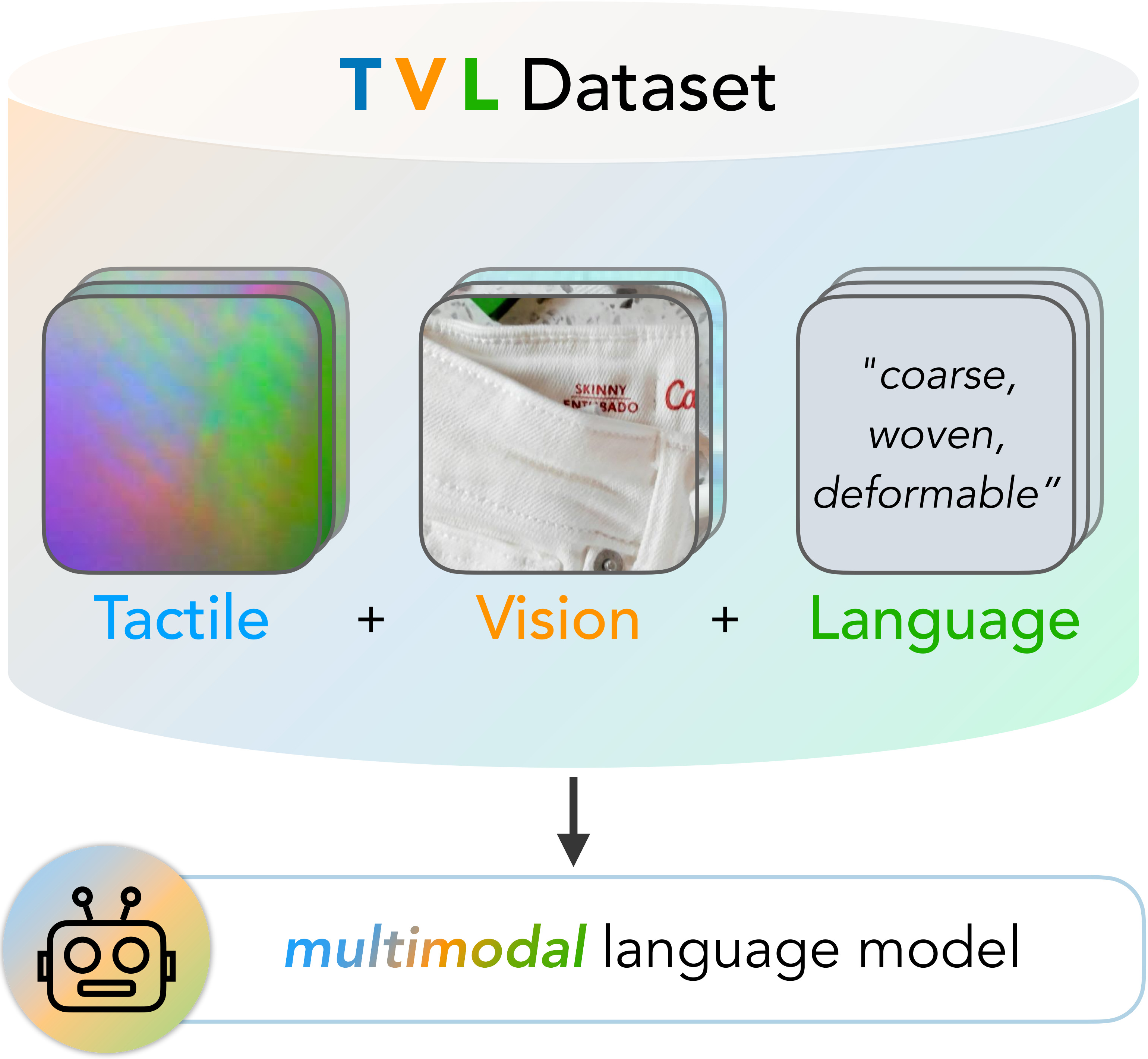
A Touch, Vision, and Language Dataset for Multimodal Alignment

