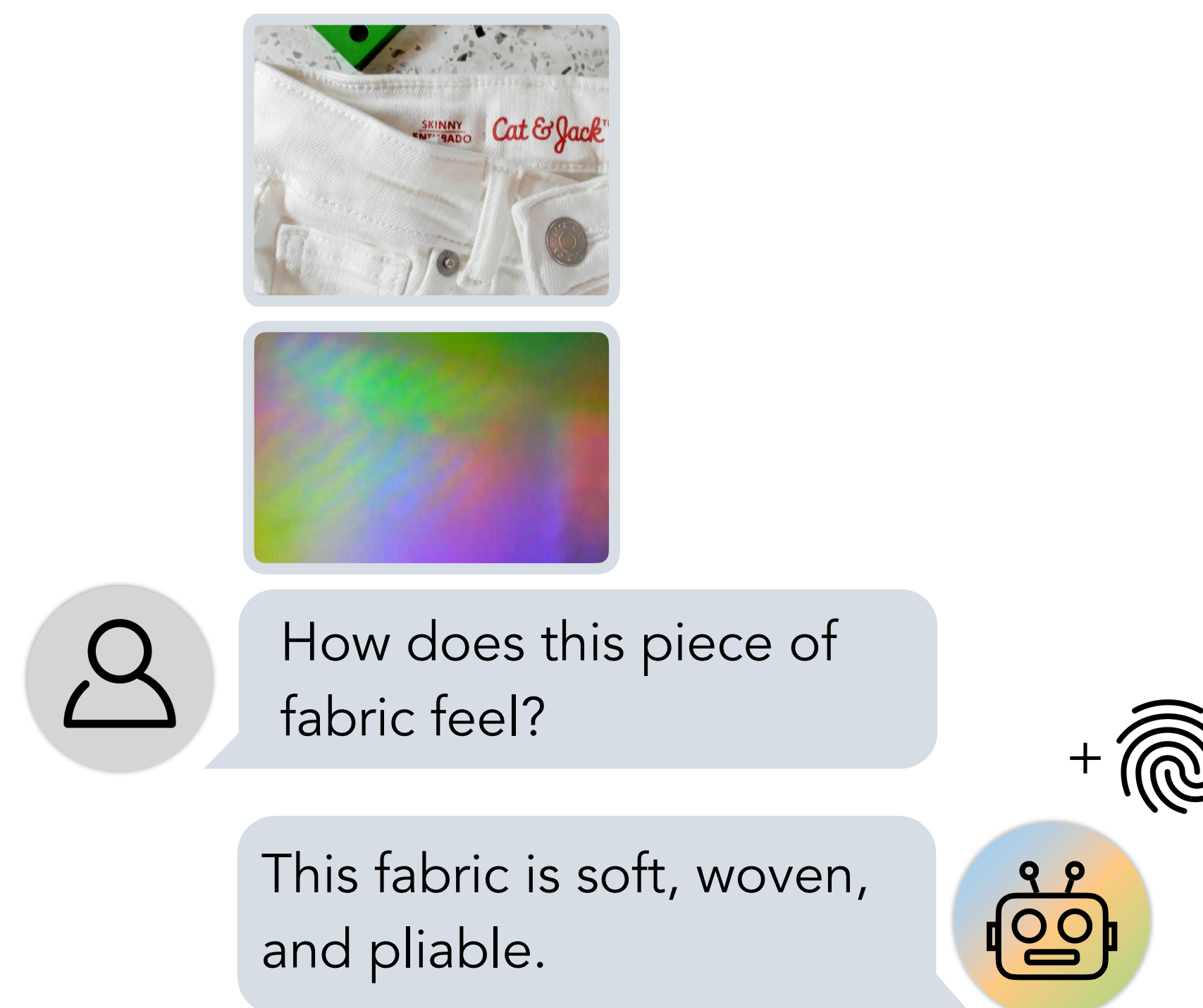
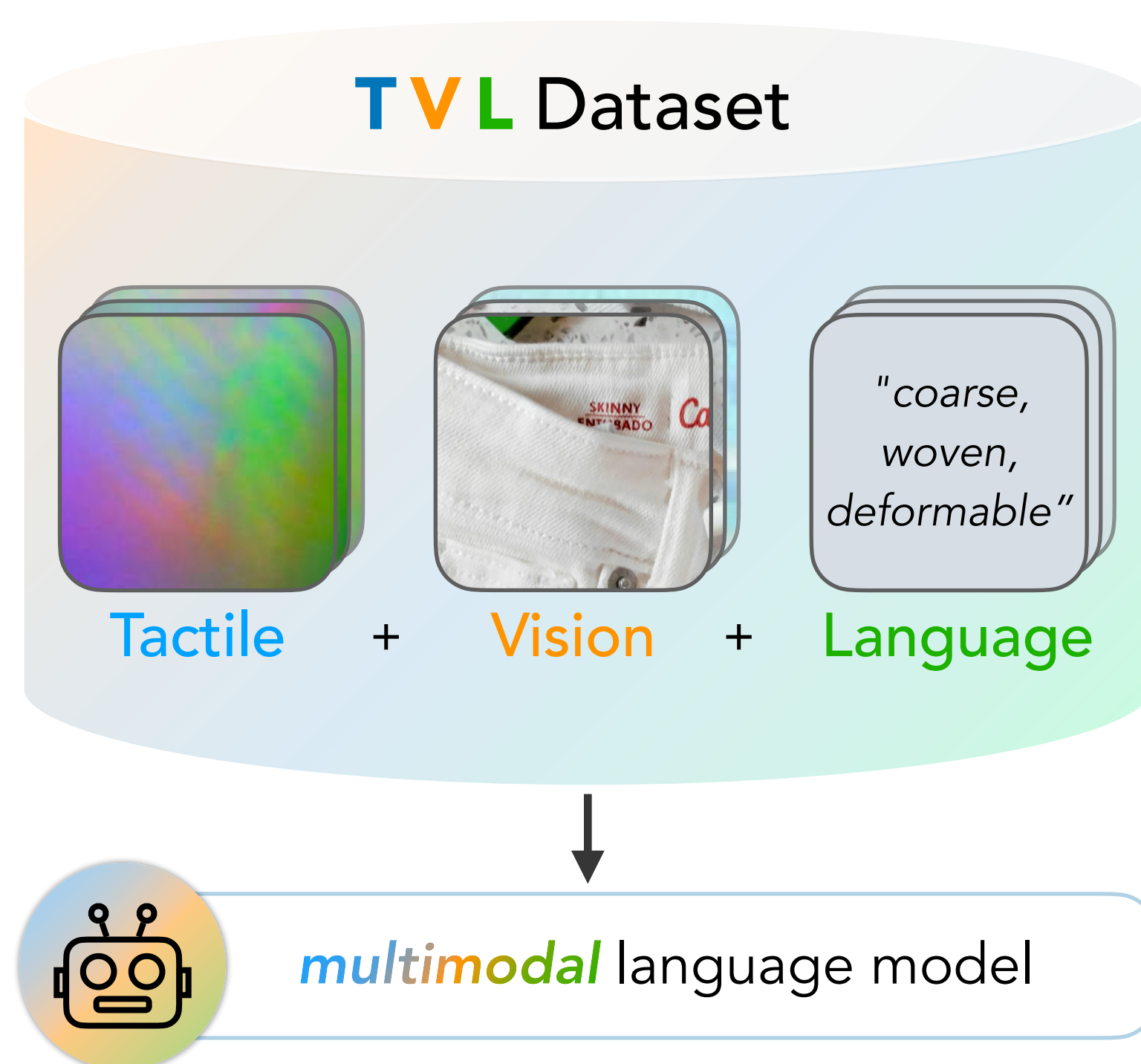


A Touch, Vision, Language Dataset for Multimodal Alignment

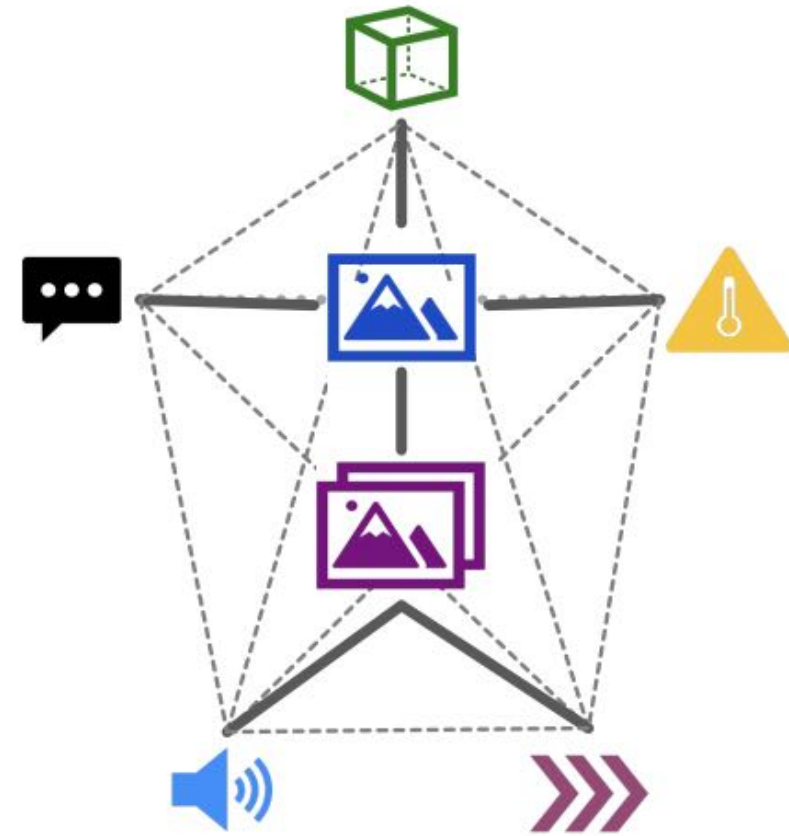
Max (Letian) Fu, Gaurav Datta*, Raven (Huang) Huang*, Will Panitch*, Jaimyn Drake*,
Joseph Ortiz, Mustafa Mukadam, Mike Lambeta, Roberto Calandra, Ken Goldberg



Multimodal Alignment



CLIP [1]



ImageBind [2]



GPT-4V [3]



LLaVA [4]



Flamingo [5]



[1] Radford, Alec et al. "Learning transferable visual models from natural language supervision." ICML 2021.

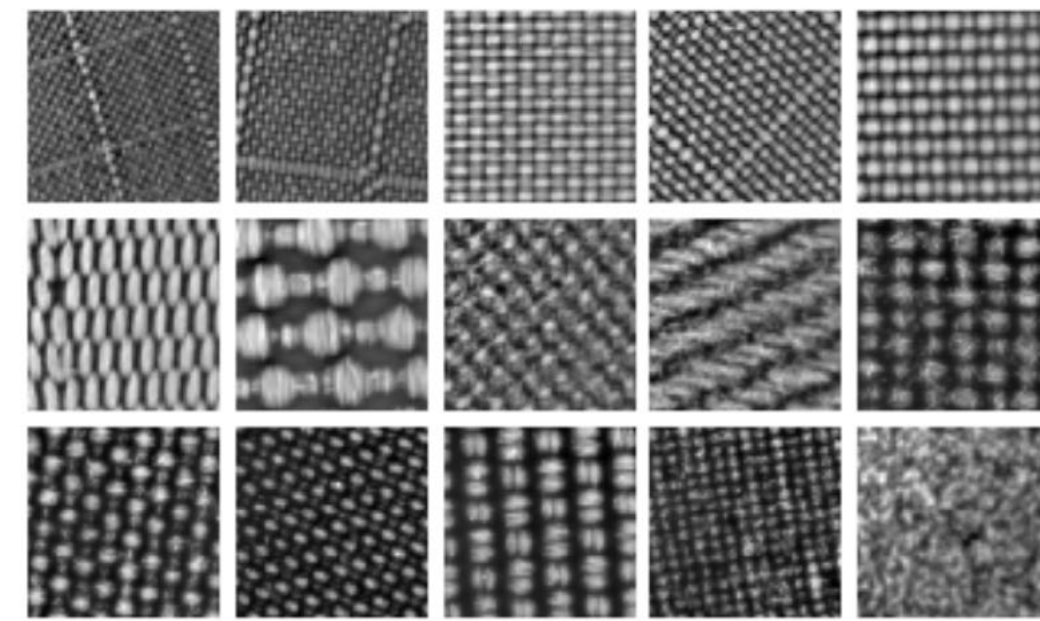
[2] Girdhar, Rohit et al. "Imagebind: One embedding space to bind them all." CVPR 2023.

