Scene Understanding in Deformable Object Manipulation via Taxonomy-Guided Vision-Language Models

Gawtam Chithra Ramesh[1,*], David Blanco-Mulero[2,*], Yifei Dong[1],
Júlia Borràs[2], Carme Torras[2], Florian T. Pokorny[1]

***Abstract*— Vision-Language Models (VLMs) can describe scenes in natural language, supporting tasks such as robot planning and action grounding. However, they struggle in deformable object manipulation (DOM), where reasoning about motion, interaction, and deformation is critical. In this work, we investigate whether guiding language models with a taxonomy for DOM can provide a structured reasoning about DOM tasks. We evaluate the performance of our approach on three challenging DOM tasks: towel twisting, meat phantom transport, and cloth edge tracing. Our results demonstrate the potential of taxonomy-guided VLMs to interpret these tasks without fine-tuning or curated datasets.**

I. Introduction

Recent advances in Vision-Language Models (VLMs) have enabled robots to interpret images and scenes in natural language, supporting tasks such as scene understanding for robot manipulation [1], or task and motion planning (TAMP) [2]. By leveraging vision and language, VLMs can describe the environments that robots interact with, parse textual commands, and translate these into action commands [3]. However, VLMs often fail in tasks that require deeper physical reasoning. To extend the capabilities of these models, one recent solution is to perform fine-tuning on datasets that incorporate physical concepts [4]. Nevertheless, these approaches require substantial computational resources and carefully curated datasets for large-scale fine-tuning. It limits their applicability to challenging tasks like deformable object manipulation (DOM), where measuring the internal state of deformable objects is challenging [5].

To systematically describe manipulation tasks, prior works have proposed several taxonomies of manipulation. These taxonomies provide a structure for key manipulation aspects such as motion at contact [6], robot trajectories [7], and deformation [8]. By providing structured, interpretable categories, these taxonomies can enable systematic reasoning about robot actions and object states.

In this extended abstract, we investigate whether these taxonomy frameworks can guide VLM reasoning, see Fig. 1. Specifically, we evaluate whether VLMs guided by the Taxonomy for Deformable Object Manipulation (T-DOM) [8] can generate structured scene descriptions of DOM tasks, which may support downstream applications such as TAMP. We evaluate this by comparing VLM outputs against human expert annotations using (i) BERT-based similarity for free-text descriptions, and (ii) Hamming distance for T-DOM taxonomy codes. Our results demonstrate that taxonomy-guided VLMs can produce structured and interpretable analyses of DOM tasks, while also revealing limitations in handling subtle deformations and occlusions.

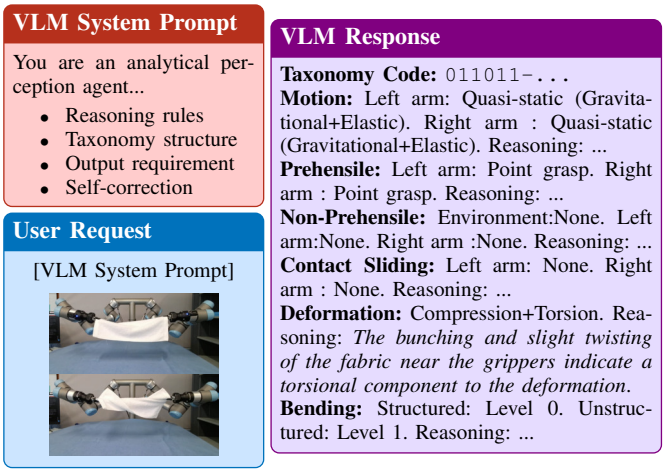


Fig. 1: Example of a VLM response guided by Taxonomy for Deformable Object Manipulation (T-DOM) [8] for analyzing a cloth manipulation scene, providing a structured output for the categories: motion, prehensile grasp, non-prehensile interaction, contact sliding, deformation, and bending.