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

Berkeley UNIVERSITY OF CALIFORNIA
Meta
TECHNISCHE UNIVERSITÄT DRESDEN

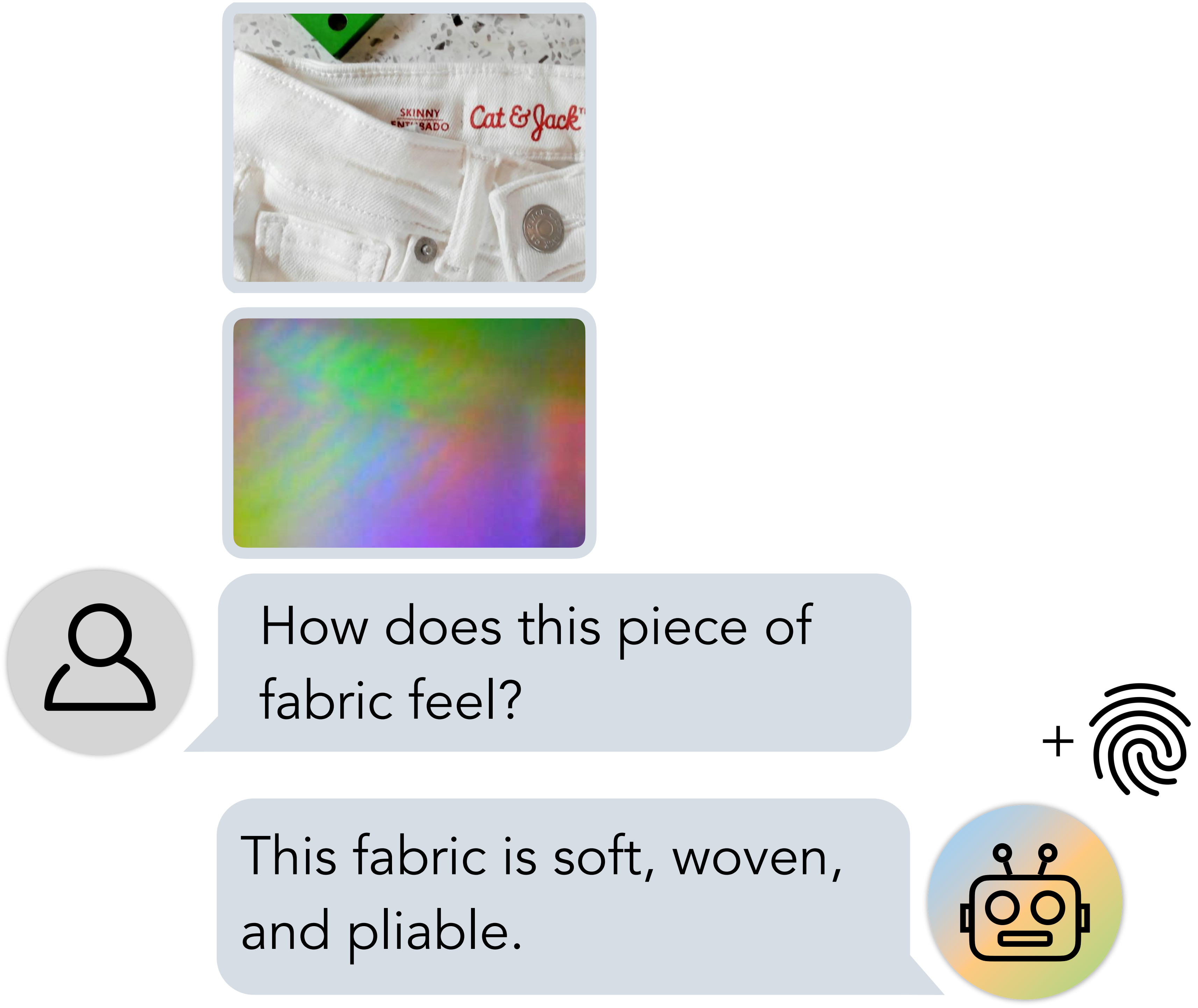


TVL Dataset

"coarse, woven, deformable"

Tactile + Vision + Language

multimodal language model



How does this piece of fabric feel?

This fabric is soft, woven, and pliable.

How?

Cross-Modal Supervision

"Don't play with sharp scissors!"

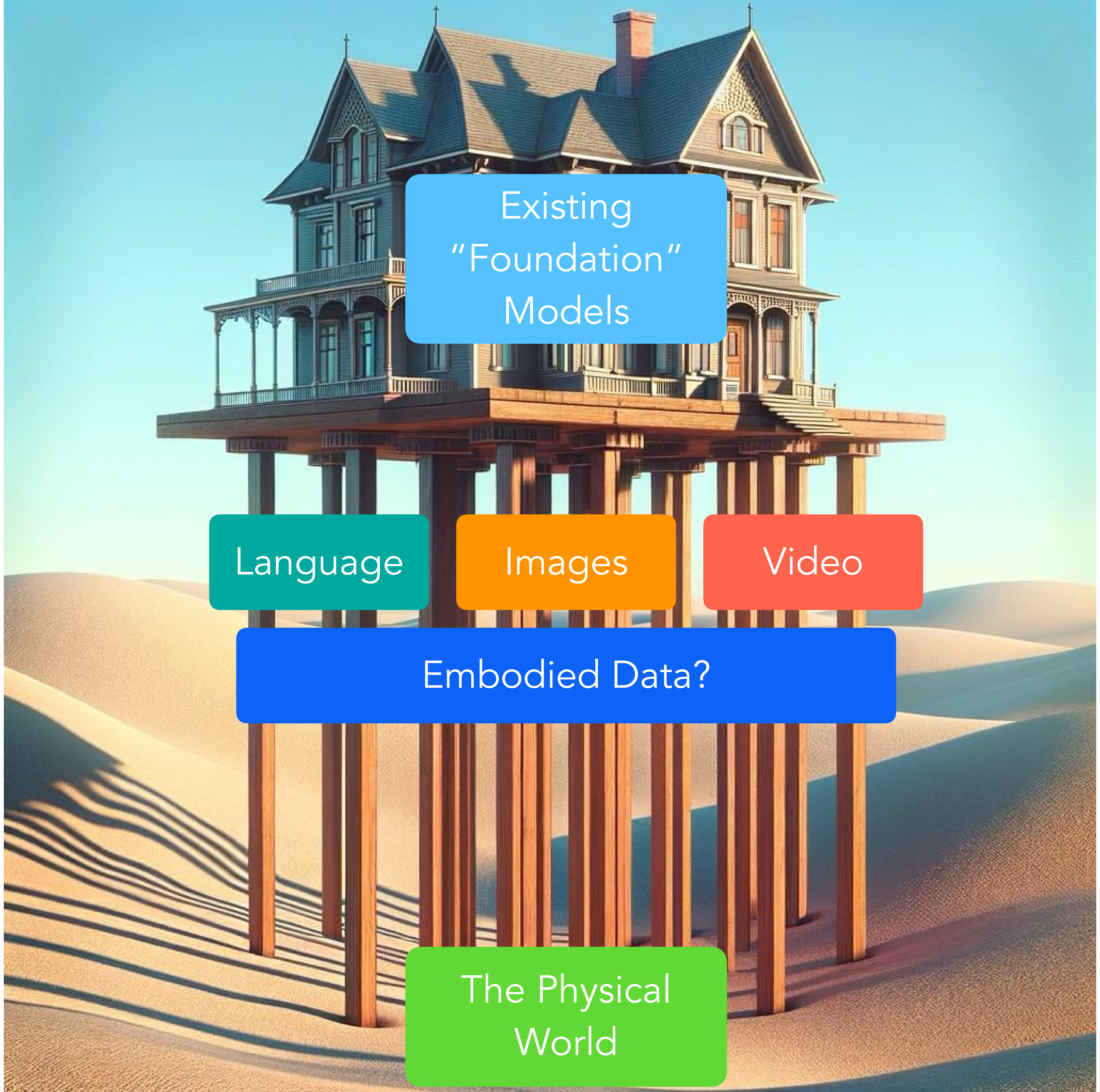
Multimodal Alignment

CLIP [1] ImageBind [2] GPT-4V [3] LLaVA [4] Flamingo [5]

Touch as a sensing modality is **missing** in multimodal models

[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.

