

A Supplemental Material

A.1 Limitations

While VTV-LLM demonstrates promising capabilities for visuo-tactile understanding, several limitations should be acknowledged. First, our method is constrained by the diversity of tactile sensors used (GelSight Mini, DIGIT, and Tac3D), which may limit generalization to other sensor types with different data formats or physical properties. Second, the VTV150K dataset, while extensive, only covers 100 objects, representing a limited subset of real-world materials and textures. This may impact performance on novel objects with unique tactile properties. Third, computational requirements increase significantly with model size (particularly with the 14B parameter variant), potentially limiting deployment on resource-constrained robotic platforms. Finally, our current implementation focuses primarily on short-term tactile interactions rather than extended temporal sequences, which may be necessary for more complex manipulation tasks requiring long-term tactile memory.

A.2 Broader Impacts

This research has several potential positive societal impacts, including enhancing human-robot interaction through improved tactile understanding, enabling more sophisticated assistive technologies for visually impaired individuals, and advancing industrial applications in manufacturing and quality control. By bridging the gap between tactile perception and natural language, our work could facilitate more intuitive communication about physical properties between humans and embodied agents. However, there are also potential negative implications to consider. The development of advanced tactile sensing technologies could raise privacy concerns if tactile signatures could be used to identify individuals or sensitive objects. Additionally, the specialized hardware requirements might create access inequalities if the technology remains expensive and inaccessible to broader communities. As with many AI technologies, there is also potential for automation that could impact employment in fields requiring skilled tactile assessment. We encourage responsible development that considers these ethical implications.

A.3 Dataset: VTV150K

A.3.1 Attribute Annotation

In our experimental protocol, we established a systematic annotation framework for both static and dynamic object attributes with precisely defined classification levels. For hardness, we designated level 0 to indicate objects where force application produces large deformation, level 1 for slight deformation, and level 2 for negligible deformation. Protrusion was classified as level 0 (no surface protrusion), level 1 (shallow surface protrusion), and level 2 (deep surface protrusion). Elasticity was categorized as level 0 (non-elastic surface), level 1 (minimal rebound after compression), and level 2 (complete rebound to original position). Friction was evaluated as level 0 (minimal or no friction), level 1 (moderate resistance during surface sliding), and level 2 (significant resistance impeding surface sliding). The annotation process involved three independent annotators evaluating tactile videos of the objects, with final classifications determined by majority consensus. The comprehensive annotation results are presented in Tab. 4.

A.3.2 Template Generation

In the template generation phase, we formulated diverse problem templates addressing multiple reasoning domains for training, including tactile feature assessment (TFA), surface feature distinction (SFD), surface optimality identification (SOI), and object sensation correlation (OSC). These question templates are comprehensively presented in Tab. 5. Such a structured approach ensures systematic coverage of various tactile reasoning capabilities required for robust model performance.

A.4 More Experimental Results

A.4.1 Implementation Details

For VTV enhancement, we employ the raft-things checkpoint [51] to extract optical flow. The model was trained for 200 epochs utilizing the AdamW optimizer [61] with a learning rate of 1e-3, batch

Table 4: Attribute annotation ratings across different objects.