[3] OpenAI. GPT-4V. 2023.

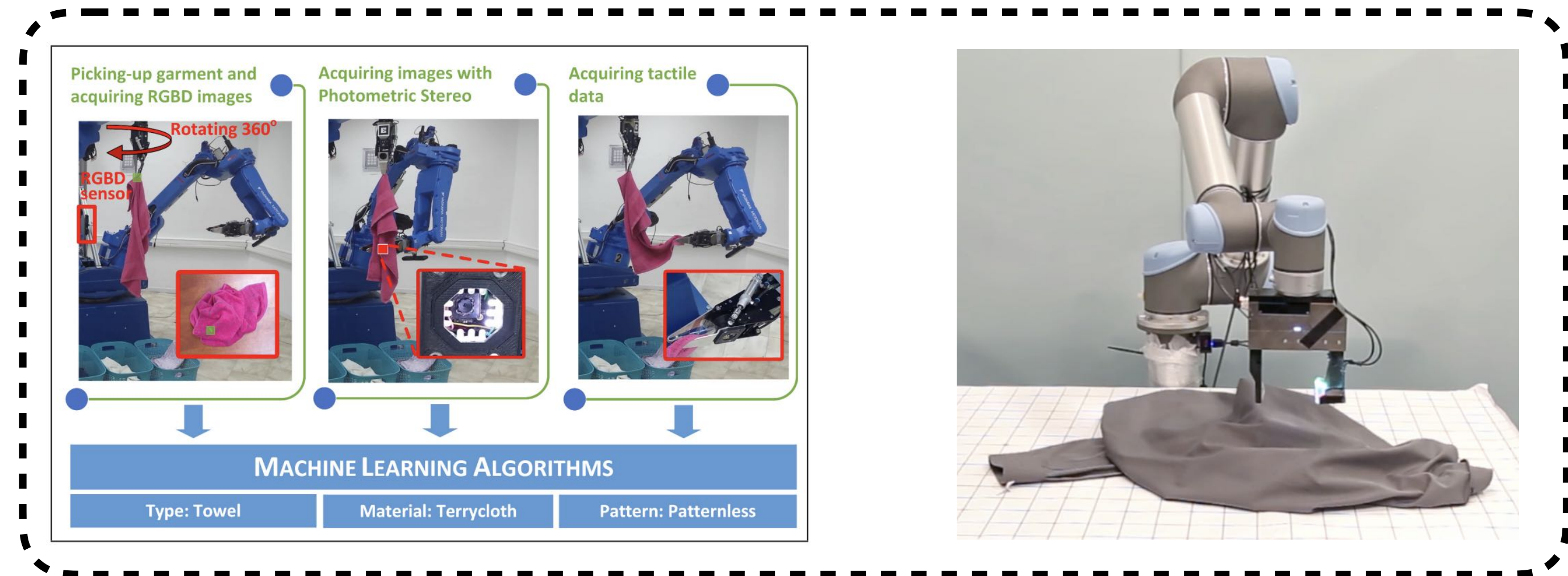
[4] Liu, Haotian et al. "Visual Instruction Tuning." NeurIPS 2023.

[5] Alayrac, Jean-Baptiste et al. "Flamingo: a Visual Language Model for Few-Shot Learning." NeurIPS 2022.

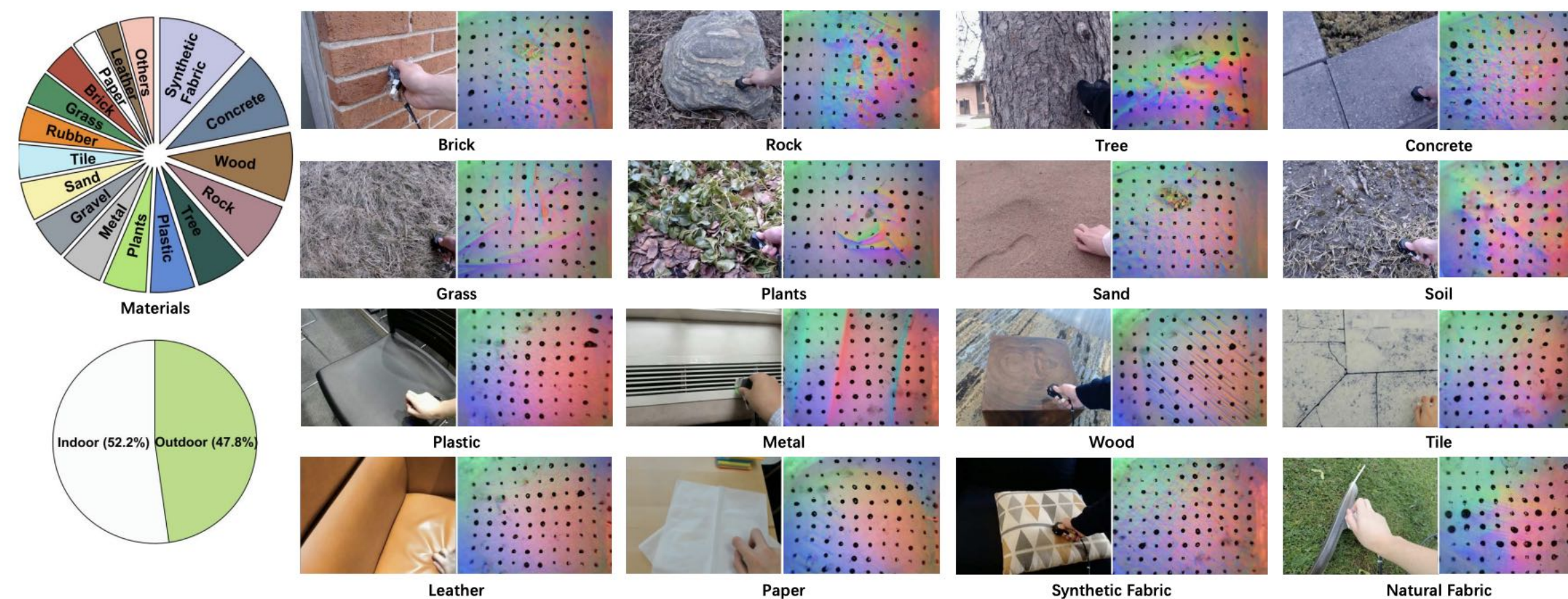
💡 Touch was not yet associated with open vocabulary descriptions



Texture Classification [1]



Cloth Classification [2,3]



“In-the-wild” Texture Classification [4]

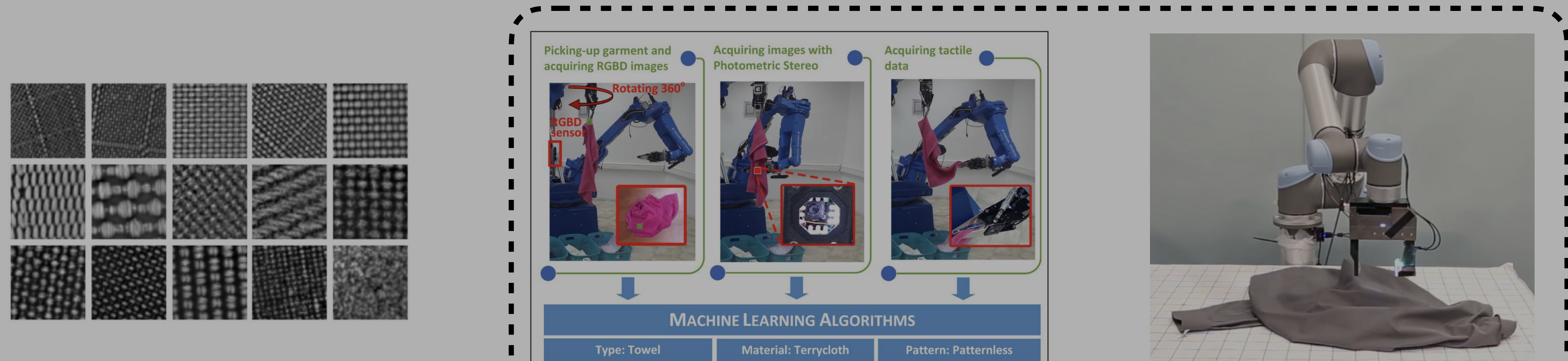
[1] Li, Rui and Edward H. Adelson. "Sensing and recognizing surface textures using a gelsight sensor." CVPR 2013.

[2] Kampouris, Christos et al. "Multisensorial and explorative recognition of garments and their material properties in unconstrained environment." ICRA 2016.

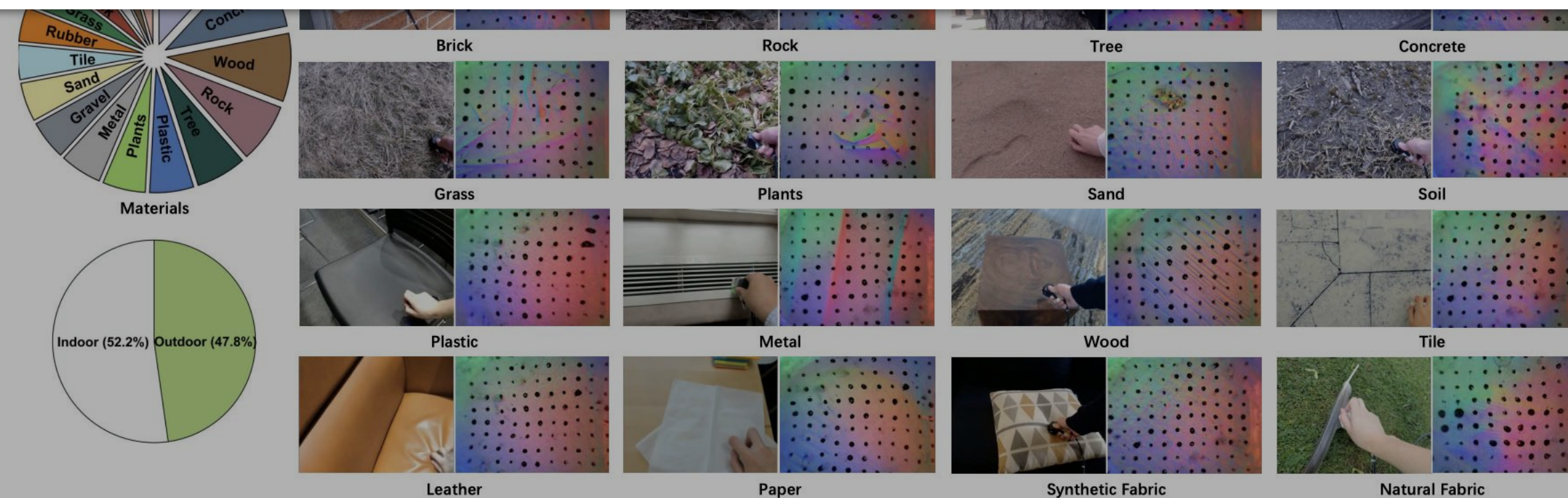
[3] Yuan, Wenzhen et al. "Active clothing material perception using tactile sensing and deep learning." ICRA 2018.

[4] Yang, Fengyu et al. "Touch and Go: Learning from Human-Collected Vision and Touch." NeurIPS 2022.