Existing
“Foundation”
Models



Existing "Foundation" Models

Language

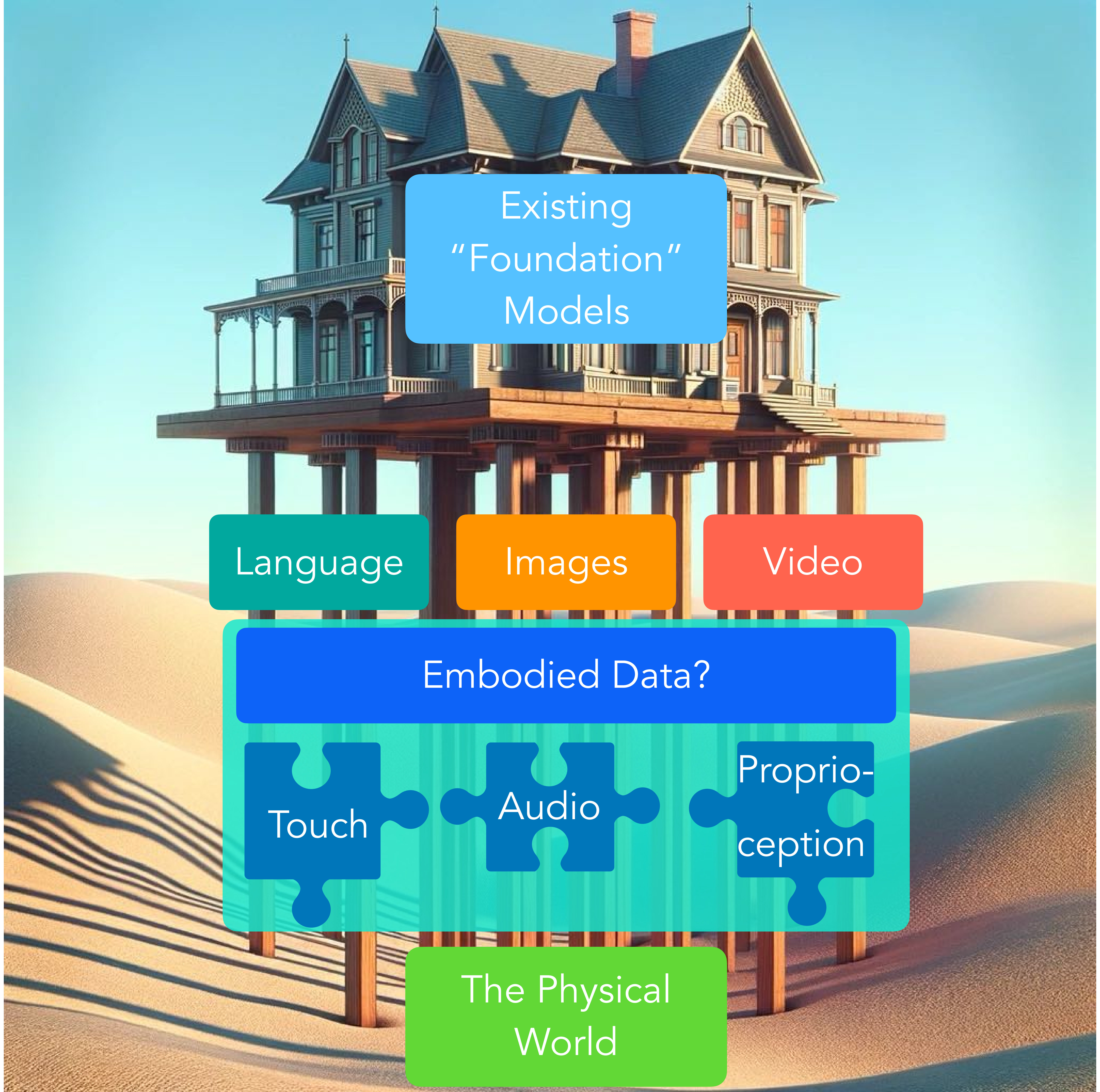
Images

Video

Embodied Data?

The Physical World

CR: Andrew Owens, DALLE 3



Existing
“Foundation”
Models

Language

Images

Video

Embodied Data?

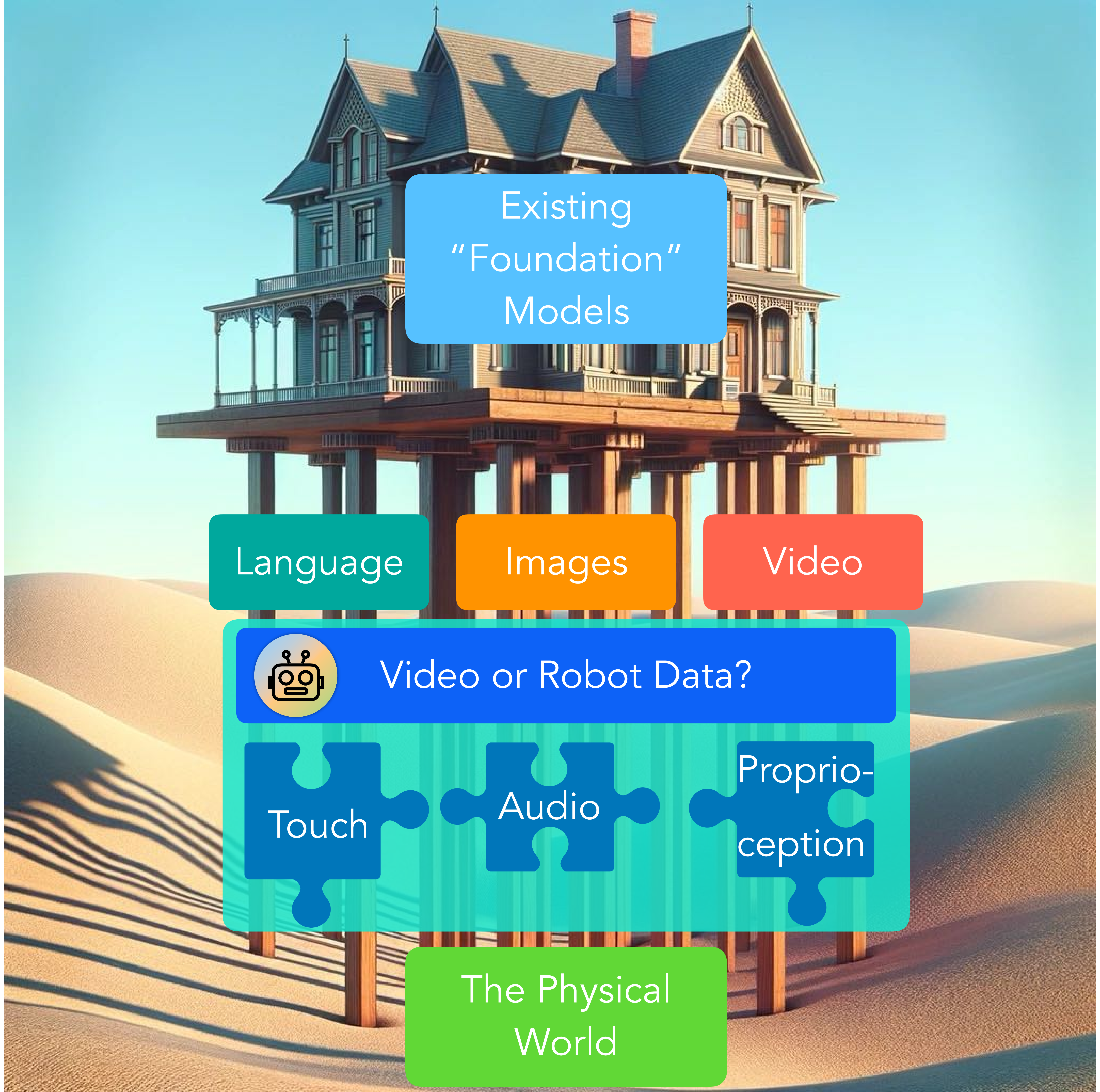
Touch

Audio

Proprio-
ception

The Physical
World

CR: Andrew Owens, DALLE 3



Existing "Foundation" Models

Language

Images

Video

Video or Robot Data?