Object	Hardness	Protrusion	Elasticity	Friction	Object	Hardness	Protrusion	Elasticity	Friction
tangerine	0	1	1	2	carton	2	0	0	1
playing card	2	0	0	0	orange	1	1	1	2
rice spoon inside	1	2	2	2	nectarine	0	0	1	2
tomato	0	0	1	0	rubber glove	1	0	1	2
pen pad	1	0	2	0	cotton cloth	1	0	0	1
velvet	1	0	0	0	sandpaper	2	1	0	2
chip bag	2	0	1	2	rubber slipper sole	1	2	2	2
candle	2	0	0	0	suede	1	0	1	2
silk scarf	1	0	0	0	fascia ball	1	0	2	2
chalk	2	0	2	0	clay	0	0	0	0
avocado	1	2	1	1	lemon	1	1	1	1
banana	0	0	1	2	kiwi	0	1	1	1
pineapple	1	2	1	2	plastic bottle	0	0	2	0
waffle	0	2	1	1	sponge	0	1	2	2
plastic basket	0	0	2	0	balloon	0	0	2	2
leather glove	1	0	1	2	building block	2	2	0	0
jelly	0	0	2	0	piano key	2	0	0	0
blanket	0	0	1	0	ceramic cup	2	0	0	0
oven glove	0	1	1	1	bark	2	2	0	2
scouring pad	0	1	2	2	fur	0	1	1	0
pine cone	0	2	2	2	ping pong ball	2	0	0	1
plastic building block	2	2	0	0	cork	2	1	0	2
wooden ruler	2	0	0	0	eraser	1	0	2	0
velcro	1	2	2	2	leather wallet	1	1	1	0
toilet paper	1	1	1	1	shower mat	1	1	2	1
baseball	1	0	2	2	golf ball	2	2	0	0
sticky note	1	0	1	1	silicone pad	1	0	2	2
yoga mat	0	0	2	1	masking tape	1	1	1	2
rubber band	0	2	2	2	cotton ball	0	0	1	0
gauze	1	0	1	0	mouse	2	0	0	1
headphone	2	0	0	0	face towel	1	1	0	1
woven watch strap	1	2	1	1	rubber watch strap	1	0	1	0
metal watch strap	2	1	0	2	wooden block	2	0	0	2
marble	2	0	0	2	claw inside	0	2	2	2
keyboard	0	2	2	2	remote button	0	2	2	2
toothbrush head	0	2	2	1	vitamin table	2	2	0	1
tennis ball	1	1	2	1	towel	1	1	1	1
absorbent cloth	2	1	0	2	wrist guard	1	1	2	1
fine bubble film	0	1	1	2	coarse bubble film	0	1	1	1
rice	2	2	0	2	ridge cup	2	2	0	2
key	2	2	0	2	screw	2	2	0	2
circuit board	2	2	0	2	mold	2	2	0	0
sponge sheet	1	1	2	2	wrench	2	2	0	0
screw knife handle	2	2	1	2	aluminum tube	2	1	0	2
wire	2	1	0	0	injection tube	1	0	2	0
tape	2	0	0	0	iron clip	2	2	0	1
iron ruler	2	0	0	0	scissor	2	2	0	0
hairpin	1	2	2	1	steel wool	0	2	1	2
huamei	2	2	0	2	candy	2	2	0	1
grid bag	2	1	0	2	ridge plastic bottle	0	2	2	1

Table 5: Question templates across different tasks. ' $\langle \text{comparison} \rangle \langle \text{attribute} \rangle$ ' represents the comparative and superlative level of the attribute.

Task	Template
TFA	"Can you detail the surface characteristics shown in this video $\langle \text{video} \rangle$?"
TFA	"What are the tactile features of the object presented in the video $\langle \text{video} \rangle$?"
TFA	"Describe the physical properties of the object in the video $\langle \text{video} \rangle$."
TFA	"How does the object in this tactile video $\langle \text{video} \rangle$ feel?"
SFD	"I have tactile videos of two objects. Which one is $\langle \text{comparison} \rangle \langle \text{attribute} \rangle$? $\langle \text{video1} \rangle$, $\langle \text{video2} \rangle$."
SFD	"Between these two videos, $\langle \text{video1} \rangle$ and $\langle \text{video2} \rangle$, which object feels $\langle \text{comparison} \rangle \langle \text{attribute} \rangle$?"
SFD	"Comparing the objects in $\langle \text{video1} \rangle$ and $\langle \text{video2} \rangle$, which one is $\langle \text{comparison} \rangle \langle \text{attribute} \rangle$?"
SFD	"Is the object in the first video $\langle \text{video1} \rangle$ $\langle \text{comparison} \rangle \langle \text{attribute} \rangle$ than the one in the second video $\langle \text{video2} \rangle$?"
SOI	"Given three tactile videos: a) $\langle \text{video1} \rangle$, b) $\langle \text{video2} \rangle$, c) $\langle \text{video3} \rangle$. Select the $\langle \text{comparison} \rangle \langle \text{attribute} \rangle$ one."
SOI	"You have tactile videos of three objects: a) $\langle \text{video1} \rangle$, b) $\langle \text{video2} \rangle$, c) $\langle \text{video3} \rangle$. Which object is the $\langle \text{comparison} \rangle \langle \text{attribute} \rangle$?"
SOI	"Among these three videos: a) $\langle \text{video1} \rangle$, b) $\langle \text{video2} \rangle$, c) $\langle \text{video3} \rangle$, identify the $\langle \text{comparison} \rangle \langle \text{attribute} \rangle$ object."
OSC	"Given three tactile videos: a) $\langle \text{video1} \rangle$, b) $\langle \text{video2} \rangle$, c) $\langle \text{video3} \rangle$. Match each video (a, b, c) to one of the following objects in alphabetical order: 1) $\langle \text{object1} \rangle$, 2) $\langle \text{object2} \rangle$, 3) $\langle \text{object3} \rangle$."
OSC	"You have tactile videos of three different objects: a) $\langle \text{video1} \rangle$, b) $\langle \text{video2} \rangle$, c) $\langle \text{video3} \rangle$. Assign each video (a, b, c) to one of these objects listed alphabetically: 1) $\langle \text{object1} \rangle$, 2) $\langle \text{object2} \rangle$, 3) $\langle \text{object3} \rangle$."

