Supplementary Material for the paper "Scene Understanding in Deformable Object Manipulation via Taxonomy-Guided Vision-Language Models"

Table II reports the performance of all evaluated VLMs across the three DOM tasks, with BERT measuring textual reasoning quality and Hamming distance capturing taxonomy code accuracy.

The figures Fig. 3, Fig. 4, Fig. 5, and Fig. 6, provide illustrative examples of VLM responses guided by T-DOM for the three analyzed DOM tasks, including towel twisting (sliding and twisting actions), meat transport, and cloth edge tracing.

TABLE II: Quantitative Results of Taxonomy-guided VLMs on three DOM tasks

	BERT ↑			HAMMING \downarrow		
	Gemini2.5Pro	Qwen-VL-72B	Qwen-VL-32B	Gemini2.5Pro	Qwen-VL-72B	Qwen-VL-32B
Task 1 - Twisting Towel						
Motion	0.68 ±0.24	0.49 ± 0.23	0.51±0.24	0.10 ±0.16	0.17±0.17	0.32±0.23
Prehensile Interaction	0.69 ± <u>0.09</u>	0.57 ± 0.17	0.68 ± 0.16	0.13 ± 0.22	0.13 ± 0.22	0.24 ± 0.19
Non P. Interaction	0.76 ± 0.08	0.55 ± 0.14	0.67 ± 0.18	0.23 ± 0.28	0.17 ± 0.21	0.17 ± 0.21
Contact Sliding	0.62 ± 0.14	0.72 ± 0.30	0.59 ± 0.20	0.13 ± 0.22	0.13 ± 0.22	0.13 ± 0.22
Deformation	0.63 ± 0.26	0.77 ± 0.18	0.66 ± 0.17	0.20 ± 0.24	0.06 ± 0.11	0.06 ± 0.11
Bending	0.52 ± 0.18	0.68 ± 0.25	0.63 ± 0.24	0.25 ± 0.20	$0.06 \pm \underline{0.11}$	0.03 ± 0.08
Task 2 - Transporting Meat						
Motion	0.65 ±0.12	0.56±0.20	0.59±0.19	0.06 ±0.10	0.11±0.13	0.10±0.12
Prehensile Interaction	0.59 ± 0.08	0.52 ± 0.18	0.48 ± 0.13	0.07 ± 0.12	0.10 ± 0.13	0.21 ± 0.18
Non P. Interaction	0.61 ± 0.08	0.46 ± 0.07	0.51 ± 0.08	0.17 ± 0.18	0.21 ± 0.14	0.24 ± 0.14
Contact Sliding	0.74 ± 0.20	0.81 ± 0.17	0.80 ± 0.15	0.11 ± 0.13	0.08 ± 0.12	0.10 ± 0.15
Deformation	0.37 ± 0.22	0.58 ± 0.37	$0.45\pm \overline{0.30}$	0.21 ± 0.18	0.17 ± 0.19	0.24 ± 0.16
Bending	0.68 ± 0.20	0.77 ± 0.23	0.64 ± 0.30	0.01 ± <u>0.06</u>	0.06 ± 0.14	0.07 ± 0.12
Task 3 - Cloth Edge Tracing						
Motion	0.69±0.08	0.70 ±0.05	0.58±0.06	0.37±0.19	0.33 ±0.00	0.33 ±0.00
Prehensile Interaction	0.65 ± 0.05	$0.55 \pm \overline{0.06}$	0.72 ± 0.08	0.03 ± 0.08	$0.00 \pm \overline{0.00}$	$0.13 \pm \overline{0.13}$
Non P. Interaction	$0.69 \pm \overline{0.07}$	0.78 ± 0.02	0.78 ± 0.05	0.32 ± 0.05	$0.17 \pm \overline{0.00}$	0.17 ± 0.00
Contact Sliding	0.65 ± 0.04	$0.55 \pm \overline{0.05}$	$0.54 \pm \underline{0.03}$	0.25 ± 0.00	0.25 ± 0.00	0.25 ± 0.00
Deformation	0.69 ± 0.05	0.21 ± 0.01	0.64 ± 0.23	0.00 ± 0.00	0.25 ± 0.00	0.00 ± 0.00
Bending	0.65 ± 0.08	$0.48 \pm \overline{0.10}$	0.64 ± 0.09	0.33 ± 0.17	0.63 ± 0.13	0.38 ± 0.13



VLM Response

Taxonomy Code:

100000-0100-000000-0000-0100-0001

Motion: Left arm: None. Right arm: Dynamic. Reasoning: Right arm lifts object with fast motion. Left arm static. No sliding detected.