Touch

Audio

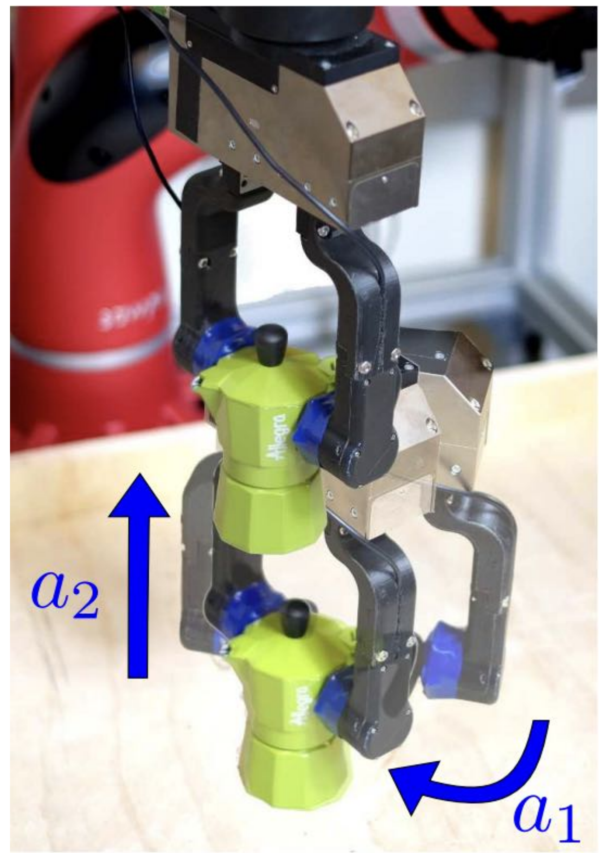
Proprio-ception

The Physical World

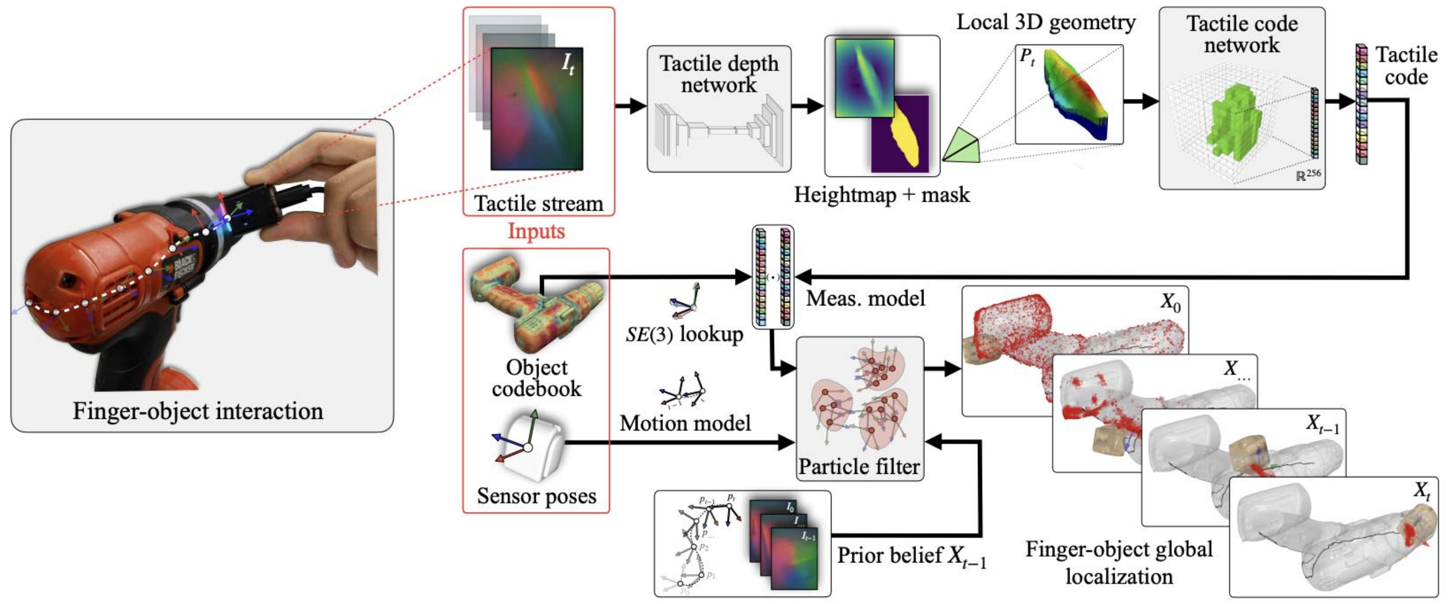
CR: Andrew Owens, DALLE 3

Tactile Perception for Robotics

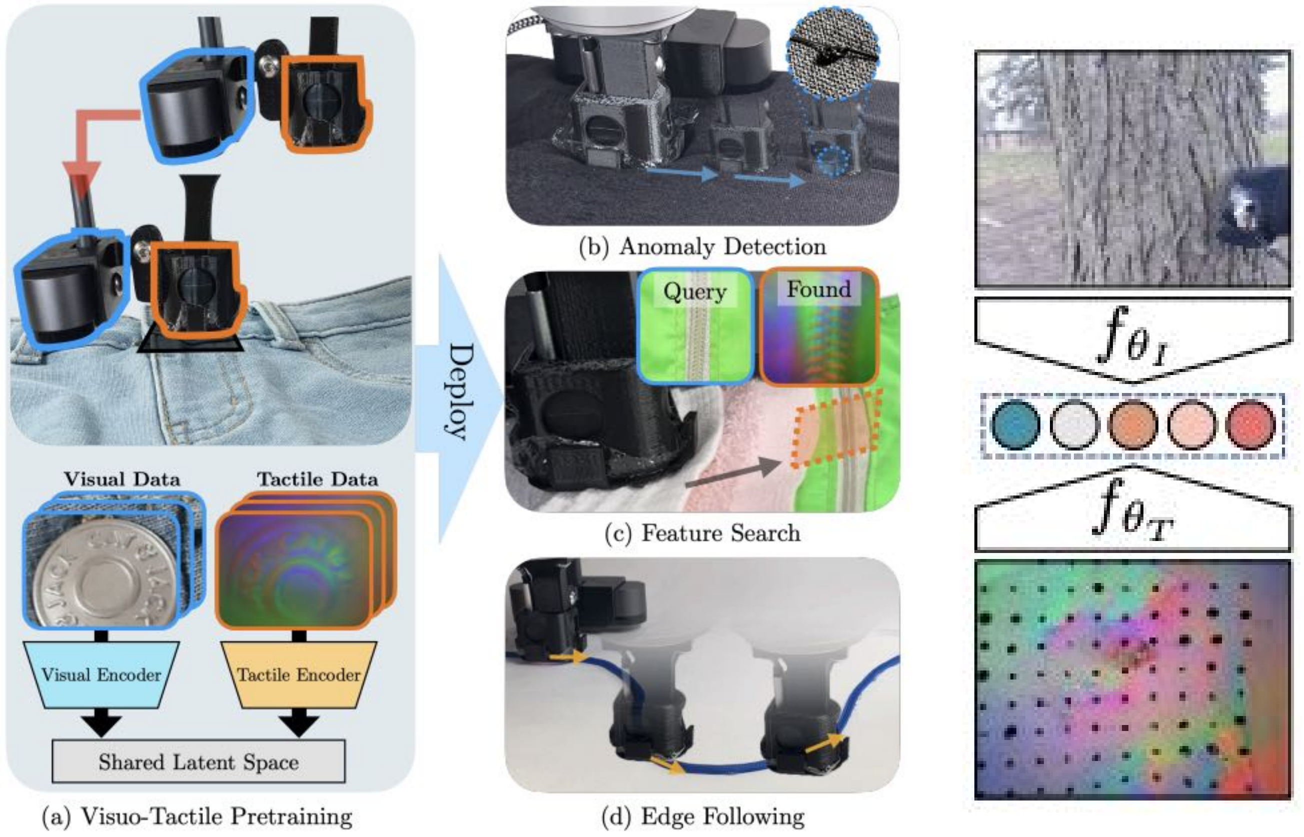
Cloth Classification [1]



