中国第四届力触觉技术及应用会议2025

Cross-Modal Robotic Perception for Physical Property Inference

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

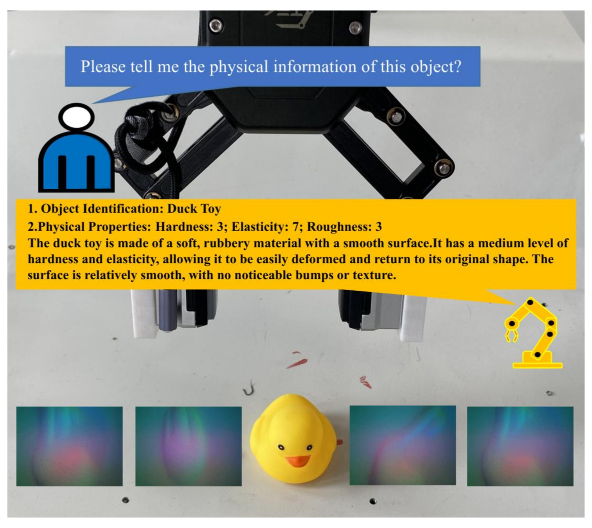
[1] College of Big Data and Internet, Shenzhen Technology University, China,

[2] Sino-German College of Intelligent Manufacturing, Shenzhen Technology University, China,

[3] Volkswagen Group, Machine Learning Research Lab, 80805, Munich, DE

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

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 according to the rules.

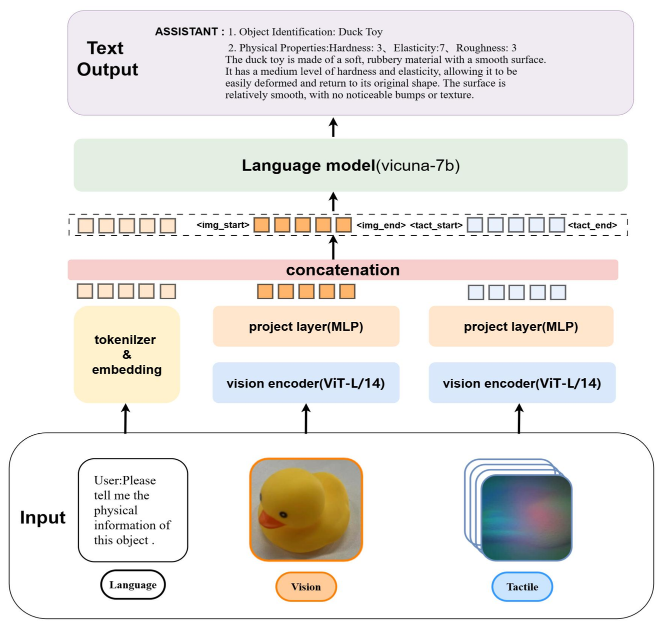


ABSTRACT

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 thatintegrates visual observations with tactile representations within a multi-modal vision-language model. We physical reasoning framework that em-ploys a hierarchical feature alignment mechanism and a refined prompt-ing strategy, our model has property-specific predictions that stronglycorrelate with ground-truth measurements. Evaluated on a dataset of 30diverse objects, our approach outperforms existing baselines.

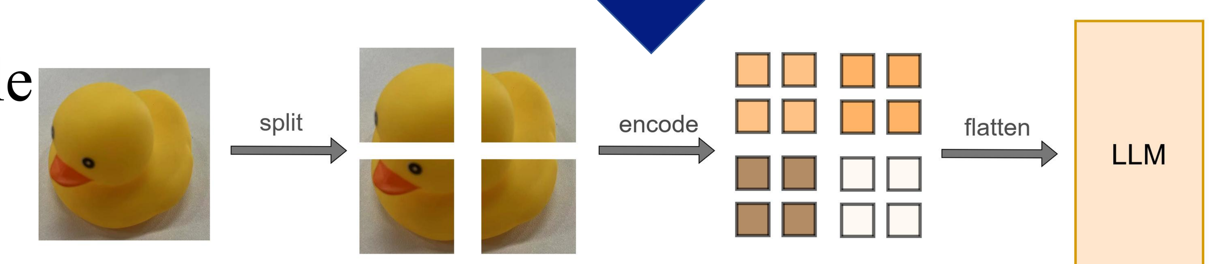
METHODOLOGY

Vision Processing



The architecture of a multimodal large model. After embedding and tokenizing the object image and tactile image alongside the text, the resulting vectors are con catenated and input into the large language model. This enables the model to interpret diverse inputs.

The image is first segmented into multiple regions using the segmentation module.



The encoder then extracts a feature matrix, which is subsequently flattened into a one-dimensional vector. Finally, this processed representation is fed into the large language model (LLM) for semantic analysis and reasoning.