Prehensile: Left arm: None. Right arm: Point grasp. Reasoning: Right gripper visibly constrains object. Left arm not in contact.

Non-Prehensile: Environment: None. Active Left arm: None. Active Right arm: None. Reasoning: No other contacts visible besides grasp.

Contact Sliding: Left arm: No Sliding. Right arm: No Sliding. Reasoning: No relative motion at contact points.

Deformation: Type(s): Tension. Reasoning: Object elongates as it's lifted, showing tension.

Bending: Structured: Level 0. Unstructured: Level 0. Reasoning: Minor curvature from lifting, no complex folds.

Ground Truth

Taxonomy Code:

010000-0100-000000-0000-1000-0100

Motion: Left arm: Quasistatic Potential motion upwards and inwards. Right arm: None. Reasoning: The left arm moved upwards to lift the meat and inwards to bring it above the white tray.

Prehensile: Left arm: Point Grasp. Right arm: No Grasp. Reasoning: the left gripper continue to maintain the grasp on the meat.

Non-Prehensile: Environment: None. Active Left arm: None. Active Right arm: None. Reasoning: The meat is lifted off from the table using the point grasp. **Contact Sliding:** Left arm: None. Right arm: None. Reasoning: There is no relative motion anywhere.

Deformation: Compression. Reasoning: The meat is being squeezed by the left gripper.

Bending: Structured: Level 0. Unstructured: None. Reasoning: The meat is lifted on one end creating a slight bend.

Fig. 3: Example of a VLM response guided by T-DOM for analyzing the transport object action in the meat transport task.

VLM Response

Taxonomy Code:

011011-0101-000000-0000-1010-0110

Motion: Left arm: Quasistatic / Gravitational + Elastic. Right arm: Quasistatic / Gravitational + Elastic. Reasoning: ... Prehensile: Left arm: Point grasp. Right arm: Point grasp. Reasoning: Both the robot's left (image-right) and right (image-left) grippers are visibly closed on the towel, maintaining a stable point grasp from the reference state.

Non-Prehensile: Environment: None. Active Left arm: None. Active Right arm: None. Reasoning: The towel is held suspended in the air by both grippers and does not make contact with the table or any other part of the environment or robot.

Contact Sliding: Left arm: No Sliding. Right arm: No Sliding. Reasoning: The grippers appear to maintain a firm grasp on the same points of the towel as in image_1. Although the material has bunched up, no relative motion between the gripper fingers and the cloth surface is evident.

Deformation: Type(s): Compression + Torsion. Reasoning: The inward motion of the arms has caused clear horizontal compression, resulting in buckling. The bunching and slight twisting of the fabric near the grippers indicate a torsional component to the deformation.

Bending: Structured: Level 0. Unstructured: Level 1. Reasoning: The towel now has a single, large, simple fold in the middle, characteristic of Structured Level 0 bending. Moderate uncontrolled wrinkles have also formed near the grasp points, corresponding to Unstructured Level 1.

Ground Truth

Taxonomy Code:

011011 - 0101 - 000000 - 0000 - 0110 - 0000

Motion: Left arm: Quasistatic Potential rotation + Quasistatic Elastic. Right arm: Quasistatic Potential rotation + Quasistatic Elastic. Reasoning: ... Prehensile: Left arm: Point Grasp. Right arm: Point Grasp. Reasoning: both the grippers have a point grasp of the towel.

Non-Prehensile: Environment: None. Active Left arm: None. Active Right arm: None. Reasoning: The towel is hanging above the table using the point grasps.

Contact Sliding: Left arm: None. Right arm: None. Reasoning: There is no relative motion anywhere.

Deformation: Tension + Torsion. Reasoning: The towel is being twisted inducing torsion along with the existing tension due to gravity.

Bending: Structured: None. Unstructured: None. Reasoning: No Signification bending observed.



User Request [System Prompt for VLM]

User Request

[System Prompt for VLM]

VLM Response

Taxonomy Code:

000000-0101-010000-0000-0000-0001

Motion: Left arm: None. Right arm: None. Reasoning: A comparison between image_1 and image_2 reveals no discernible change in the position or orientation of either robot arm or the object. The entire system is static.

Prehensile: Left arm: Point grasp. Right arm: Point grasp. Reasoning: The robot's left (image-right) and right (image-left) grippers are both visibly closed around the corners of the white towel, maintaining a stable point grasp.