Table 6: Additional ablation study on VTV encoder settings using the VTV-LLM-7B model.

Settings	Hardness	Protrusion	Elasticity	Friction	Combined
VideoMAE (w/o train)	29.7	9.4	13.0	14.4	0.0
VideoMAE (w/ train)	59.4	74.6	59.4	47.1	22.4
Ours (w/o cls)	63.7	73.1	63.0	39.8	20.2
Ours	73.9	75.0	67.3	56.5	35.6

Table 7: Additional ablation study on three-stage training paradigm settings using the VTV-LLM-7B model.

Settings	Hardness	Protrusion	Elasticity	Friction	Combined
w/o stage 2	52.8	74.6	60.1	50.7	22.4
w/o stage 3	53.6	69.5	51.4	36.9	14.4
Same dataset	68.1	76.0	64.4	53.6	25.3
Ours	73.9	75.0	67.3	56.5	35.6

Table 8: Ablation study on temporal modeling strategies using the VTV-LLM-7B model.

Strategies	Hardness	Protrusion	Elasticity	Friction	Combined	SFD	SOI	OSC	TSA	Average
Selecting 5 frames [15]	62.6	60.7	41.3	31.7	10.8	58.1	47.5	27.1	51.0	43.4
Ours	73.9	75.0	67.3	56.5	35.6	71.3	57.6	43.2	64.0	60.4

size of 16, mask ratio of 0.9, and a cosine annealing learning rate schedule. Training was conducted on 4 NVIDIA RTX 6000 Ada GPUs with an approximate duration of 24 hours for this phase.

VTV-text alignment was executed for 1 epoch employing the AdamW optimizer [61] without weight decay, a learning rate of $2e-4$, batch size of 16, and a cosine annealing learning rate schedule. During text prompt fine-tuning, both the projector and LLM parameters were optimized using newly generated question-answer pairs for 1 epoch. We implemented the AdamW optimizer [61] without weight decay, batch size of 16, and a cosine annealing learning rate schedule. For LoRA [36], we utilized a learning rate of $2e-4$, scaling factor of 256, rank of 128, and dropout rate of 0.05. The computational resources consisted of 1 NVIDIA RTX 6000 Ada GPU for VTV-LLM-3B/7B variants and 2 NVIDIA RTX 6000 Ada GPUs for VTV-LLM-14B, with both stages requiring approximately 3 hours of training time.

A.4.2 More Ablation Studies

Tactile Feature Performance As demonstrated in Tab. 2 and 3, the evaluation metrics primarily focus on high-level reasoning tasks. We further conducted comprehensive ablation studies to assess tactile feature performance, with results presented in Tab. 6 and 7. The experimental findings clearly indicate that each proposed setting contributes positively to enhancing the overall system performance.

Temporal Modeling To further demonstrate the value of temporal modeling in tactile perception, we adopted the data reading method of Octopi [15] (selecting 5 frames) and train using the identical dataset and pipeline. The results shown in the Tab. 8 demonstrate the importance of tactile video understanding, especially for dynamic attributes.

Masking Ratio VideoMAE [27, 28] shows that videos contain significant temporal redundancy, with consecutive frames changing slowly and being highly correlated. It allows excellent reconstruction even at 0.9-0.95 masking ratios, as models can infer masked content from limited visible tokens. Tactile videos have even stronger structural advantages than natural videos. Deformations are localized yet constitute important tactile features, and tactile interactions follow predictable physical laws. Our optical flow-guided masking strategy (Fig. 3) preserves key deformation information consistently, while multi-frame sequences provide rich spatiotemporal context even when single frames are limited.

We conduct dedicated masking ratio ablation experiments specifically for tactile videos. Results can be found in the Tab. 9, where lower performance at 0.7 masking ratio, optimal performance at 0.9. This proves that high masking ratio provides sufficient challenge for effective representation learning without causing information insufficiency.