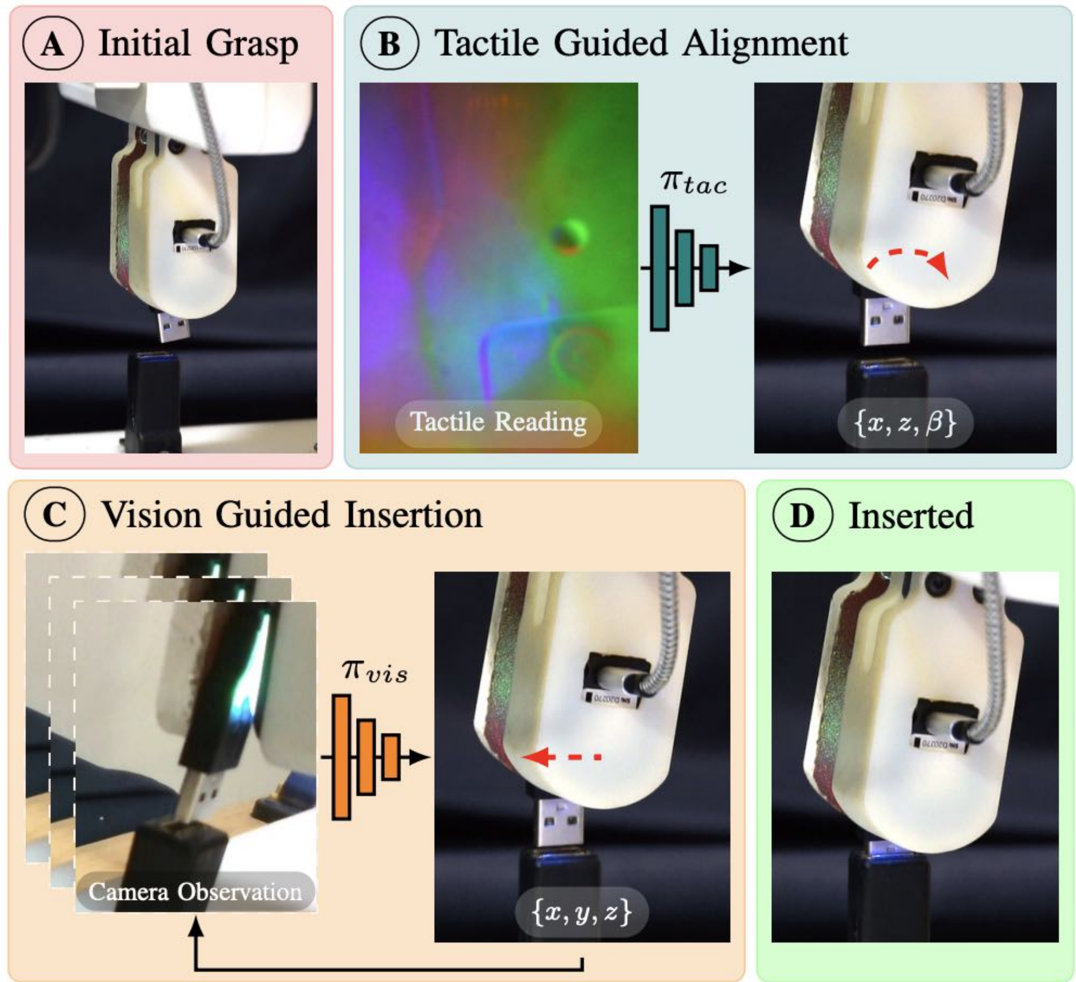
Grasping and Regrasping [2]