Non-Prehensile: Environment: Soft surface contact. Active Left arm: None. Active Right arm: None. Reasoning: The white towel rests on a blue cloth, which is a soft environmental surface. Both arms are engaged in a prehensile grasp, with no non-prehensile contact.

Contact Sliding: Left arm: No Sliding. Right arm: No Sliding. Reasoning: The scene is static. There is no relative motion between the grippers and the towel at the points of contact.

Deformation: Type(s): None. Reasoning: The towel is held in a resting state without any actively induced deformation. There are no visual indicators of compression, tension, torsion, or shearing.

Bending: Structured: None. Unstructured: Level 0. Reasoning: The towel has no distinct, structured folds. It does exhibit minor, uncontrolled wrinkles and a gentle sag consistent with its draped state, which corresponds to Unstructured Level 0.

Ground Truth

Taxonomy Code:

 $\tt 010010-0000-101010-0101-0000-0000$

Motion: Left arm: Quasistatic Potential motion to right. Right arm: Quasistatic Potential motion to left. Reasoning: The arms moved towards and under the towel preparing for a possible point grasp.

Prehensile: Left arm: No Grasp. Right arm: No Grasp. Reasoning: Both arms are positioned to grasp. **Non-Prehensile:** Environment: Rigid Surface. Active Left arm: Rigid Surface. Active Right arm: Rigid Surface. Reasoning: The towel is resting on the table, and both arms are in contact with the table.

Contact Sliding: Left arm: Active Sliding. Right arm: Active Sliding. Reasoning: The arms are sliding under the towel.

Deformation: None. Reasoning: No Signification deformation observed.

Bending: Structured: None. Unstructured: None. Reasoning: No Signification bending observed.

Fig. 5: Example of a VLM response guided by T-DOM for analyzing the sliding action in the towel twisting task.

User Request [System Prompt for VLM] Perception and Mary Group

VLM Response

Taxonomy Code:

000010-0101-010000-0000-0100-0010

Motion: Left arm: None (Static). Right arm: Quasistatic / Gravitational potential energy. Reasoning: The left arm (image-right) is stationary between images. The right arm (image-left) has moved, indicating at least quasistatic motion against gravity.

Prehensile: Left arm: Point grasp. Right arm: Point grasp. Reasoning: In image_2, both the left (image-right) and right (image-left) grippers are visibly closed on the cloth, constraining it at two distinct points.

Non-Prehensile: Environment: Soft surface contact. Active Left arm: None. Active Right arm: None. Reasoning: The lower portion of the cloth rests on the blue, soft-looking table surface. Both arms interact with the cloth only via their prehensile grasps.

Contact Sliding: Left arm: No Sliding. Right arm: No Sliding. Reasoning: The grasps on both arms appear stable. The cloth's texture does not seem to shift relative to the gripper fingers between image_1 and image_2, indicating no sliding at the contact points. Deformation: Type(s): Tension. Reasoning: The cloth is held at two points and hangs under its own weight,

Deformation: Type(s): Tension. Reasoning: The cloth is held at two points and hangs under its own weight, creating tension. The movement of the right arm (image-left) has slightly increased the separation between grippers, enhancing this tension.

Bending: Structured: Level 0. Unstructured: Level 1. Reasoning: There are no sharp, structured folds. The cloth exhibits moderate, uncontrolled draping and curvature due to gravity, consistent with Unstructured Level 1.

Ground Truth

Taxonomy Code:

011000-0101-100000-0010-0100-0111

Motion: Left arm: Quasistatic Potential motion to left + Quasistatic Elastic. Right arm: None. Reasoning: The left arm pulls the cloth to the left while the right arm remains static. The elastic forces of the cloth also contribute to the motion.

Prehensile: Left arm: Point Grasp. Right arm: Point Grasp. Reasoning: Both arms have a point grasp on the cloth.

Non-Prehensile: Environment: Rigid Surface. Active Left arm: None. Active Right arm: None. Reasoning: Even though the cloth is freely hanging, it is still in contact with the table, whereas both arms are not in contact with the table.

Contact Sliding: Left arm: None. Right arm: Passive Sliding. Reasoning: The cloth is passively sliding at the grasp of the right arm as the left arm pulls the cloth to the left.

Deformation: Tension. Reasoning: The cloth is being pulled creating tension along the horizontal direction. **Bending:** Structured: Level 0. Unstructured: Level 2. Reasoning: The cloth is bent around the left grasp point creating a simple level 0 fold, and the rest of the cloth is bent in an unstructured manner creating level 2 bends.

Fig. 6: Example of a VLM response guided by T-DOM for analyzing the action of tracing in the cloth edge tracing task.