A.4.3 More LLMs

To validate the generalizability of our framework, we additionally implemented the open-source LLaMA-based LLM, Vicuna [62], as an alternative backbone architecture. Comparative results between this implementation and Qwen 2.5 [4, 5] are presented in Fig. 6. These experimental

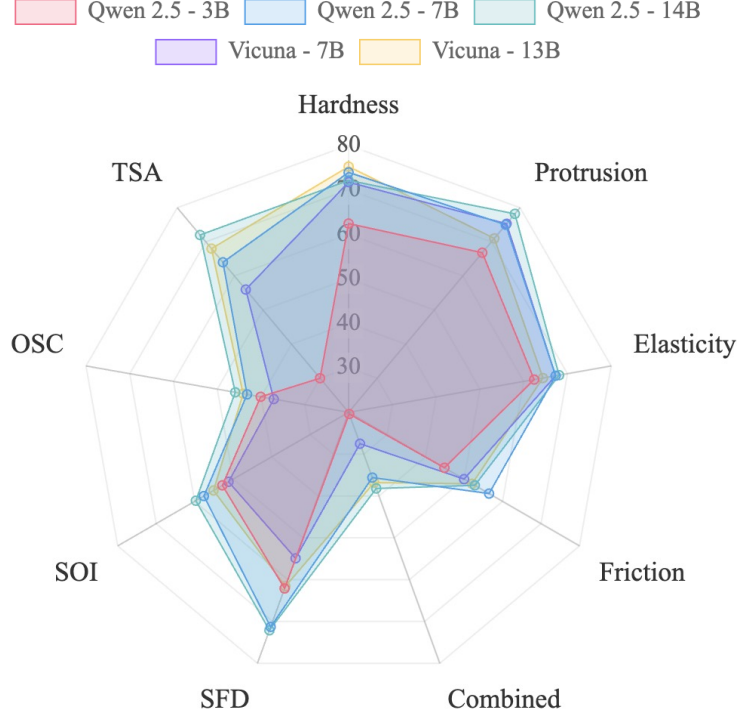


Figure 6: Performance comparison between Qwen-based and LLaMA-based LLMs across multiple metrics.

Table 9: Ablation Studies on masking ratio using the VTV-LLM-7B model.

Masking Ratio	Hardness	Protrusion	Elasticity	Friction	Combined	SFD	SOI	OSC	TSA	Average
0.7	68.4	61.9	58.2	45.3	14.8	59.7	43.7	38.6	49.0	48.8
0.8	73.7	77.3	66.1	54.3	28.8	70.5	52.2	40.3	59.0	58.8
0.9	73.9	75.0	67.3	56.5	35.6	71.3	57.6	43.2	64.0	60.4
0.95	72.4	76.8	62.4	48.9	30.4	72.1	51.4	40.1	60.0	57.1

outcomes demonstrate that LLaMA-based LLMs can be effectively integrated within our proposed framework, confirming the versatility of our method across different foundation models.

A.4.4 Robustness Test

To evaluate cross-sensor generalization, we employed three distinct visuo-tactile sensors with inherently different data characteristics. To rigorously assess the robustness of our approach to novel data distributions, we conducted additional testing using several original Tac3D videos, which feature 20×20 force fields with significantly different signal patterns from those in our training dataset. The experimental results, presented in Fig. 7, demonstrate that our method maintains strong performance across these heterogeneous input modalities, confirming its resilience to sensor-specific variations and its potential for deployment across diverse tactile sensing platforms.

To further verify the effectiveness, we also conducted a zero-shot experiment on the XENSE sensor (a variant of the GelSlim sensor [63]). In the experimental setup, we use the XENSE sensor to perform the same data collection steps for 20 objects as in the original experiment and

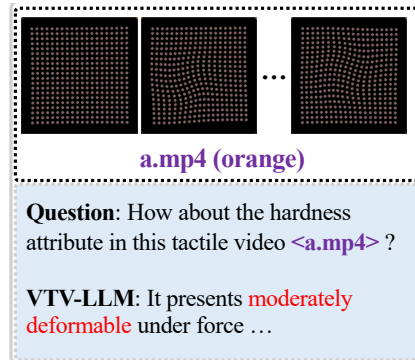


Figure 7: Example about robustness test.

Table 10: Robustness test on the XENSE sensor using the VTV-LLM-7B model.

