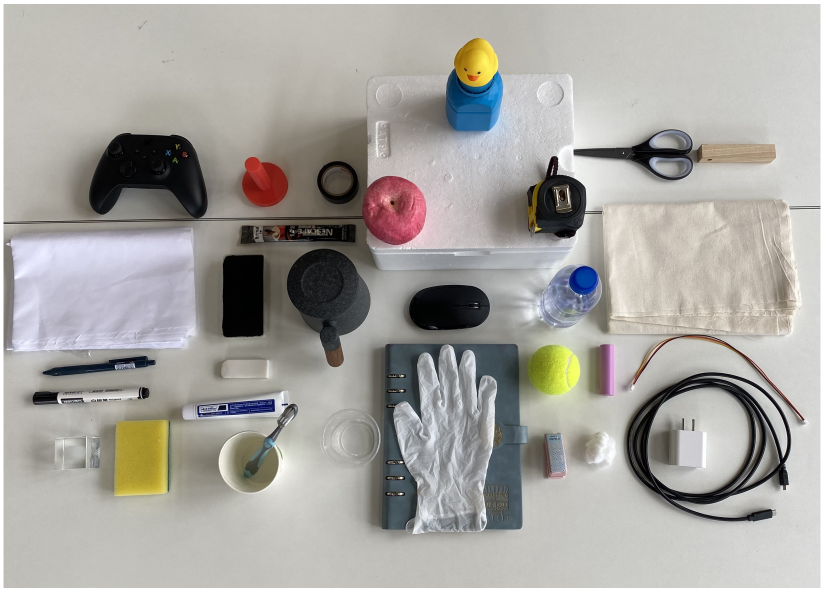

Fig. 5: Object set of 35 common household items, spanning nine major material categories—plastic, rubber, metal, wood, ceramic, glass, foam, paper, and textile—to validate the generalizability of the experimental data across diverse materials, for evaluating multimodal models in physical property reasoning.

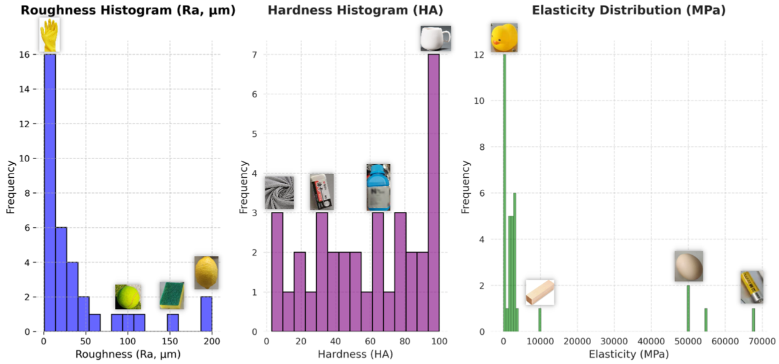

Fig. 6: During the experiment, we selected 35 objects in the laboratory, measured their roughness, hardness and elastic modulus, and drew the histogram shown in the figure above.

Each object was annotated with ground truth physical properties measured by professional instruments: hardness (Shore scale) with PosiTector SHD, elastic modulus with C610H Auto Tensile Tester, and surface roughness (Ra) with RUGOSURF 20 roughness tester (Fig.6).

4.2 Data Collection

The data collection process was designed to systematically capture multimodal information for comprehensive physical property analysis. For tactile data acquisition, we employed a GelSight Mini sensor operating at 20 fps to record six-second videos of each interaction, encompassing the complete contact cycle from approach to retraction. These videos were subsequently sampled at 250ms intervals to obtain representative frames while maintaining temporal coherence. Ground truth measurements were obtained following established protocols to ensure accuracy and repeatability. Hardness measurements were conducted using a PosiTector SHD durometer, with three tests performed at predetermined locations on each object and a standardized 5-second dwell time. Elastic modulus characterization was performed using a C610H tensile tester, analyzing stress-strain responses in the linear deformation regime. Surface roughness was quantified with a RUGOSURF 20 profilometer, executing multiple scans per object with carefully controlled parameters.

4.3 Experimental results

To evaluate our model, we employed designed prompts to assess the physical properties of 35 objects through both our method and the Octopi framework [20], while simultaneously obtaining ground truth measurements of the objects' attributes using specialized instrumentation. Specifically, true hardness was measured using a Shore hardness tester, the elastic modulus was determined using a universal material testing machine, and surface roughness was quantified using a portable surface roughness instrument. These instruments ensured that our ground truth data were both accurate and reproducible.