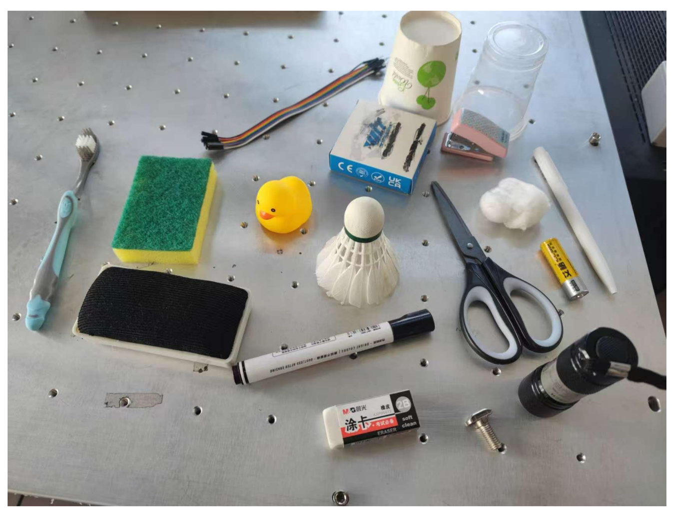
EXPERIMENTS

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

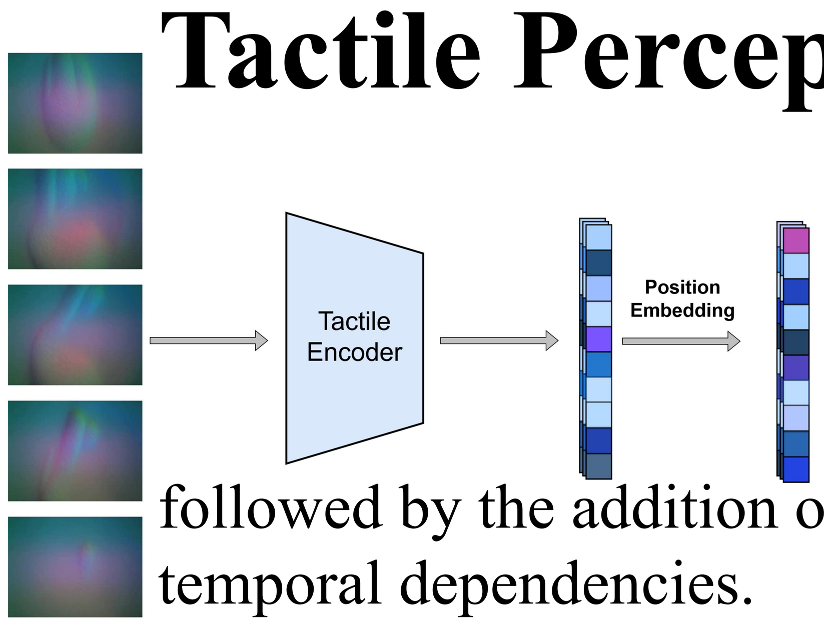
Table 2. Comparison of Correlation Coefficients Between Models and Ground Truth

Attribute	Method	Correlation Coefficient	P-value
Hardness	Our Model	0.501*	0.005**
	Octopi	0.307	0.099
	Octopi (3 levels)	0.307	0.099
Elasticity	Our Model	0.530*	0.003**
	Octopi	0.053	0.781
	Octopi (3 levels)	−0.060	0.753
Roughness	Our Model	0.643*	0.0001**
	Octopi	−0.010	0.959
	Octopi (3 levels)	0.118	0.534



Experimental evaluations on 30 diverse objects show that our approach significantly outperforms baseline methods. Our model achieves Spearman coefficients of 0.501 for hardness, 0.530 for elasticity, and 0.643 for roughness, showing improvements in alignment with ground-truth measurements compared to existing approaches.

Tactile Perception



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.

CONCLUSION

We presented a novel approach to enhance tactile perception through visual compensation and optimized prompt engineering. Our method addresses key limitations of tactile-only systems by incorporating visual information and structuring language model interactions more effectively. Experimental results demonstrate significant improvements in physical property inference, with particularly strong performance in roughness estimation. The success of our approach highlights the importance of compensating for tactile sensory limitations through complementary visual information and carefully designed language model prompts. Future work will extend this framework to robotic grasping applications, where multimodal tactile-visual reasoning could enable adaptive manipulation of objects with different material properties.

深圳技术大学 SHENZHEN TECHNOLOGY UNIVERSITY SZTU

大数据与互联网学院 COLLEGE OF BIG DATA AND INTERNET

中德智能制造学院 SINO-GERMAN COLLEGE OF INTELLIGENT MANUFACTURING

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

2 0 2 5年 4月 18 – 2 0 日 江西 南昌