💡 Touch was not yet associated with open vocabulary descriptions



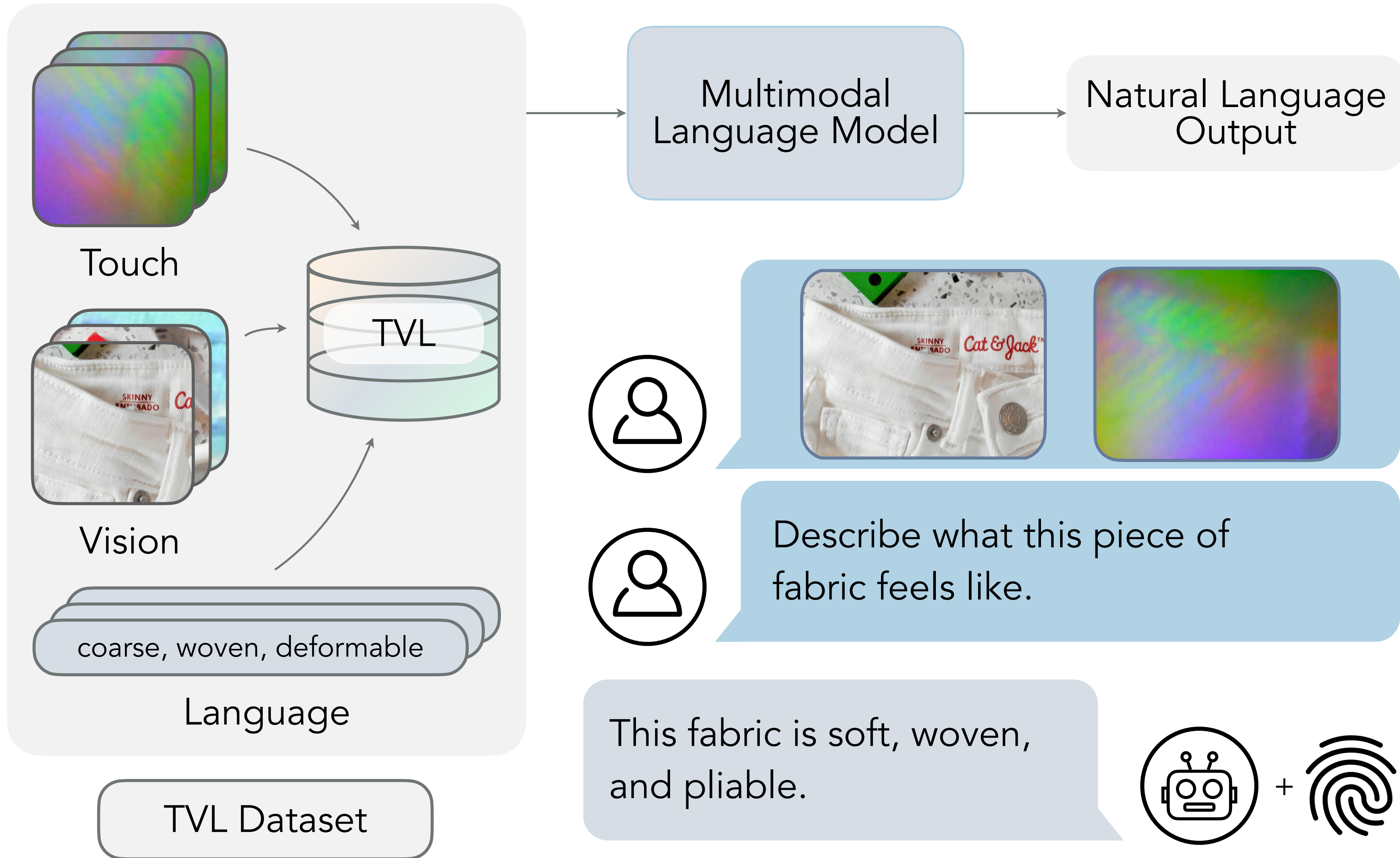
Touch, Vision, Language?



“In-the-wild” Texture Classification [4]

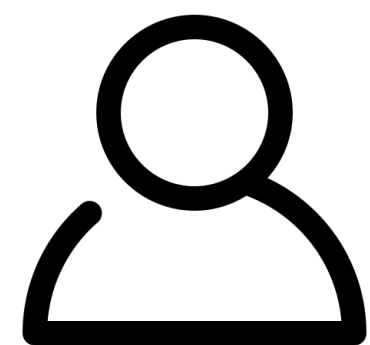
- [1] Li, Rui and Edward H. Adelson. "Sensing and recognizing surface textures using a gelsight sensor." CVPR 2013.
- [2] Kampouris, Christos et al. "Multisensorial and explorative recognition of garments and their material properties in unconstrained environment." ICRA 2016.
- [3] Yuan, Wenzhen et al. "Active clothing material perception using tactile sensing and deep learning." ICRA 2018.
- [4] Yang, Fengyu et al. "Touch and Go: Learning from Human-Collected Vision and Touch." NeurIPS 2022.

Framework

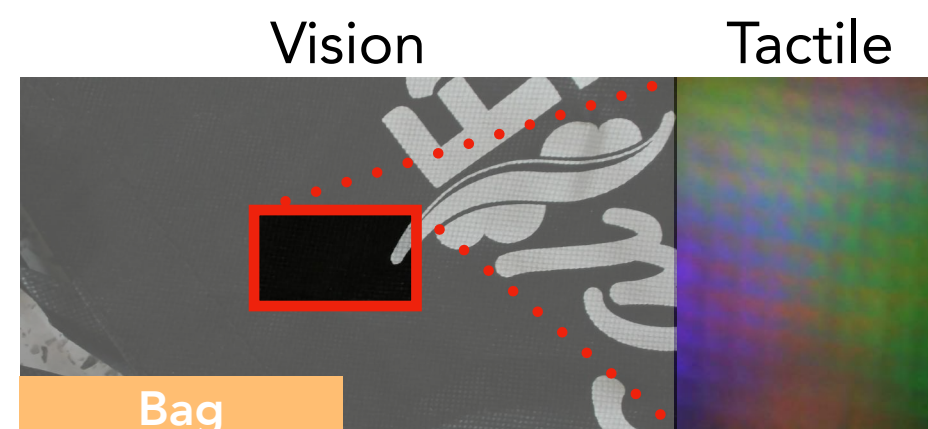


TVL Dataset

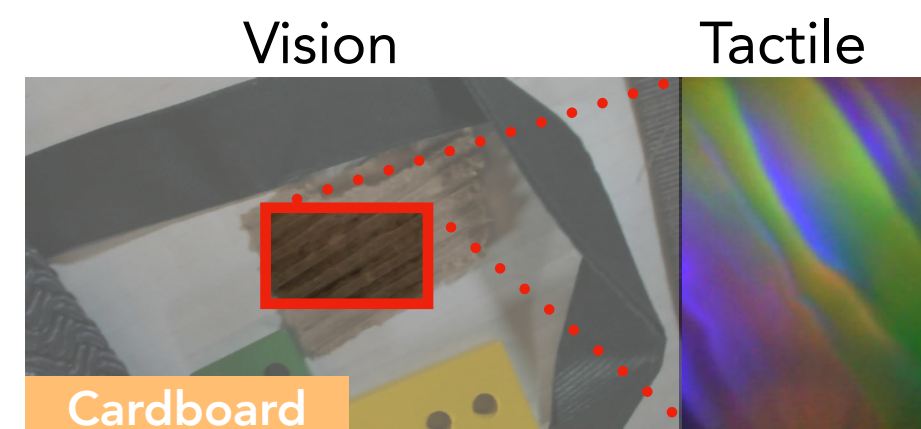
SSVTP [1]