Following data collection, we normalized each dataset and computed Spearman correlation coefficients between the model scores and the ground truth measurements. These analyses allowed us to quantitatively assess the predictive accuracy of the physical property inference for hardness, elasticity, and roughness, highlighting significant differences between our approach and the baseline Octopi method. As can be seen from the Table 2, the correlation coefficients between the models and ground truth measurements reveal significant differences in the performance of our model compared to Octopi across the three physical attributes: hardness, elasticity, and roughness.

For hardness, our model exhibits a moderate and statistically significant positive correlation with the ground truth (Spearman's $\rho = 0.501$, p = 0.005), demonstrating its capability to integrate visual and tactile cues effectively. In comparison, the pure vision model yields a weaker correlation (ρ = 0.307, p = 0.099), failing to reach statistical significance. Interestingly, both Octopi (fine-grained)

Table 2: Zero-shot evaluation: Comparison of Spearman's rank correlation between models and ground truth (Octopi as the tactile-only model; Octopi-ViTaL is our model)

Attribute	Method	Correlation Coefficient	P-value
Hardness	Octopi-ViTaL	**0.501**	**0.005**
	Octopi-ViTaL (vision only)	0.307	0.099
	Octopi (fine-grained)	0.307	0.099
	Octopi (original)	0.015	0.935
Elasticity	Octopi-ViTaL	**0.530**	**0.003**
	Octopi-ViTaL (vision only)	0.452	0.012
	Octopi (fine-grained)	0.053	0.781
	Octopi (original)	-0.060	0.753
Roughness	Octopi-ViTaL	**0.643**	**0.0001**
	Octopi-ViTaL (vision only)	0.413	0.023
	Octopi (fine-grained)	-0.010	0.959
	Octopi (original)	0.118	0.534

and Octopi (original) perform worse: while the fine-grained version—Octopi uses our prompt to score the physical properties of objects.—achieves a similar correlation to pure vision (ρ = 0.307, p = 0.099), the original version—which relies on Octopi's predefined three-level classification system—shows virtually no correlation with the ground truth (ρ = 0.015, p = 0.935). These results confirm that our multimodal model surpasses both unimodal and tactile-only baselines, and benefits significantly from combining sensory modalities.

In the case of elasticity, because elastic modulus is inversely proportional to perceived "elasticity," we report the absolute value of Spearman's ρ to reflect predictive strength regardless of sign. Our model achieves a Spearman correlation of 0.530 (p = 0.003), clearly outperforming the vision-only baseline (ρ = 0.452, p = 0.012). Meanwhile, Octopi (fine-grained) shows a negligible correlation (ρ = 0.053, p = 0.781), and Octopi (original) displays a weak negative trend (ρ = -0.060, p = 0.753). The poor performance of both Octopi variants indicates that tactile input alone lacks sufficient expressiveness for elasticity estimation, even when adapted to more descriptive prompts.

The comparison is most striking for roughness, where our model achieves a strong and statistically robust correlation with ground truth (ρ = 0.643, p = 0.0001). Although the pure vision model also yields a moderate correlation (ρ = 0.413, p = 0.023), it falls short of our model's performance. Octopi (fine-grained) shows no meaningful correlation (ρ = -0.010, p = 0.959), and the original Octopi version fares only slightly better (ρ = 0.118, p = 0.534), with neither result statistically significant. This further demonstrates that a tactile-only approach—even with fine-tuned prompts or structured rating schemes—fails to adequately capture surface texture without visual context.

When applied in a zero-shot fashion to our new setup, the pretrained Octopi model failed to produce meaningful predictions (e.g., Spearman's $\rho < 0.1$; see Table 2). This failure arises from multiple domain shifts: we use a GelSight Mini with different resolution and calibration compared to Octopi's original high-resolution GelSight; lighting and camera angles differ. These combined shifts in sensor modality, resolution, and lighting prevent Octopi from succeeding zero-shot on our data.

Overall, these results highlight the clear advantage of our multimodal approach. By fusing vision and touch, our model consistently achieves statistically significant and higher correlations with ground truth across all three physical attributes. In contrast, both the vision-only and tactile-only methods—particularly the Octopi framework in its original and adapted forms—fall short, reinforcing the value of cross-modal integration in physical property understanding.

5 Conclusion

We proposed a novel approach to enhance tactile perception through visual compensation and optimized prompt engineering, leveraging VLM for cross-modal robotic perception. By effectively integrating visual priors and structuring language model interactions, our method overcomes tactile-only limitations and significantly improves physical property inference, especially in roughness estimation. The success of our framework underscores the value of multimodal reasoning with VLMs for robotic applications. Future work will explore applying this multimodal tactile-visual approach to robotic grasping tasks involving adaptive manipulation of objects with different material properties.