II. Guiding VLMs with a Taxonomy for DOM

To investigate the scene understanding capabilities of taxonomy-guided VLMs in the context of DOM, we provide a visual observation of a manipulation scene (see Fig. 1). The visual input provides information on (i) the scene prior to manipulation, and (ii) the resulting manipulation state to be analyzed. The visual input is accompanied by a system prompt that incorporates the T-DOM categories—robot motion, prehensile grasp, non-prehensile environment and agent interactions, contact sliding, object deformation, and bending. In addition, it incorporates essential context, such as camera and robot coordinate conventions, as well as reasoning rules derived from the taxonomy, which help to eliminate ambiguity in spatial reasoning. The complete system prompt is provided in our project website.

∗ - Equal contribution. The authors are with: [1]the Division of Robotics, Perception and Learning, KTH Royal Institute of Technology, Stockholm, Sweden; [2]Institut de Robòtica i Informàtica Industrial, CSIC-UPC, Barcelona, Spain. Funded by the European Commission under the Horizon Europe Framework Programme project SoftEnable, grant number 101070600. Contact: gawtam@kth.se, david.blanco.mulero@upc.edu.

Experiment videos and other supplementary materials are available at: https://sites.google.com/view/tax-guided-vlm

To constrain the VLM output to the structured reasoning provided by the taxonomy, the system prompt requests to output with the following format: (i) a 28-bit binary string code that classifies the manipulation state across the six T-DOM categories; (ii) concise textual justification for each classification; and (iii) a self-correction step, cross-checking that its generated binary code is consistent with its textual justifications.

III. Experiments

Our experiments aim to answer the following:

- Can taxonomy-guided VLM achieve performance comparable to human expert annotations when describing DOM tasks?
- Do VLMs provide reasonable analyses of both the robot action and the resulting object state?

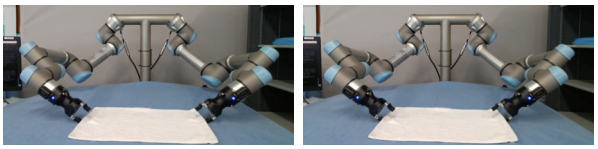
A. Tasks and Baselines

To evaluate the VLMs' capabilities of DOM scene understanding, we reused a subset of three real-world tasks from the T-DOM dataset [8]. The selected tasks cover different types of objects, motions, interactions, and deformations: *task 1* performs the twisting of a towel, *task 2* transports a meat phantom object, and *task 3* performs edge tracing for a piece of cloth. The tasks were performed by a bimanual robotic platform, with a fixed camera capturing the manipulation workspace. The resulting visual data was manually segmented into distinct states representing different phases of the manipulation (e.g., approaching, grasping, twisting). The selected frames were prepared as inputs for the models, ensuring the manipulation scenes were not occluded. Fig.2 shows some examples of the images. To assess the performance of our taxonomy-guided approach, we selected three models: Gemini2.5Pro[9], Qwen2.5-VL-32B and Qwen2.5-VL-72B[10]. This choice allows for a comparative analysis across model architectures as well as model scale.

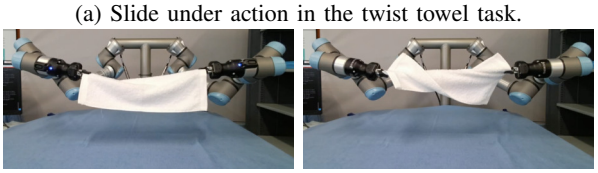
B. Evaluation Metrics

To evaluate the model accuracy, we create a human-labeled ground-truth (GT) for each representative frame, in accordance with the T-DOM framework [8]. To comprehensively assess the model performance, we employed two metrics, one for the binary code and one for the language reasoning.

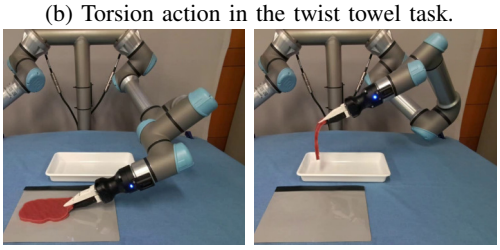
Since the taxonomy classifies each state using a binary code, for the VLM-generated binary codes, we used the Hamming distance. This directly measures the classification error by counting differing bit positions between the predicted and the GT codes. Lower values indicate higher accuracy in classifying the manipulation state according to the taxonomy categories. For the unstructured text, we used the BERT score, which leverages contextual embeddings to measure semantic similarity between the model explanations and the GT reasoning. Unlike word-overlap metrics, it can capture correct explanations even when phrased differently compared to the reference text. By using both metrics, we can distinguish between a model's ability to correctly classify a state, measured by Hamming distance, and its ability to coherently explain why it made that classification, measured by the BERT score.