Strategies	Hardness	Protrusion	Elasticity	Friction	Combined	SFD	SOI	OSC	TSA	Average
Original 3 sensors	73.9	75.0	67.3	56.5	35.6	71.3	57.6	43.2	64.0	60.4
XENSE sensor	77.0	71.0	55.0	52.0	26.0	78.0	55.0	39.0	69.0	58.0

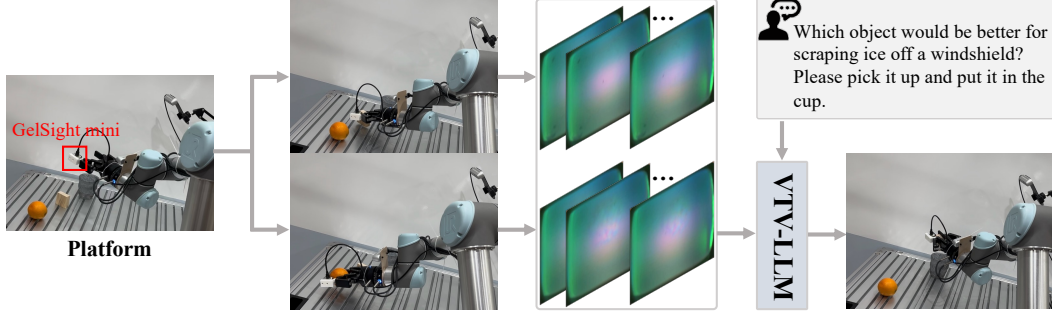


Figure 8: The workflow of real robotic test.

generated 100 templates for each task. Then, the trained VTV-LLM-7B model is used for zero-shot inference. Related results can be found in the Tab. 10, it proves that although our method is trained on only three kinds of sensor data, it generalizes well for the other sensors.

A.4.5 Robotic Test

In this experiment, we developed a multi-modal robotic manipulation system comprising a UR5 collaborative robot, a Robotiq-85 two-finger gripper, and a modified gripper system integrated with a GelSight Mini tactile sensor. As illustrated in Fig. 8, the system operates through the following workflow: Initially, a computer vision system localizes and identifies target objects within the workspace. Subsequently, the robotic arm performs sequential grasping tasks, during which the visuo-tactile sensor establishes physical contact with objects while continuously recording tactile video sequences. Upon completing the tactile data collection from all candidate objects, the robotic arm returns to its predefined home position, awaiting human-robot interaction commands. During the interaction phase, users can express task requirements through natural language queries via our VTV-LLM. The system then combines language comprehension with tactile feature analysis to select the most suitable object from the candidates and executes corresponding pick-and-place operations, precisely positioning the selected object into the designated container.

A.4.6 More Examples

To provide a more comprehensive evaluation, Figs. 9 and 10 present additional representative question-answer pairs across diverse tactile reasoning tasks. These qualitative results clearly demonstrate that our proposed VTV-LLM consistently outperforms established video-based baseline models in multiple tactile reasoning scenarios. The superior performance is particularly evident in complex reasoning tasks requiring fine-grained tactile understanding and multi-modal integration. These illustrative examples further corroborate the quantitative advantages documented in our experimental results and highlight the model’s capacity to effectively bridge visual perception with tactile reasoning in embodied interaction scenarios.



Figure 9: Comparison results on tactile feature assessment and surface feature distinction.

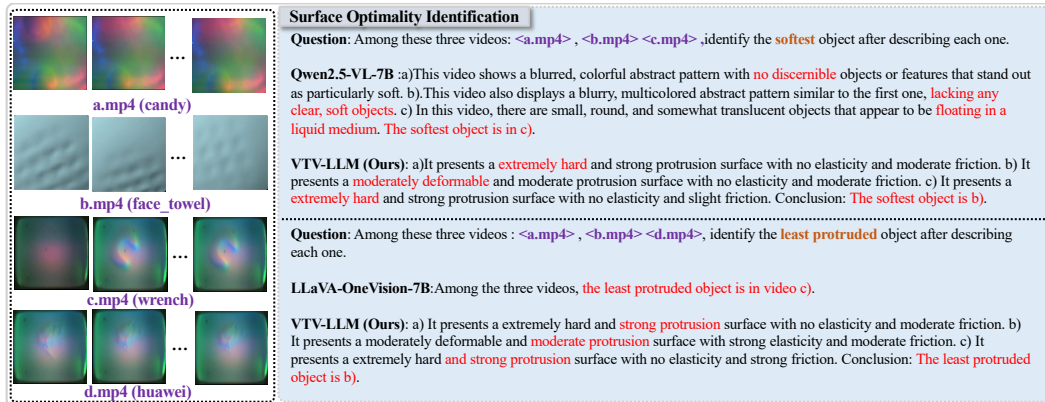


Figure 10: Comparison results on surface optimality identification.