Acknowledgment

This research is supported by the "Natural Science Foundation of Top Talent of SZTU", grant no. GDRC202411.

References

1. Jeka, J., Oie, K.S., Kiemel, T.: Multisensory information for human postural control: integrating touch and vision. *Exp. Brain Res.* **134**, 107–125 (2000)
2. Fleming, R.W.: Visual perception of materials and their properties. *Vision Res.* **94**, 62–75 (2014)
3. Chi, C., Sun, X., Xue, N., et al.: Recent progress in technologies for tactile sensors. *Sensors* **18**(4), 948 (2018)
4. Zeng, F., Gan, W., Wang, Y., et al.: Large language models for robotics: A survey. *arXiv preprint* arXiv:2311.07226 (2023)
5. Liu, H., Huang, B., Li, Q., et al.: Multi-fingered tactile servoing for grasping adjustment under partial observation. *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)* **2022**, 7781–7788 (2022)

6. Van Hoof, H., Chen, N., Karl, M., et al.: Stable reinforcement learning with autoencoders for tactile and visual data. *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)* **2016**, 3928–3934 (2016)
7. Yuan, W., Dong, S., Adelson, E.H.: Gelsight: High-resolution robot tactile sensors for estimating geometry and force. *Sensors* **17**(12), 2762 (2017)
8. Donlon, E., Dong, S., Liu, M., et al.: Gelslim: A high-resolution, compact, robust, and calibrated tactile-sensing finger. In: *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1927–1934. IEEE (2018)
9. Smith, E., Calandra, R., Romero, A., et al.: 3D shape reconstruction from vision and touch. *Adv. Neural Inf. Process. Syst.* **33**, 14193–14206 (2020)
10. Dave, V., Lygerakis, F., Rueckert, E.: Multimodal visual-tactile representation learning through self-supervised contrastive pre-training. In: *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 8013–8020. IEEE (2024)
11. Xu, W., Yu, Z., Xue, H., et al.: Visual-tactile sensing for in-hand object reconstruction. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8803–8812 (2023)
12. Taunyazov, T., Sng, W., See, H.H., et al.: Event-driven visual-tactile sensing and learning for robots. *arXiv preprint* arXiv:2009.07083 (2020)
13. Li, S., Yu, H., Ding, W., et al.: Visual–tactile fusion for transparent object grasping in complex backgrounds. *IEEE Trans. Robot.* **39**(5), 3838–3856 (2023)
14. Li, S., Tang, H.: Multimodal Alignment and Fusion: A Survey. *arXiv preprint* arXiv:2411.17040 (2024)
15. Liu, H., Yu, Y., Sun, F., et al.: Visual–tactile fusion for object recognition. *IEEE Trans. Autom. Sci. Eng.* **14**(2), 996–1008 (2016)
16. Meyer, J., Eitel, A., Brox, T., et al.: Improving unimodal object recognition with multimodal contrastive learning. In: *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5656–5663. IEEE (2020)
17. bibliography Wong, C.Y., Vergez, L., Suleiman, W.: Vision-and tactile-based continuous multimodal intention and attention recognition for safer physical human–robot interaction. *IEEE Trans. Autom. Sci. Eng.* (2023)
18. Gao, J., Sarkar, B., Xia, F., et al.: Physically grounded vision-language models for robotic manipulation. In: *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 12462–12469. IEEE (2024)
19. Lai, W., Zhang, T., Lam, T.L., et al.: Vision-language model-based physical reasoning for robot liquid perception. In: *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 9652–9659. IEEE (2024)
20. Yu, S., Lin, K., Xiao, A., et al.: Octopi: Object property reasoning with large tactile-language models. *arXiv preprint* arXiv:2405.02794 (2024)
21. Radford, A., Kim, J.W., Hallacy, C., et al.: Learning transferable visual models from natural language supervision. In: *International Conference on Machine Learning (ICML)*, pp. 8748–8763. PMLR (2021)
22. Chiang, W.L., Li, Z., Lin, Z., et al.: Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality. *See https://vicuna.lmsys.org (accessed 14 April 2023)* **2**(3), 6 (2023)
23. Liu, H., Li, C., Wu, Q., et al.: Visual instruction tuning. *Adv. Neural Inf. Process. Syst.* **36**, 34892–34916 (2023)