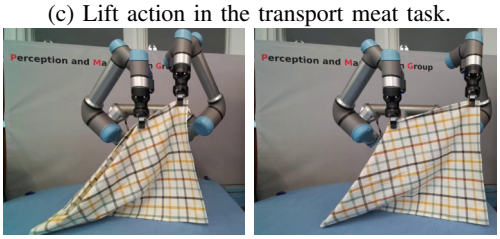
(a) Slide under action in the twist towel task.



(b) Torsion action in the twist towel task.



(c) Lift action in the transport meat task.



(d) Motion with slippage in the cloth edge tracing task.

Fig. 2: Example states provided as input to the VLM.

C. Evaluation and Analysis

Due to space limitations, detailed task analyses, including full tables and figures, are provided in the Supplementary Material and on our project website.

Our quantitative evaluation, summarized in Table I, provides several insights into the performance of the selected VLMs. In terms of textual reasoning (BERT score), Gemini2.5Pro achieves the highest average score, suggesting its explanations are often well-aligned with the ground truth, particularly for prehensile and non-prehensile interactions in the towel and meat tasks. However, the Qwen models show competitive performance, sometimes outperforming Gemini in specific categories, such as in task 3. For classification accuracy (Hamming distance), the models performed similarly on average. Notably, the smaller Qwen-VL-32B model often performs on par with or sometimes better than its larger counterpart and Gemini, indicating that model scale does not consistently correlate with higher classification accuracy in this structured task.

Beyond quantitative results, we also conducted a qualitative analysis to better understand the strengths and limitations of taxonomy-guided reasoning.We observed several recurring patterns across all experiments. A primary challenge for the

TABLE I: Quantitative results of taxonomy-guided VLMs for the three DOM tasks averaged over categories.

	BERT ↑			HAMMING ↓		
	Gemini2.5Pro	Qwen-VL-72B	Qwen-VL-32B	Gemini2.5Pro	Qwen-VL-72B	Qwen-VL-32B
Motion	**0.67**±0.15	0.58±0.16	0.56±0.16	**0.18**±0.15	0.20±0.10	0.25±0.12
Prehensile Interaction	**0.64**±0.07	0.55±0.14	0.63±0.12	**0.08**±0.14	**0.08**±0.12	0.19±0.17
Non P. Interaction	**0.69**±0.08	0.60±0.08	0.65±0.10	0.24±0.17	**0.18**±0.12	0.19±0.12
Contact Sliding	0.67±0.13	**0.69**±0.17	0.64±0.13	0.16±0.12	**0.15**±0.11	0.16±0.12
Deformation	0.56±0.18	0.52±0.19	**0.58**±0.23	0.14±0.14	0.16±0.10	**0.10**±0.09
Bending	0.62±0.15	**0.64**±0.19	**0.64**±0.21	0.20±0.14	0.25±0.13	**0.16**±0.11
Average	**0.64**±0.13	0.60±0.16	0.62±0.16	**0.17**±0.14	**0.17**±0.11	0.18±0.12

VLMs was interpreting states characterized by subtle visual cues or occlusions. For instance, identifying contact sliding was consistently difficult, as the corresponding pixel changes were often too small to capture. Similarly, pre-grasp motions were often misclassified as a completed prehensile grasp because the gripper occluded key features of the interaction.

In addition, we provide a qualitative analysis of specific cases, with the corresponding robot actions depicted in Fig. 2. The full VLM outputs for these and all other cases are available in the supplementary material; here, we highlight key observations. For the sliding action in the towel twisting task (Fig. 2b), the VLM identified deformation as compression and tension, while the ground-truth (GT) was tension and torsion. Both interpretations are physically plausible, yet this semantic ambiguity leads to a high Hamming distance of 0.5, incorrectly suggesting a model failure. In the meat transport task (Fig. 2c), the VLM classified the motion as dynamic, which is reasonable given that only the start and end frames were provided. However, the GT labeled it as quasistatic, as the human annotator possessed knowledge of the robot's speed. In this example, the VLM was able to apply the taxonomy principles to correctly identify tension due to the object elongating, whereas the GT focused only on the compression from the gripper. This highlights a fundamental information asymmetry and shows the VLM's ability to reason from visual evidence. Finally, in the cloth tracing task (Fig. 2d), the VLMs correctly predicted the level of structured bending but struggled to classify the unstructured bending, which is more ambiguous. The accurate classification across other complex categories for this task demonstrates how the taxonomy provides a structured analysis strategy, enabling the VLM to parse these challenging DOM tasks.