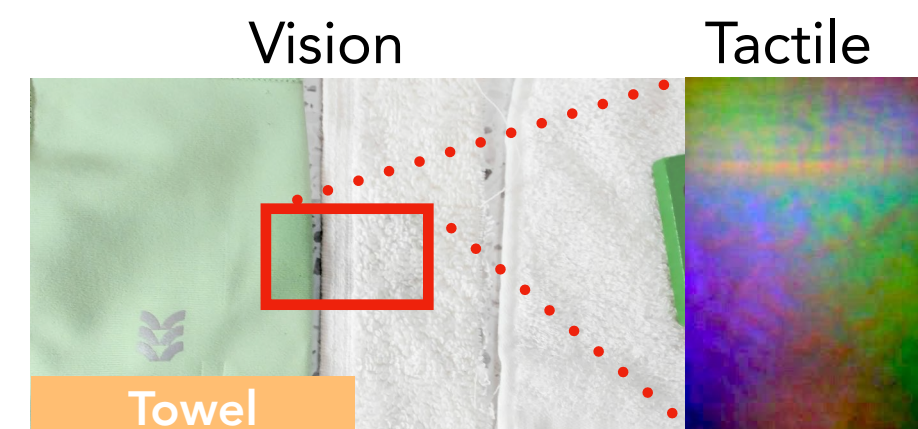
4.6K Human
Annotations



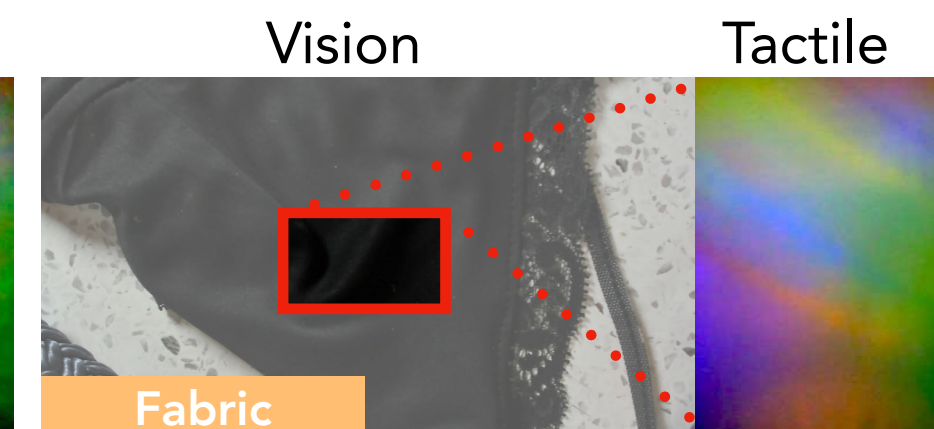
fabric, coarse



lined, cardboard, creased



fabric, bumpy

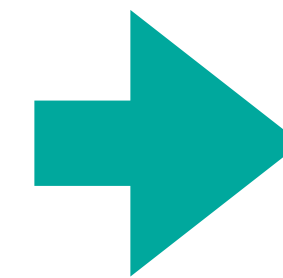
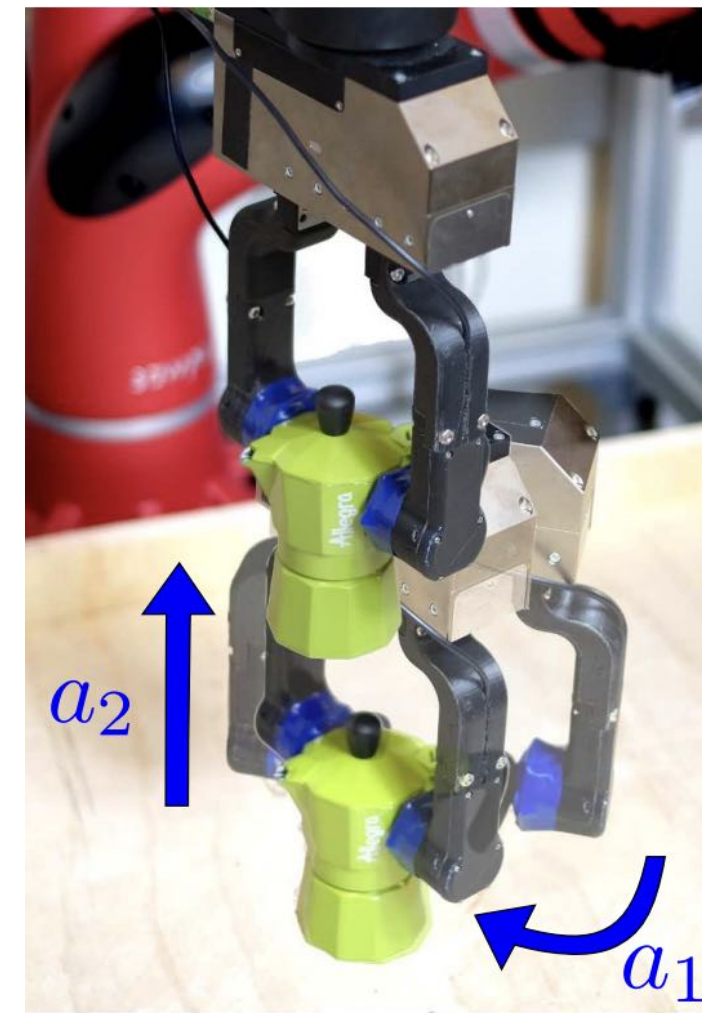
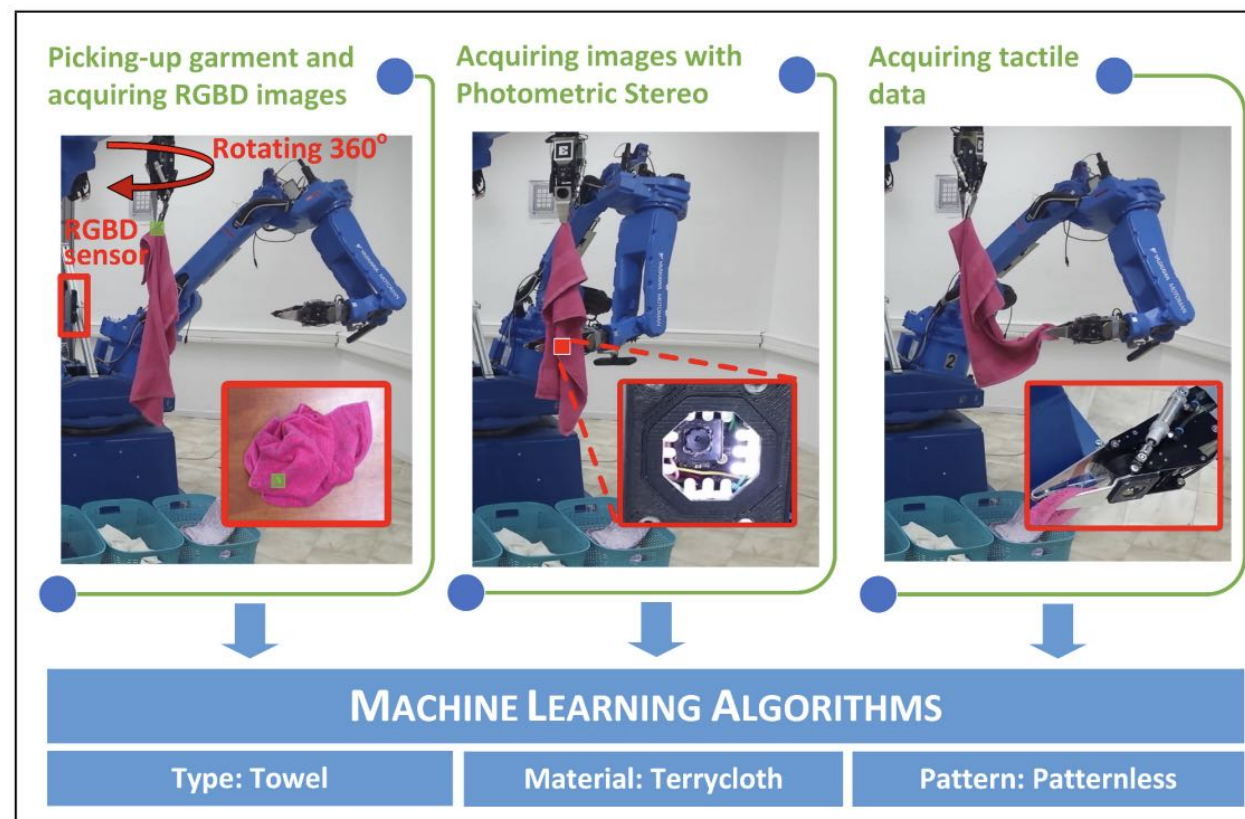


deformable, grainy

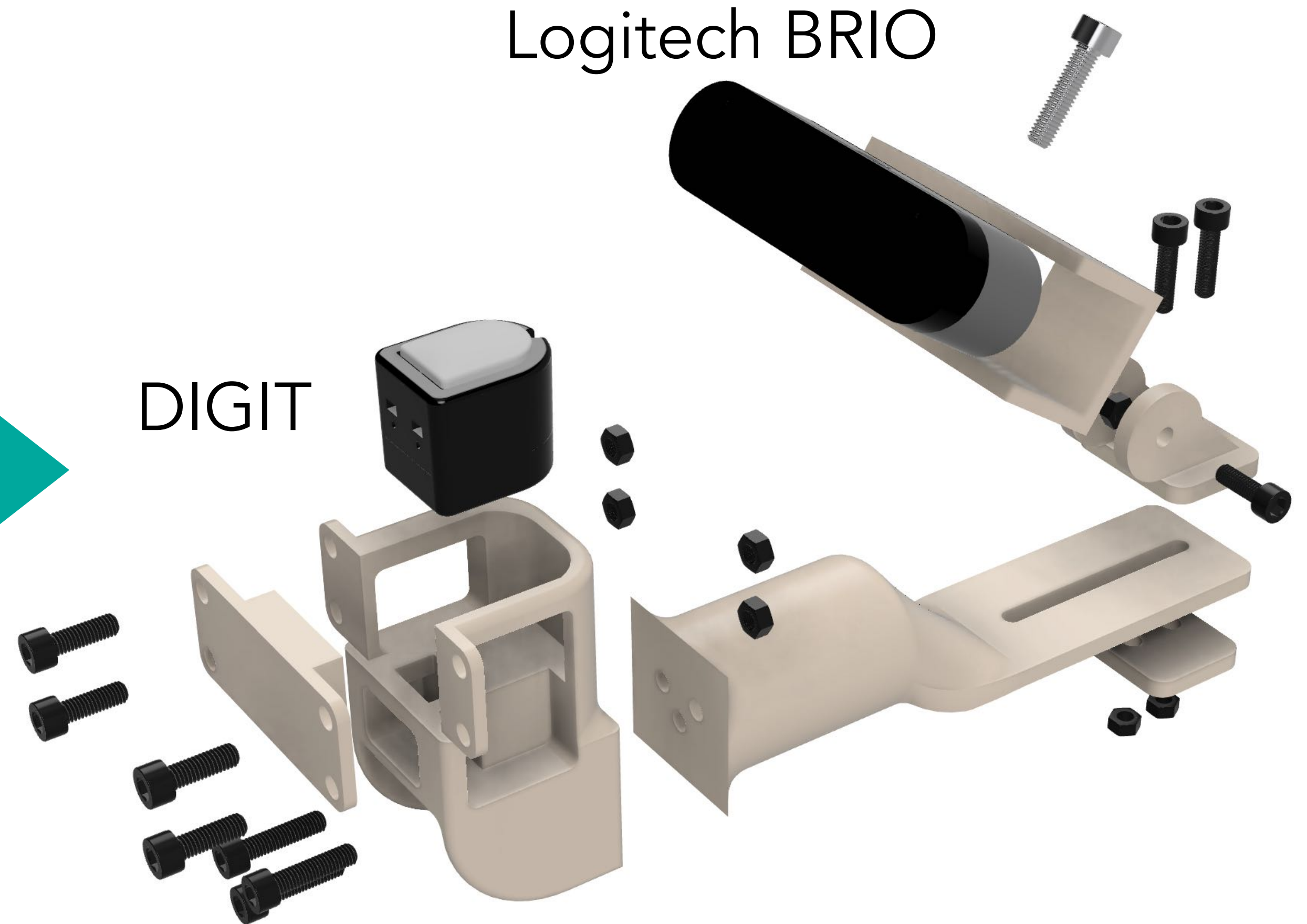
[1] Kerr, Justin et al. "Self-supervised visuo-tactile pretraining to locate and follow garment features." RSS 2023.

[2] Barnett, A.J. "400 Words to Describe Texture." 2023.