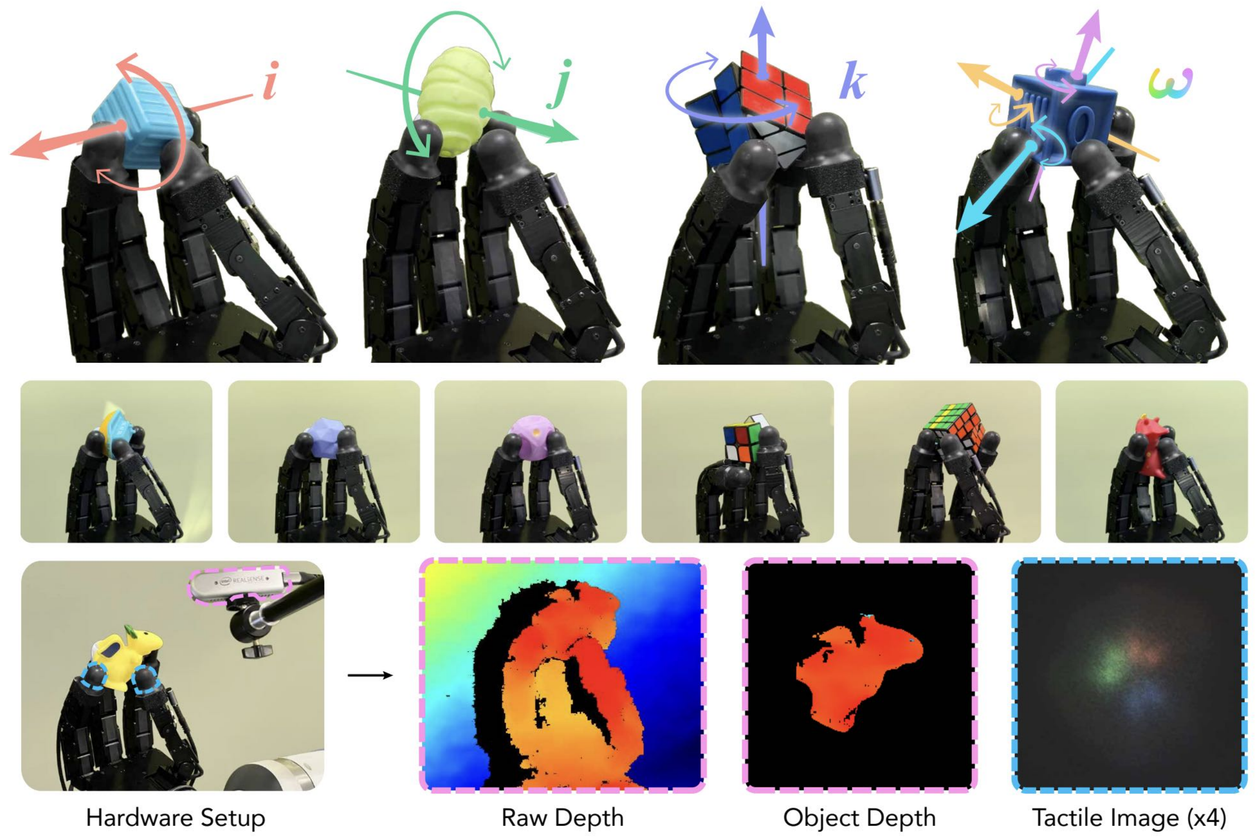
Pose Estimation [3]



Pretraining [4,5]



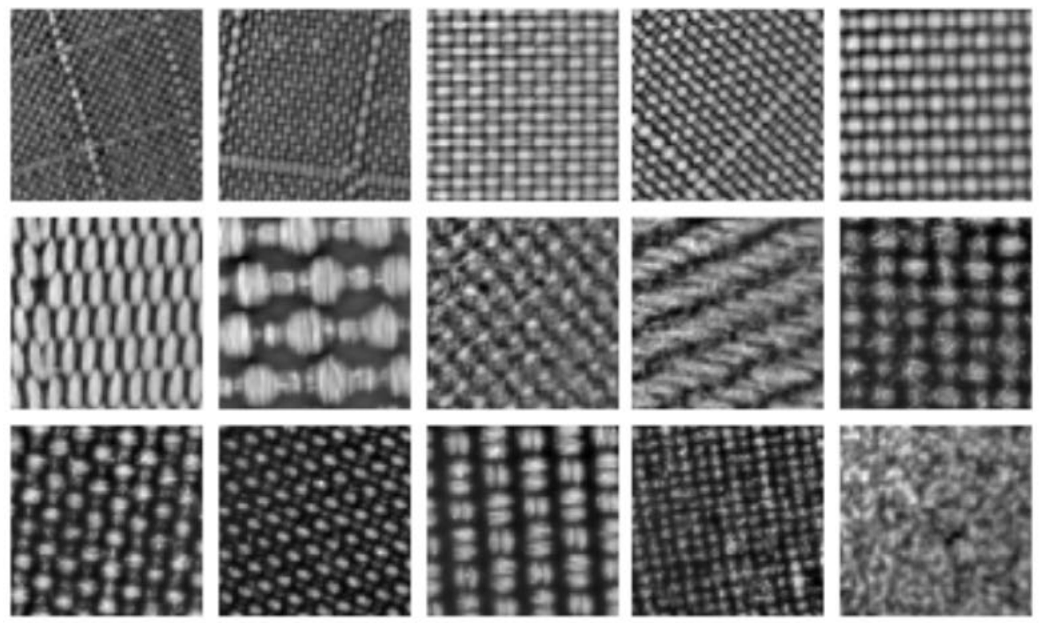
Industrial Insertion [6]



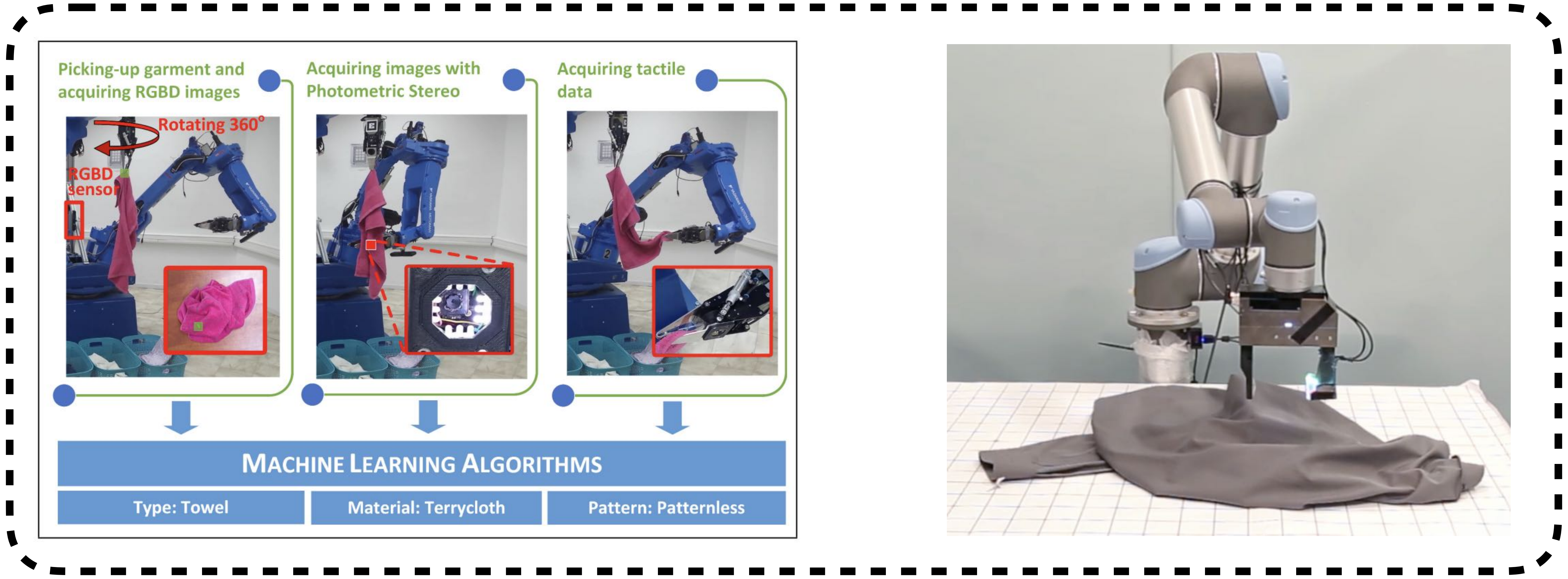
General In-Hand Rotation [7]