IV. Discussion

Our results demonstrate that a taxonomy can impose a structured and interpretable semantic layer on VLM outputs, enabling them to systematically analyze DOM scenes without fine-tuning. This highlights the potential of taxonomy-guided VLMs to facilitate annotation and support scene understanding, thereby reducing the need for exhaustive human labeling. Such capabilities open a promising direction for applying taxonomy-guided VLMs in robotic data analysis and motion planning.

At the same time, we have discussed some limitations of the approach. Discrepancies with ground-truth often arose from task ambiguity (e.g., distinguishing torsion from tension), or information asymmetry between model output and annotation. These cases suggest that differences should not always be interpreted as failure modes, but rather as signals to improve both models and human labels.

A key direction for future work is to demonstrate how the quality of the VLM's structured output impacts downstream robotic tasks. This highlights a key challenge: while a comparison to an 'unguided' VLM seems logical, a truly unguided model's free-form text is not directly robot-parsable. To be useful, any VLM output requires some structure, which is itself a form of implicit taxonomy. The central hypothesis is therefore centered on the value of semantic translation: that a detailed, domain-aware structure like T-DOM enables a more effective conversion of visual perception into executable robotic actions compared to minimally structured formats. Future work will focus on validating this by comparing the utility of T-DOM-guided outputs against those from VLMs prompted for simpler, generic formats in automating robotic task and motion planning.

References

[1] S. Liu, J. Zhang, R. X. Gao, X. Vincent Wang, and L. Wang, "Vision-language model-driven scene understanding and robotic object manipulation," in *2024 IEEE 20th International Conference on Automation Science and Engineering (CASE)*, 2024, pp. 21–26.

[2] K. Kawaharazuka, J. Oh, J. Yamada, I. Posner, and Y. Zhu, "Vision-language-action models for robotics: A review towards real-world applications," *IEEE Access*, vol. 13, pp. 162 467–162 504, 2025.

[3] J. Duan, W. Yuan, W. Pumacay, Y. R. Wang, K. Ehsani, D. Fox, and R. Krishna, "Manipulate-anything: Automating real-world robots using vision-language models," in *Proceedings of The 8th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, P. Agrawal, O. Kroemer, and W. Burgard, Eds., vol. 270. PMLR, 06–09 Nov 2025, pp. 5326–5350.

[4] J. Gao, B. Sarkar, F. Xia, T. Xiao, J. Wu, B. Ichter, A. Majumdar, and D. Sadigh, "Physically grounded vision-language models for robotic manipulation," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 12 462–12 469.

[5] H. Yin, A. Varava, and D. Kragic, "Modeling, learning, perception, and control methods for deformable object manipulation," *Science Robotics*, vol. 6, no. 54, p. eabd8803, 2021.

[6] I. M. Bullock, R. R. Ma, and A. M. Dollar, "A hand-centric classification of human and robot dexterous manipulation," *IEEE Transactions on Haptics*, vol. 6, no. 2, pp. 129–144, 2013.

[7] D. Paulius, N. Eales, and Y. Sun, "A Motion Taxonomy for Manipulation Embedding," in *Proceedings of Robotics: Science and Systems*, Corvalis, Oregon, USA, July 2020.

[8] D. Blanco-Mulero, Y. Dong, J. Borras, F. T. Pokorny, and C. Torras, "T-dom: A taxonomy for robotic manipulation of deformable objects," *arXiv preprint arXiv:2412.20998*, 2024.

[9] Deepmind, "Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities," *arXiv preprint arXiv:2507.06261*, 2025.

[10] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, and J. Lin, "Qwen2.5-vl technical report," *arXiv preprint arXiv:2502.13923*, 2025.