Data collection



DIGIT



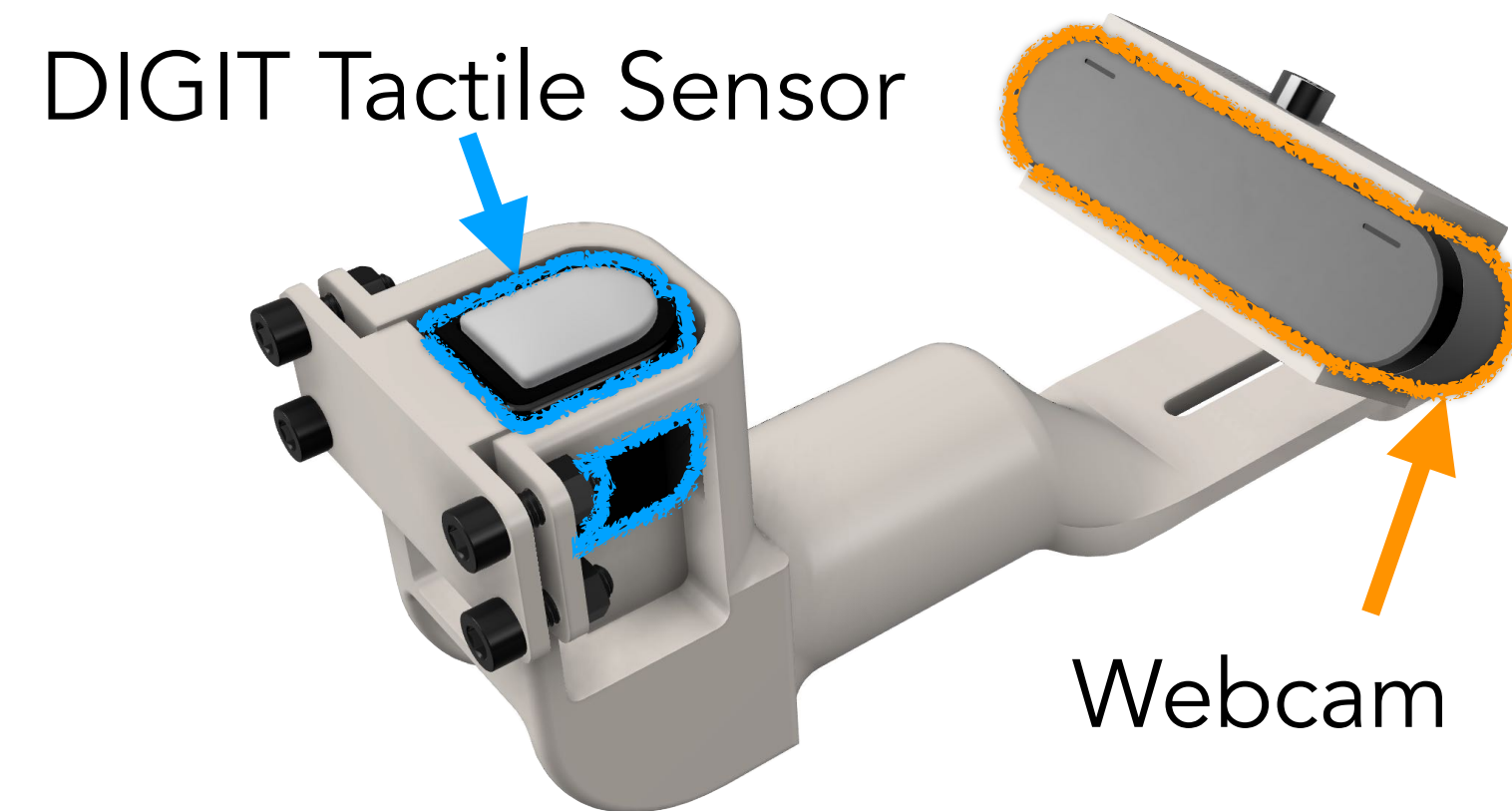
Logitech BRIO

Controlled environments and objects

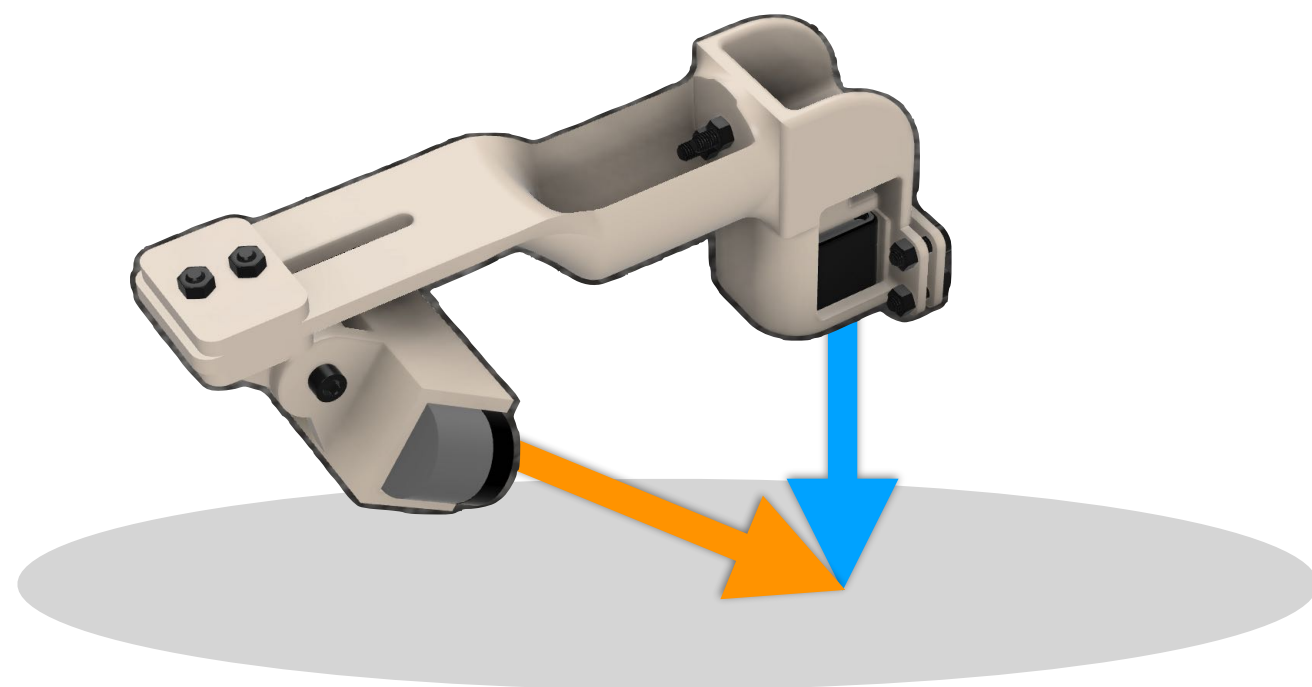
"In-the-wild" Device

Data collection

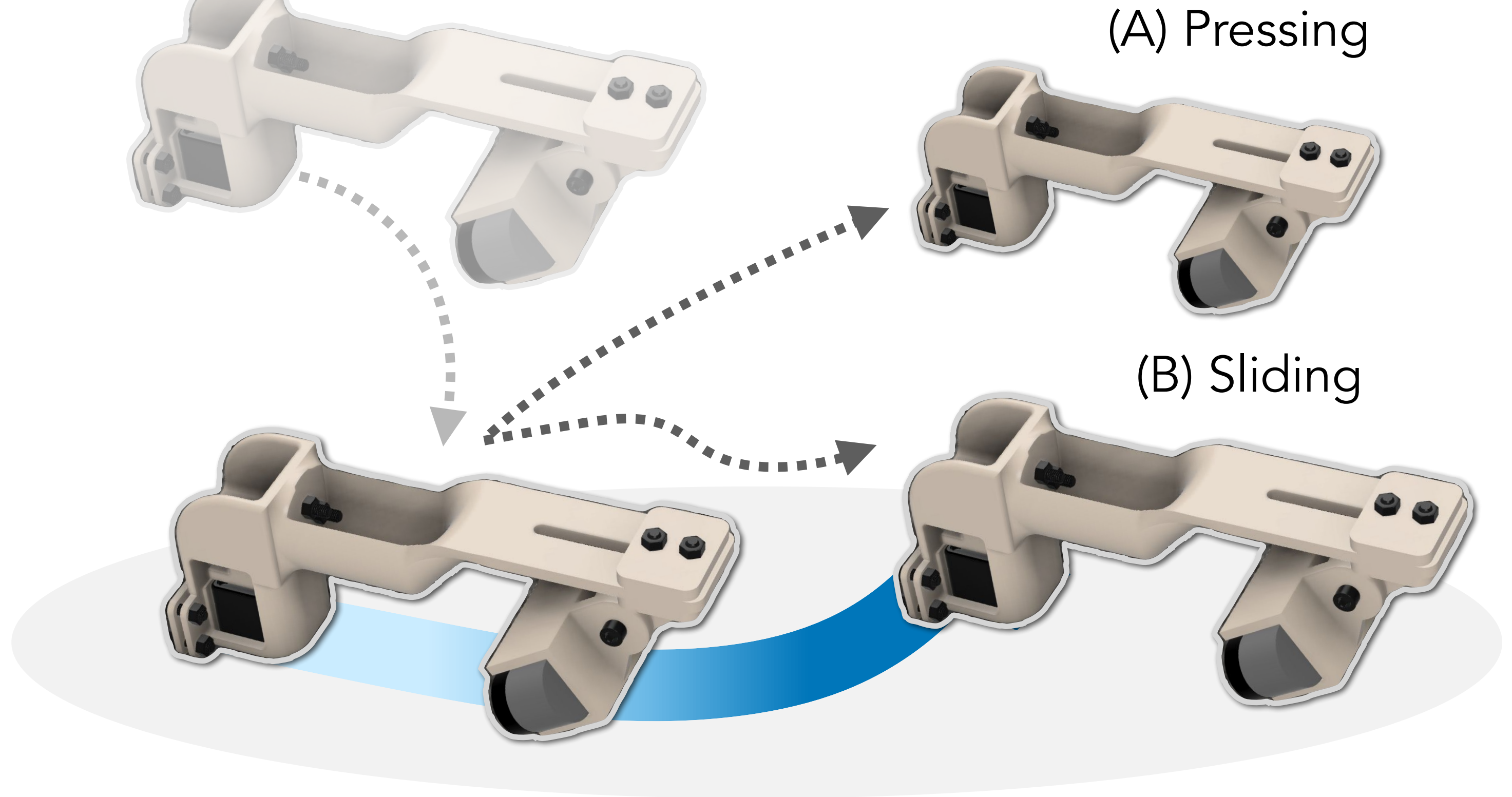
(1) Multimodal



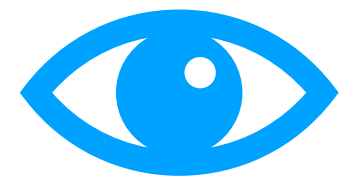
(2) Synchronous Collection



(3) Multiple Motions



GPT-4V Pseudo-labeling



Visual Obs



Prompt

Surface Type: [Specify the surface type, e.g., "metal," "fabric"]

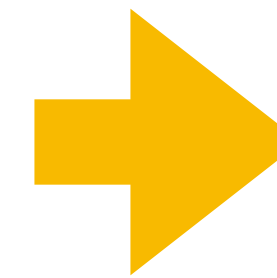
Images: The first image is from a camera observing the tactile sensor (shiny, near the top of the image) and the surface. The second image is a cropped version of the first image that focuses on the contact patch.

Example: For a smooth and cold surface, the description might be "slick, chilly, hard, unyielding, glossy."

Task: Based on these images, describe the possible tactile feelings of the contact patch using sensory adjectives. Limit your response up to five adjectives, separated by commas.

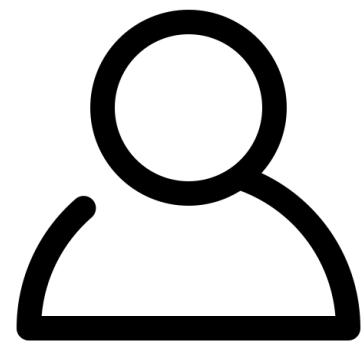


GPT-4V [1]

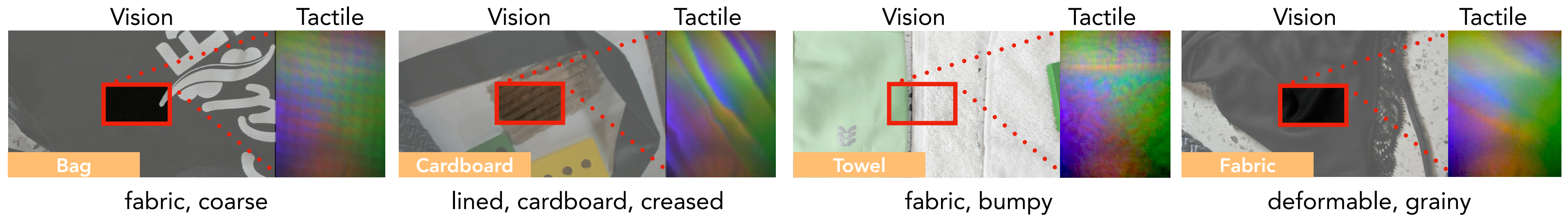


"textured, firm, worn, cool"