[1] Yuan, Wenzhen et al. "Active clothing material perception using tactile sensing and deep learning."ICRA 2018.
[2] Calandra, Roberto el al. "More than a feeling: Learning to grasp and regrasp using vision and touch." RAL 2018.
[3] Suresh, Sudharshan et al. "MidasTouch: Monte-Carlo inference over distributions across sliding touch." CoRL 2023.
[4] Kerr, Justin et al. "Self-supervised visuo-tactile pretraining to locate and follow garment features." RSS 2023.
[5] Yang, Fengyu et al. "Touch and Go: Learning from Human-Collected Vision and Touch." NeurIPS 2022.
[6] Fu, Letian et al. "Safe Self-Supervised Learning in Real of Visuo-Tactile Feedback Policies for Industrial Insertion." ICRA 2023.
[7] Qi, Haozhi et al. "General In-Hand Object Rotation with Vision and Touch." CoRL 2023.

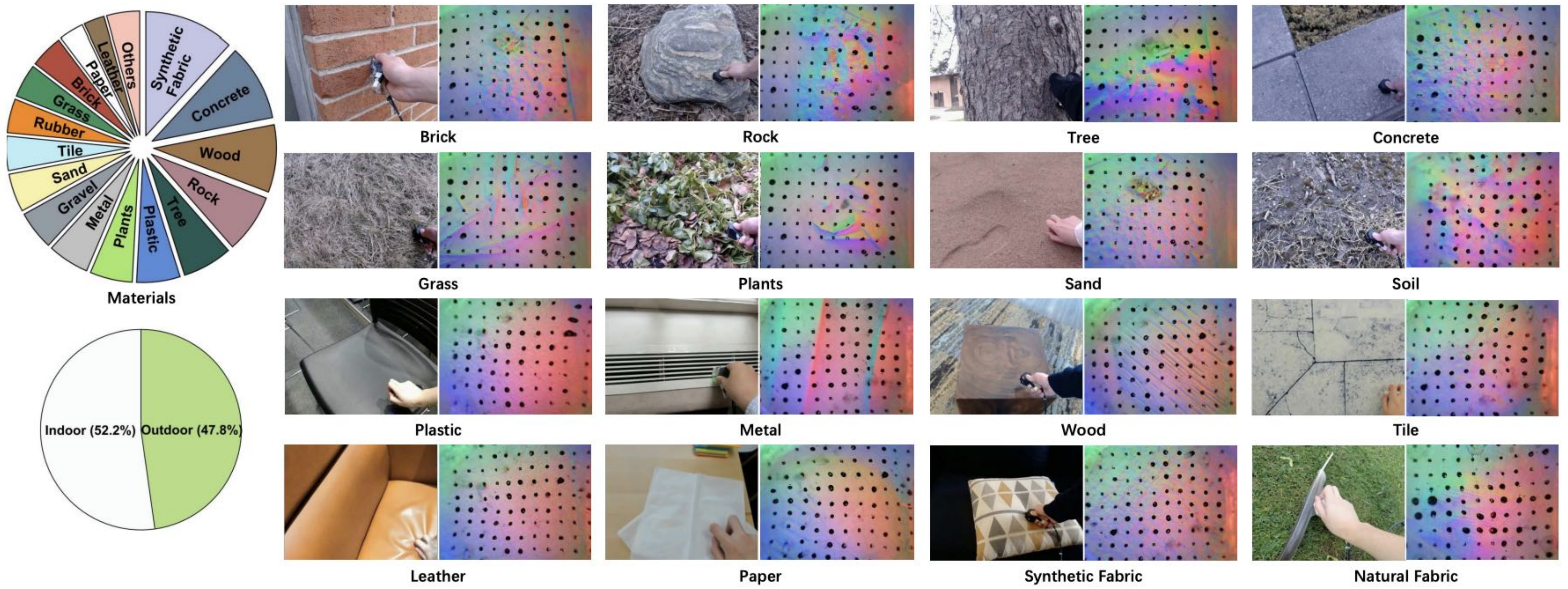
🗣️ Touch was not yet associated with open vocabulary descriptions



Texture Classification [1]


Picking-up garment and acquiring RGBD images
Rotating 360°
RGBD sensor
Acquiring images with Photometric Stereo
Acquiring tactile data
MACHINE LEARNING ALGORITHMS
Type: Towel
Material: Terrycloth
Pattern: Patternless

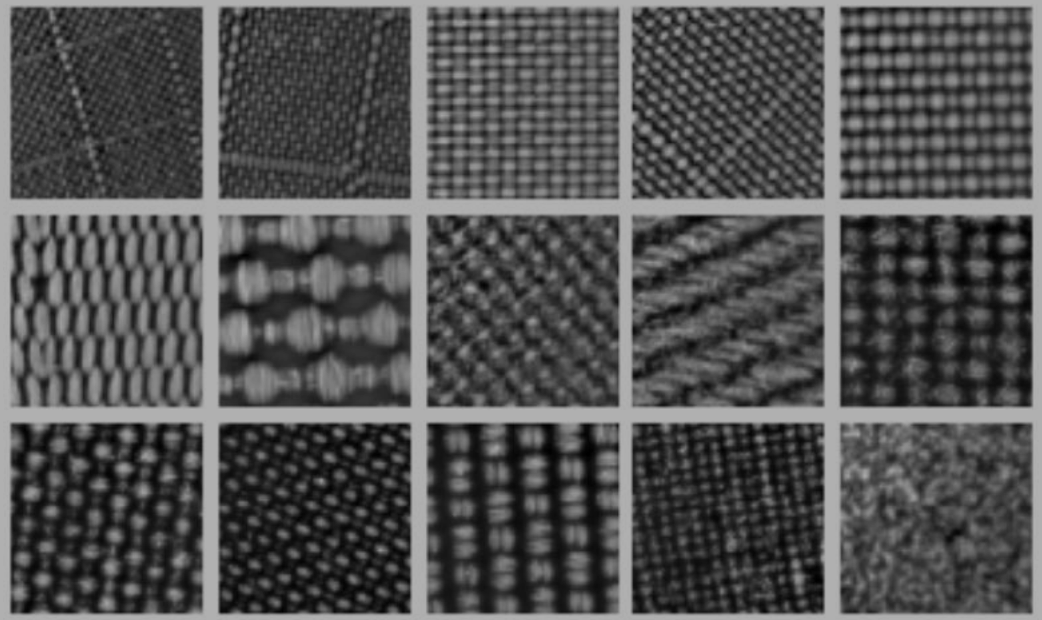
Cloth Classification [2,3]