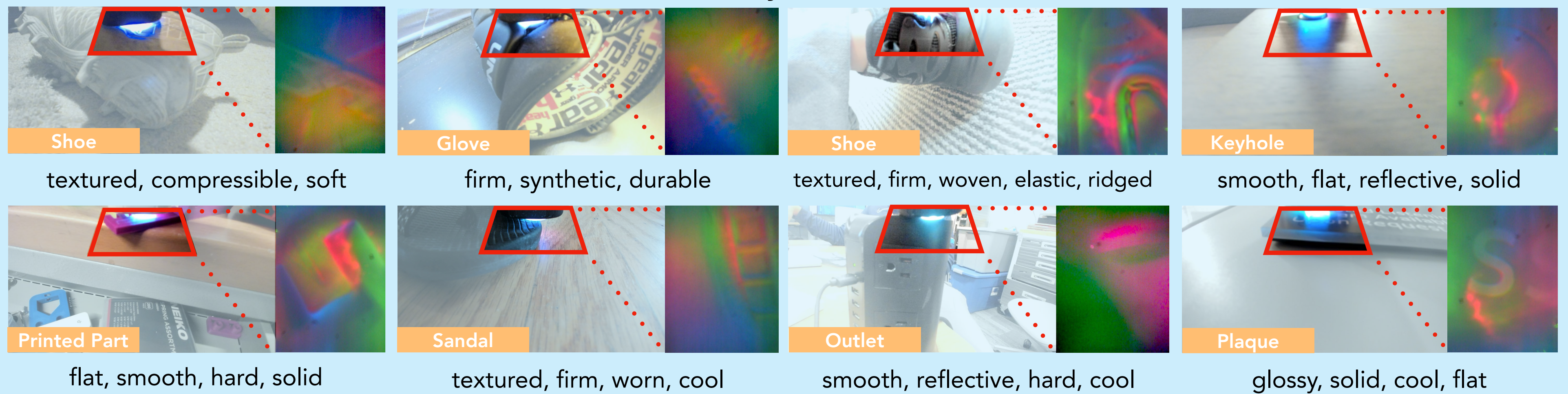
Human + VLM Pseudo-labels



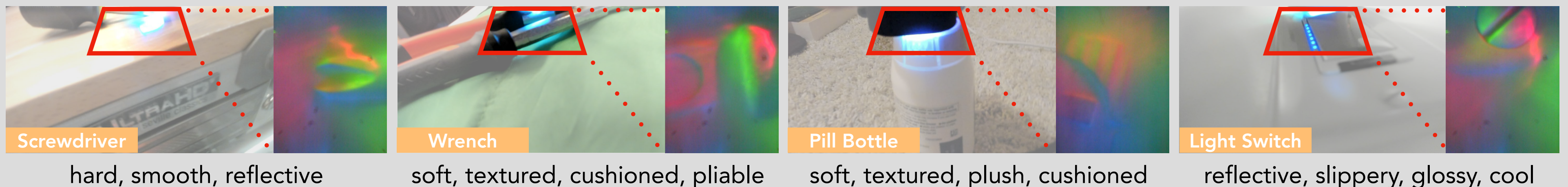
4.6K Human Annotations



Correctly Labeled

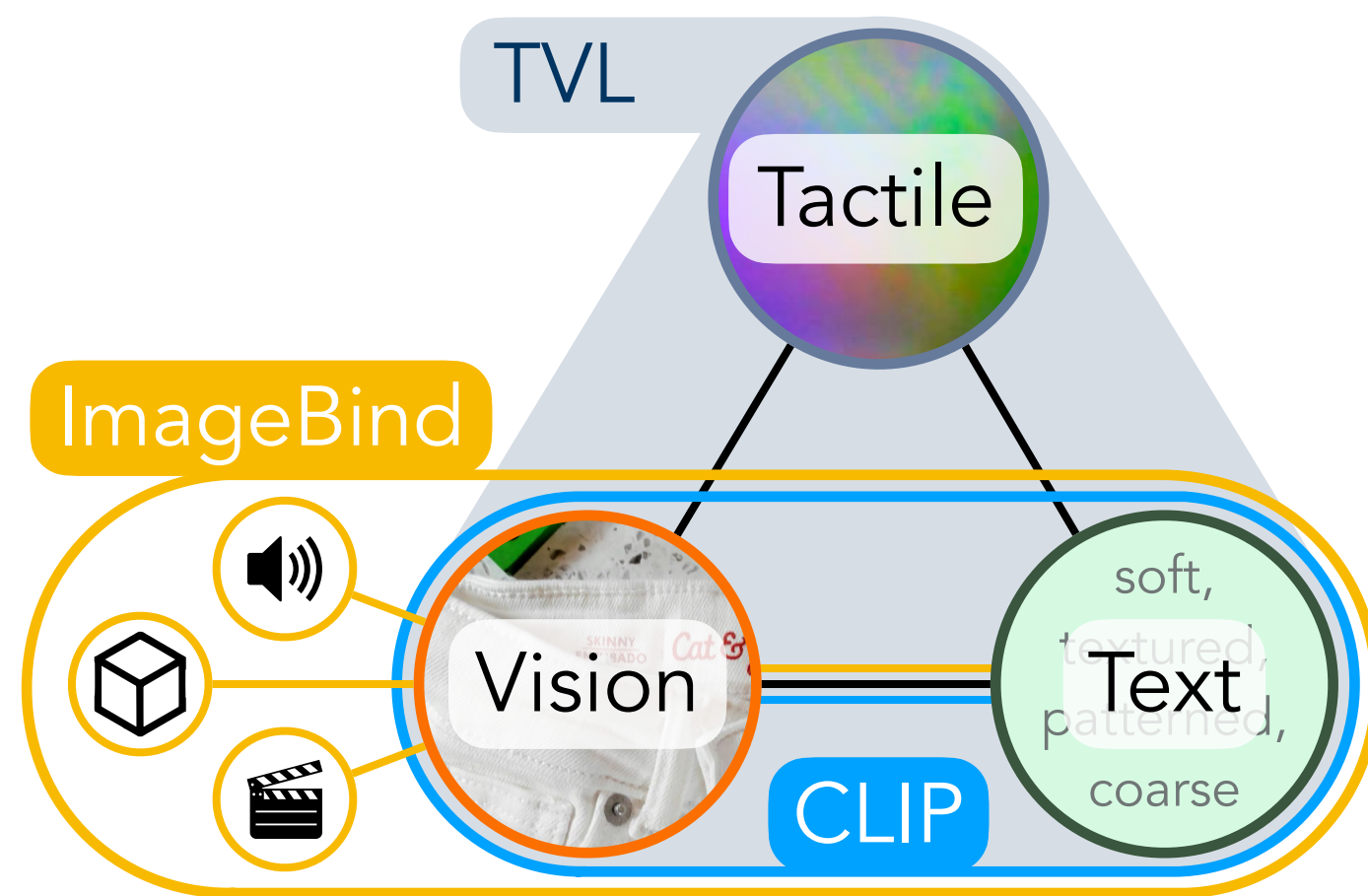


Mislabeled

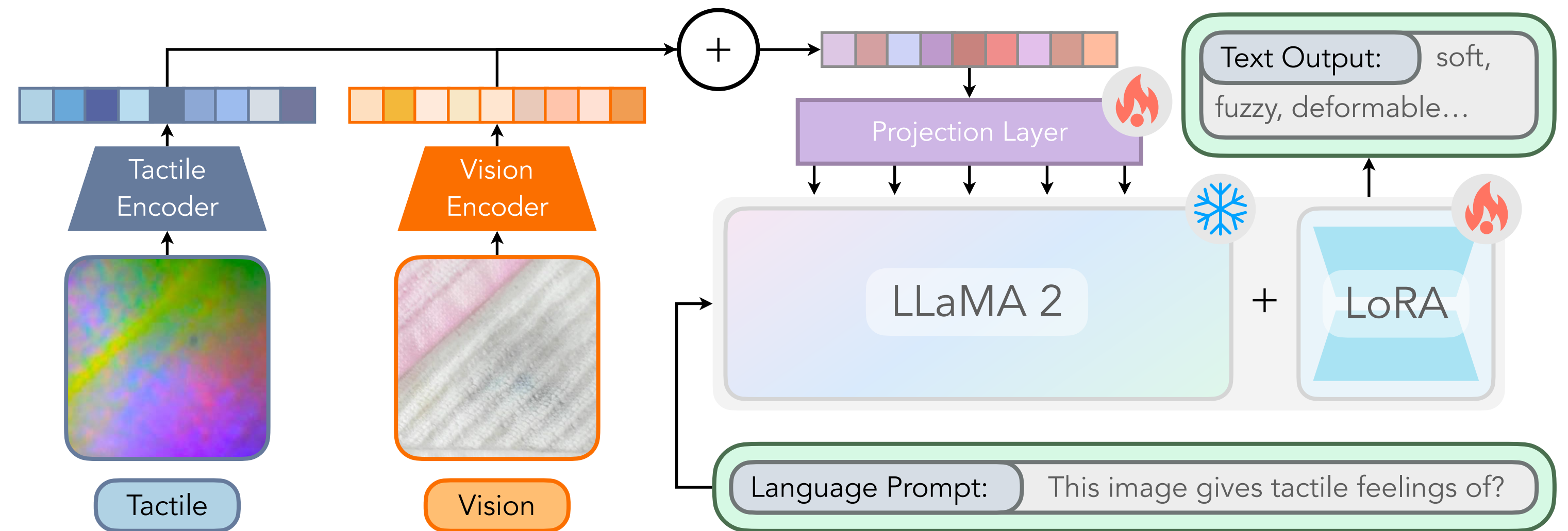


39K Pseudo-Labels
with GPT-4V

TVL Models

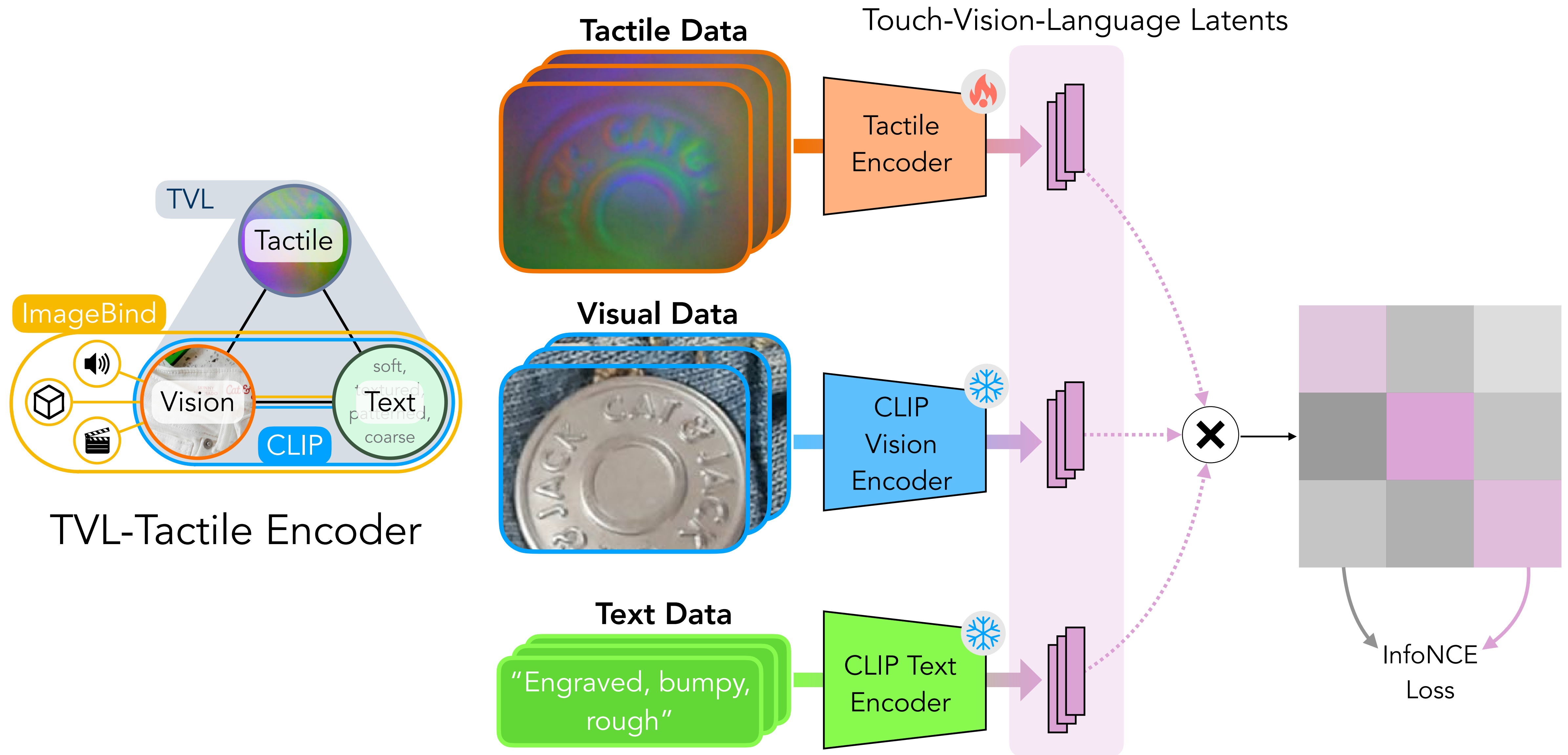


TVL-Tactile Encoder

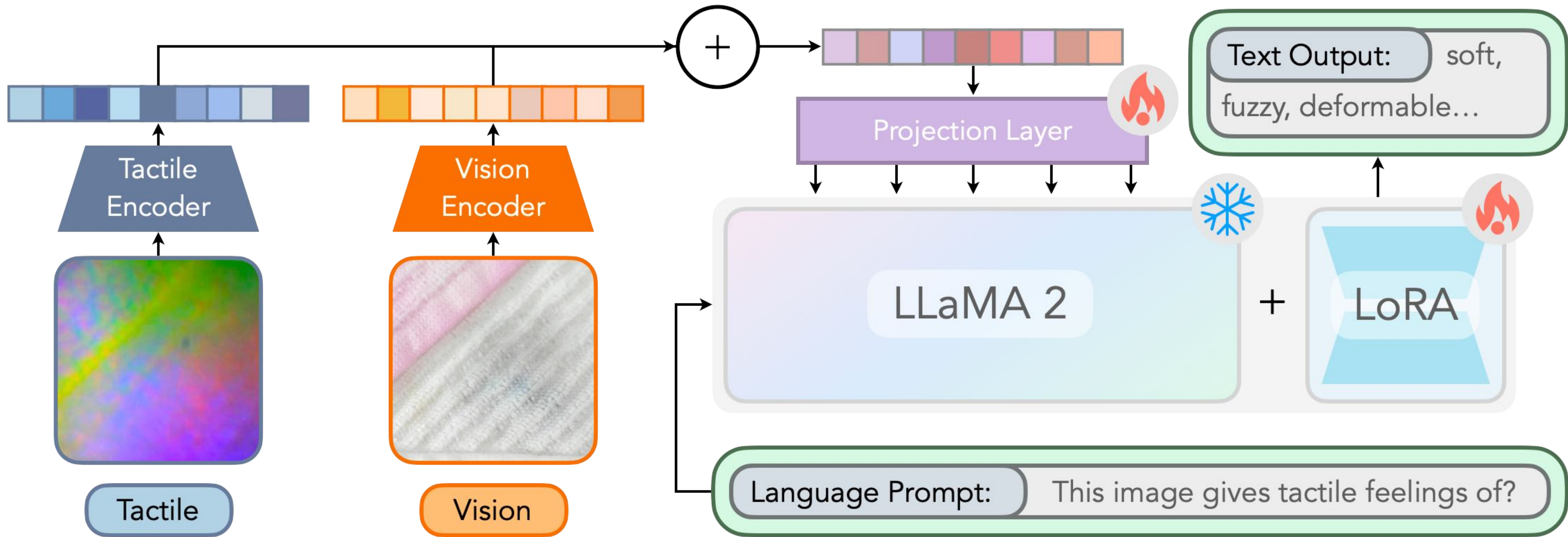


TVL-LLaMA

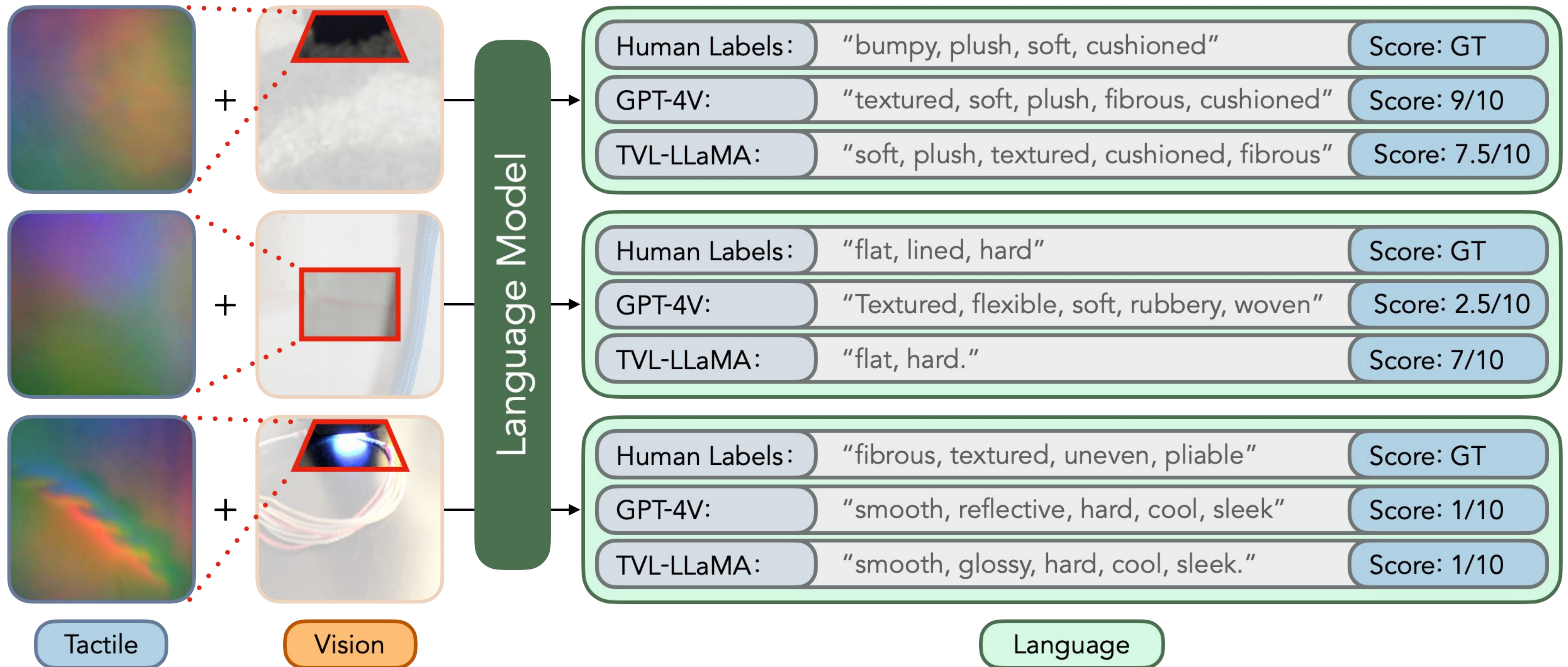
TVL-Tactile Encoder



TVL-LLaMA

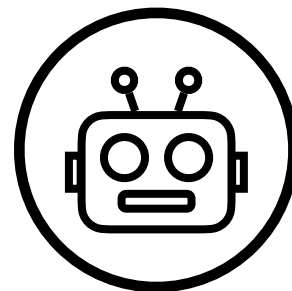


TVL-Benchmark



TVL-Benchmark

"Soft, smooth"



Given that the human labelled ground truth is "soft and fabric", rate the response from 1-10 and provide reasoning.



7

The assistant's response generally aligns with the correct response ...

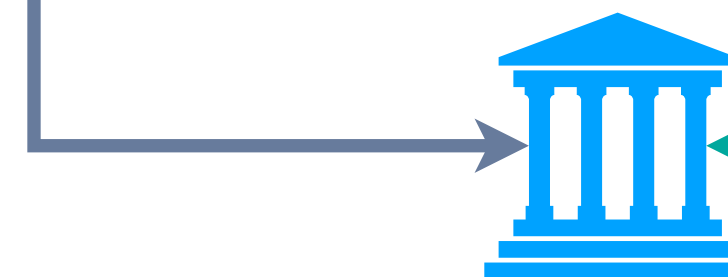


GPT4V Scores

7

(T)VLM Scores

8

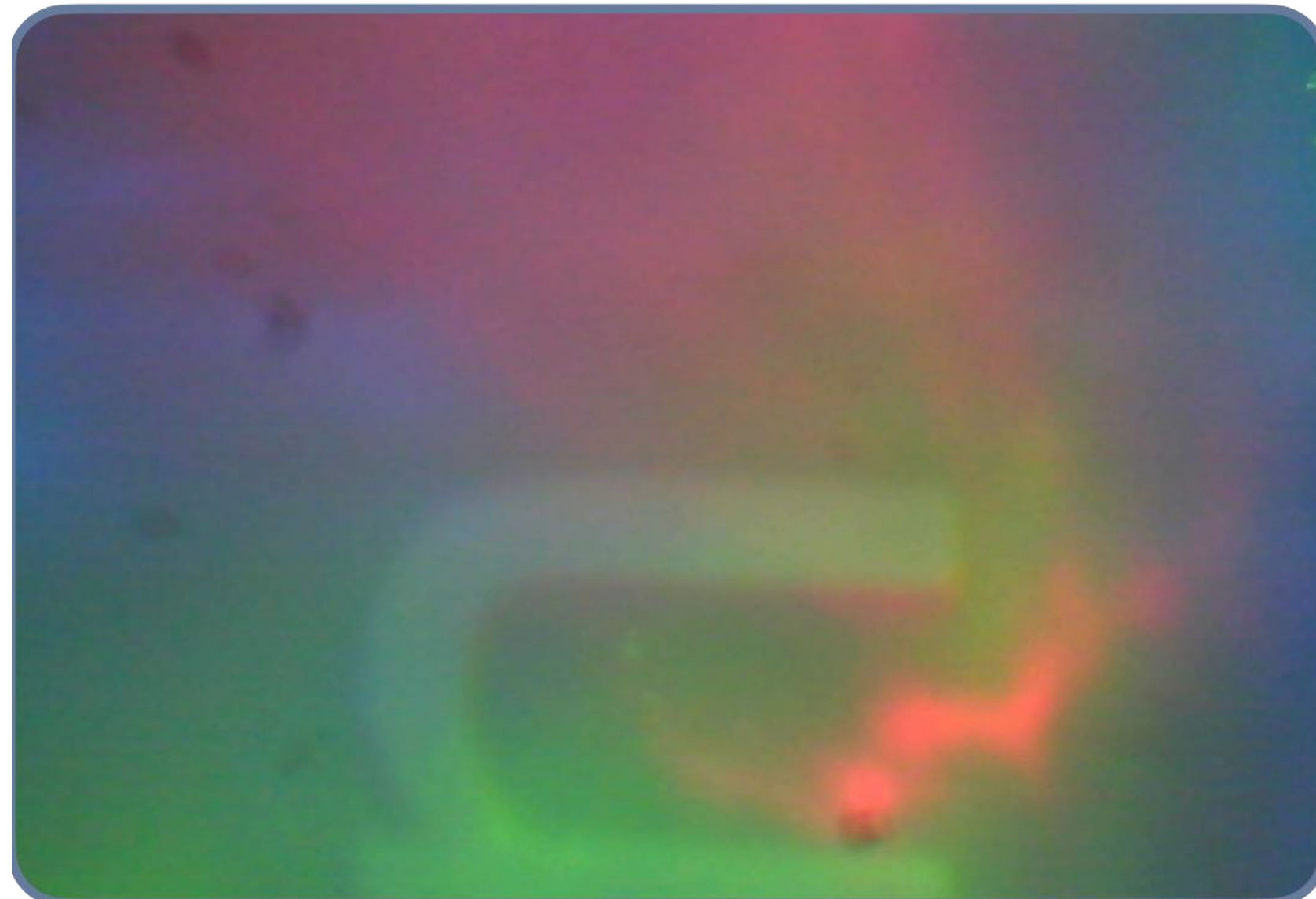
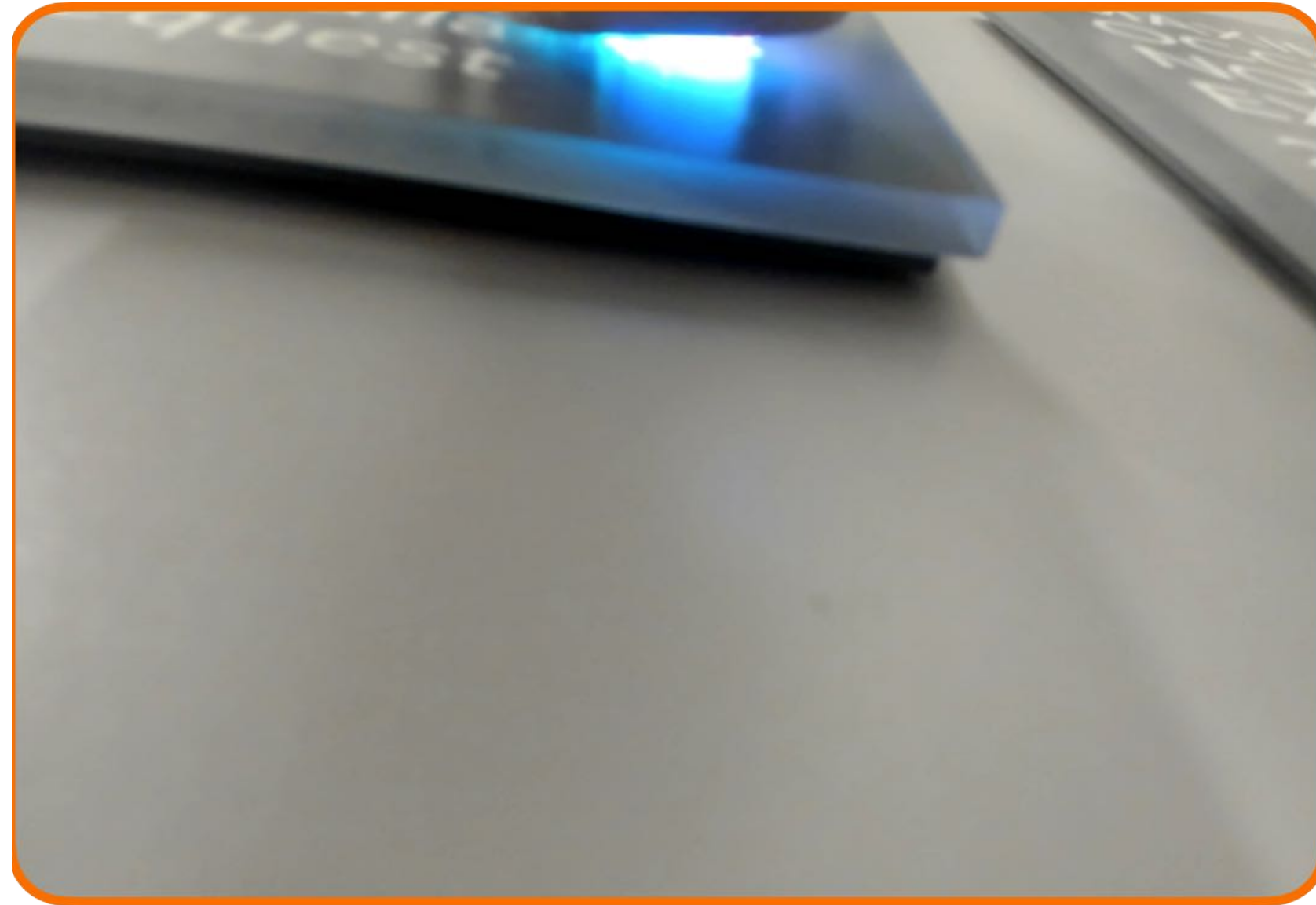


Pair Sampled t-Test

TVL-Benchmark

	Encoder Pre-training Modalities			Score (1-10)			p -value (d.f. = 401)
	Vision	Tactile	Language	SSVTP	HCT	TVL	
LLaVA-1.5 7B	✓	-	✓	3.64	3.55	3.56	1.21×10^{-9}
LLaVA-1.5 13B	✓	-	✓	3.55	3.63	3.62	1.49×10^{-9}
ViP-LLaVA 7B	✓	-	✓	2.72	3.44	3.36	8.77×10^{-16}
ViP-LLaVA 13B	✓	-	✓	4.10	3.76	3.80	1.72×10^{-6}
LLaMA-Adapter	✓	-	✓	2.56	3.08	3.02	2.68×10^{-17}
BLIP-2 Opt-6.7b	✓	-	✓	2.02	2.72	2.64	1.92×10^{-31}
InstructBLIP 7B	✓	-	✓	1.40	1.30	1.31	1.07×10^{-84}
InstructBLIP 13B	✓	-	✓	1.44	1.21	1.24	4.64×10^{-88}
GPT-4V	✓	-	✓	5.02	4.42	4.49	-
SSVTP-LLaMA	✓	✓	-	2.58	3.67	3.54	1.79×10^{-9}
TVL-LLaMA (ViT-Tiny)	✓	✓	✓	6.09	4.79	4.94	4.24×10^{-5}
TVL-LLaMA (ViT-Small)	✓	✓	✓	5.81	4.77	4.89	6.02×10^{-4}
TVL-LLaMA (ViT-Base)	✓	✓	✓	6.16	4.89	5.03	3.46×10^{-6}

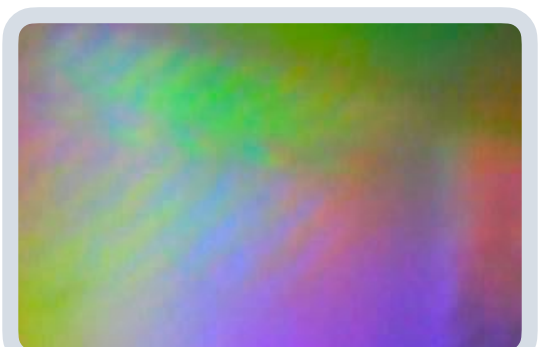
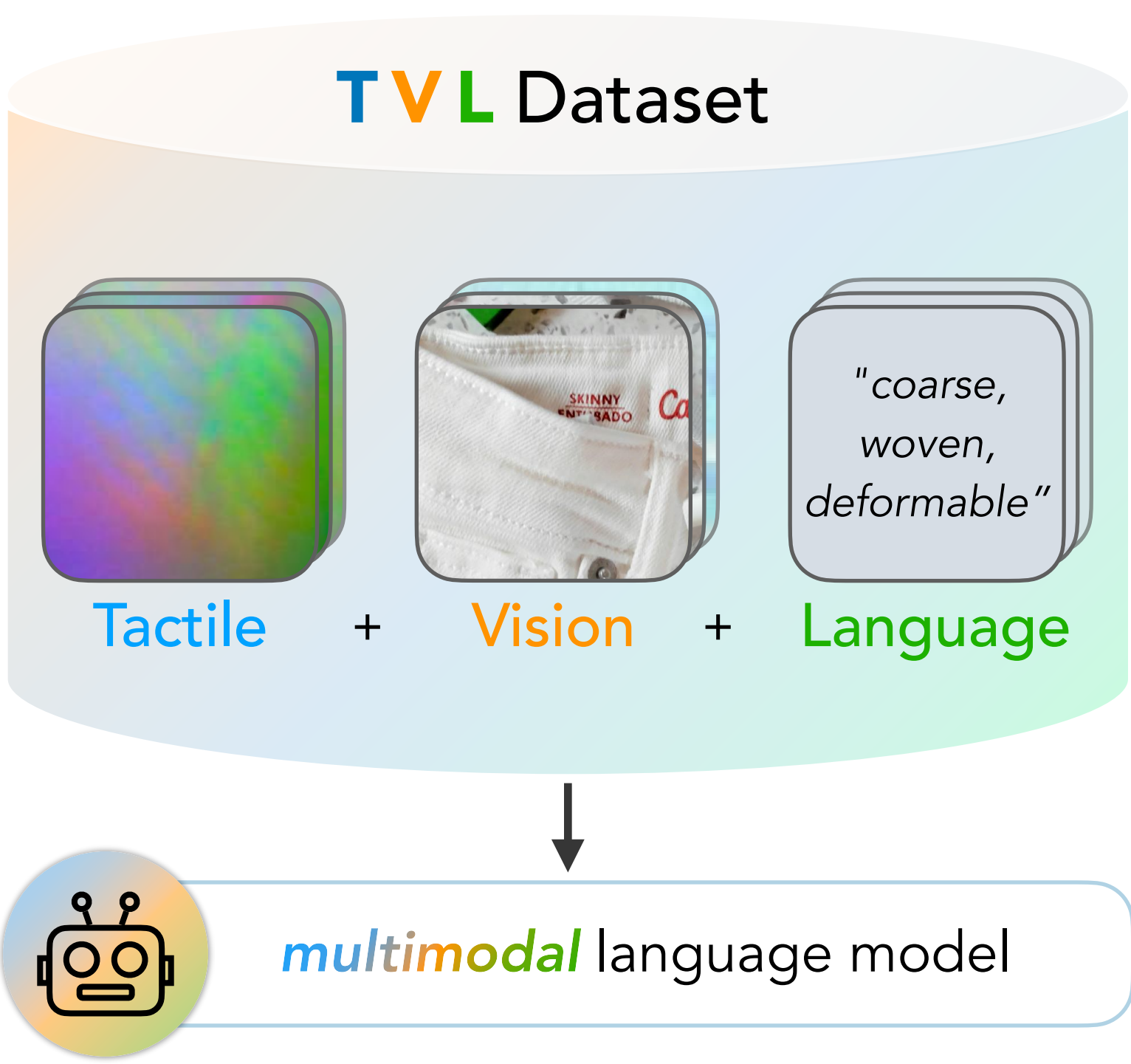
Examples




Model	Description	Score (1-10)
Human Labels	smooth, hard, sleek	GT
TVL-LLaMA	reflective, cool, sleek, smooth.	8
GPT-4V	smooth, reflective, hard, cool, sleek	8
SSVTP-LLaMA	smooth, reflective, hard, cool, sleek.	8.5
LLaVA 1.5 7B:	Slick, cold, smooth, hard, reflective.	8.5
LLaVA 1.5 13B	The surface is smooth, cold, hard, unyielding, and glossy.	8.5
ViP-LLaVA 7B	Slick, cold, smooth, shiny, reflective	7.5
ViP-LLaVA 13B	Glossy, hard, unyielding, smooth, shiny.	7.5
LLaMA-Adapter	1. Slick, hard, unyielding, cold, glossy.	8.5
BLIP-2 OPT-6.7	The surface is metallic.	6.5
InstructBLIP 7B	1 a person is using a black tablet on their lap	1
InstructBLIP 13B	a person uses a finger to press the light on a keyboard 3	1

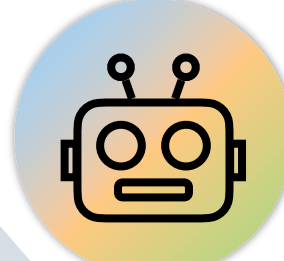
A Touch, Vision, Language Dataset for Multimodal Alignment

Max (Letian) Fu, Gaurav Datta*, Raven (Huang) Huang*, Will Panitch*, Jaimyn Drake*, Joseph Ortiz, Mustafa Mukadam, Mike Lambeta, Roberto Calandra, Ken Goldberg





How does this piece of fabric feel?



This fabric is soft, woven, and pliable.

Arxiv, Code, Dataset, Checkpoints

A Touch, Vision, and Language Dataset for Multimodal Alignment

Letian Fu¹, Gaurav Datta^{1*}, Huang Huang^{2,1}, William Chung-Ho Panitch¹, Jaimyn Drake¹, Joseph Ortiz², Mustafa Mukadam², Mike Lambeta², Roberto Calandra², Ken Goldberg¹


Abstract

Touch is an important sensing modality for humans, but it has not yet been incorporated into a multimodal generative language model. This is partially due to the difficulty of obtaining natural language labels for tactile data and the complexity of aligning tactile readings with both visual observations and language descriptions. As a step towards bridging that gap, this work introduces a new dataset of 44K in-the-wild vision-touch pairs, with English language labels annotated by humans (10%) and textual pseudo-labels from GPT-4V (90%). We use this dataset to train a vision-language-aligned tactile encoder for open-vocabulary classification and a touch-vision-language (TVL) model for text generation using the trained encoder. Results suggest that by incorporating touch, the TVL model improves (+29% classification accuracy) touch-vision-language alignment over existing models trained on any pair of these modalities. Although only a small fraction of the dataset is human labeled, the TVL model demonstrates improved visual tactile understanding over GPT-4V (+12%) and open-source vision-language models (+32%) on a new touch-vision understanding benchmark. Code and data: <https://tactile-vlm.github.io>.

1. Introduction

Almost all biological perception is inherently multimodal (Harrison & DeGeler, 2004; Turk, 2014; Brack et al., 2022), enabling agents to reason and make decisions based on multiple streams of information. Recent research in artificial multimodal representation learning has explored linking modalities such as vision, language, audio, temperature, and robot actions (Radford et al., 2021; Gribner et al., 2021; Gribner et al., 2021; Branson et al., 2021; Radourovic et al., 2023). However, the tactile modality remains underexplored in multimodal understanding. Touch enables humans to distinguish surface textures, object materials, dimensions, and contact forces (Johansson & Flanagan, 2009; Doherty et al., 2009; Klatzky & Lederman, 2000). Tactile perception has also proven useful in robotic applications, particularly for contact-rich manipulation tasks (Lambeta et al., 2020; Doherty et al., 2009; Calandra et al., 2018; Yuan et al., 2017; Dave et al., 2024; Qi et al., 2023).

Many works also explore visual tactile association, build cross-modal generators, and leverage cross-modal prompting for material property, surface texture, and cloth classification on a closed set of vocabularies (Yang et al., 2022; Dave et al., 2024; Li & Adelson, 2013; Ojala et al., 2002; Kamposoris et al., 2016; Yuan et al., 2018; Kerr et al., 2023). However, human tactile perception captures more than tactile-visual associations; the tactile modality captures diverse semantic information and demonstrates deep integration with language (Schmidt et al., 2019; Speed et al., 2021; Miller et al., 2018; gharrett, 2023). One major obstacle to the integration of touch and language is the scarcity of diverse data. While recent work has collected both datasets of paired tactile and visual observations and human-labeled datasets for tactile-based texture or material classification, we are not aware of any tactile dataset that contains open vocabulary language labels. Therefore, we develop a custom



arXiv:2402.13232v1 [cs.CV] 20 Feb 2024