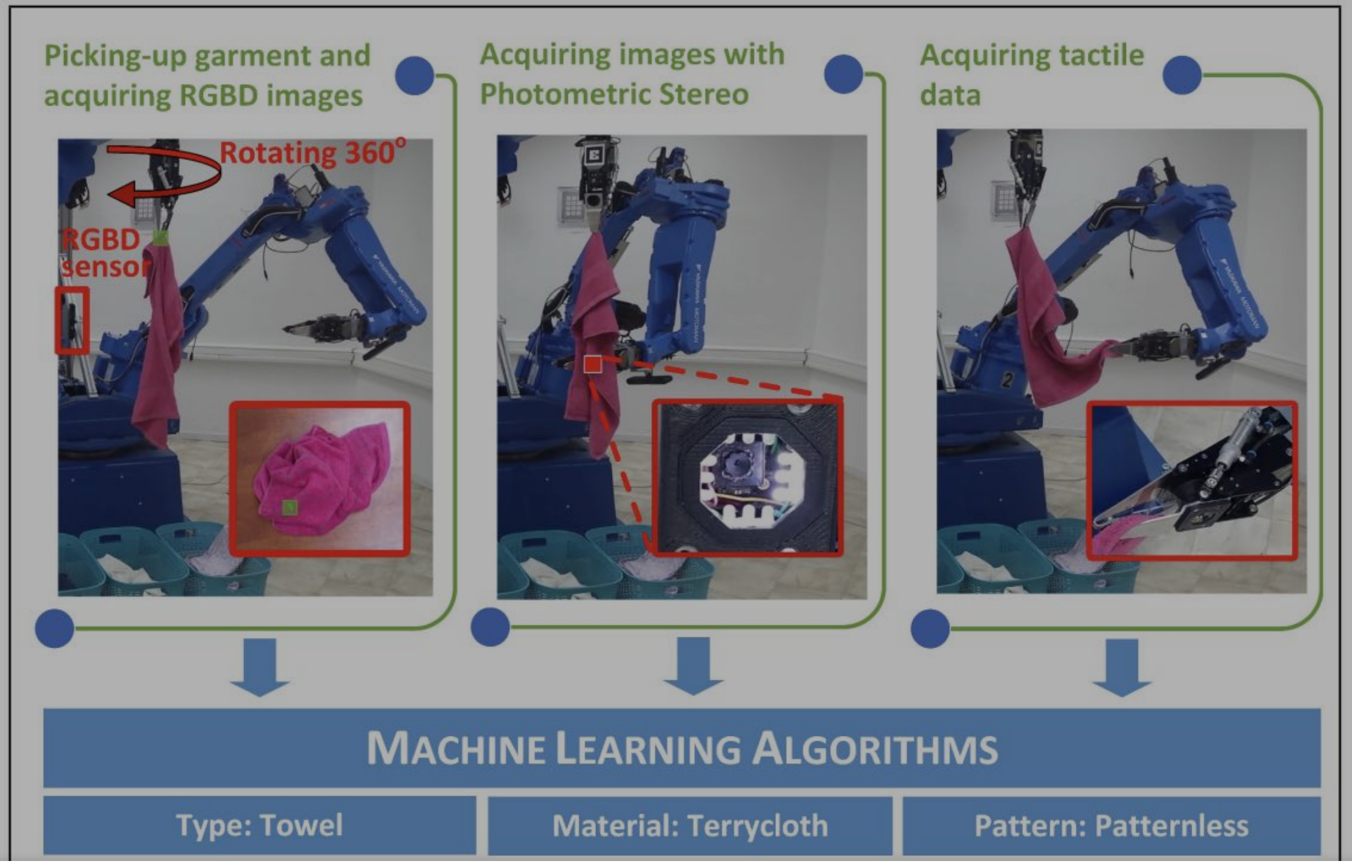


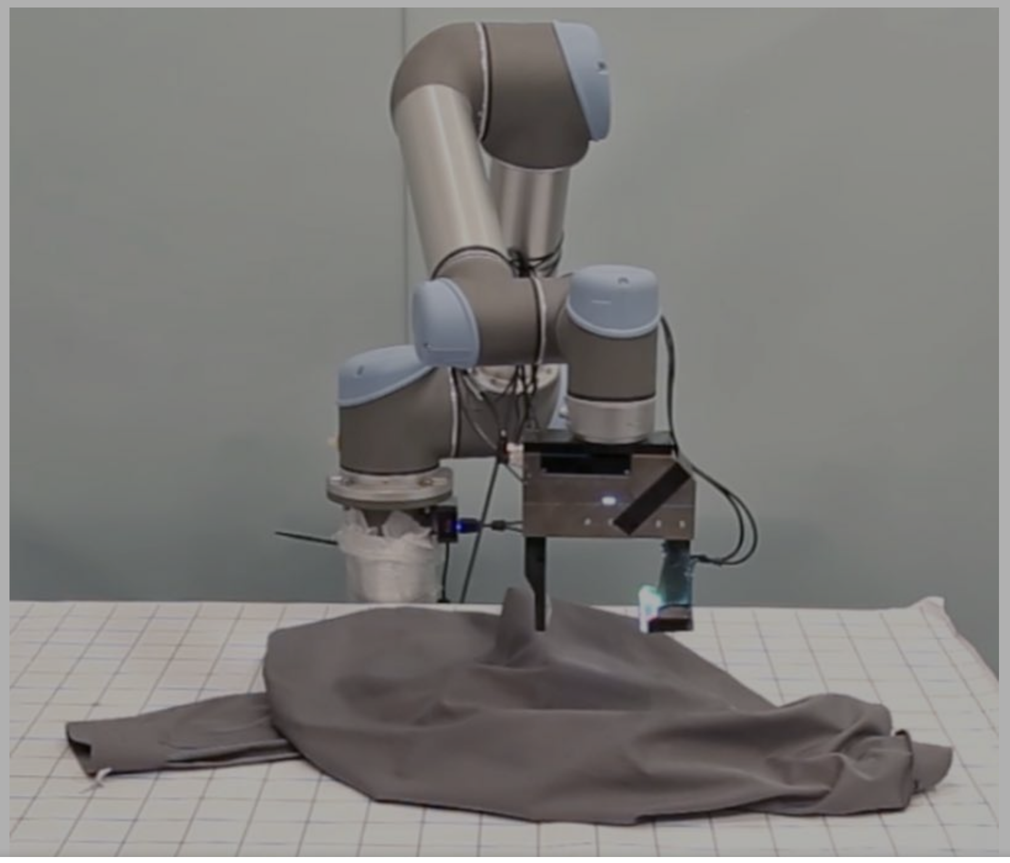
"In-the-wild" Texture Classification [4]

[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.

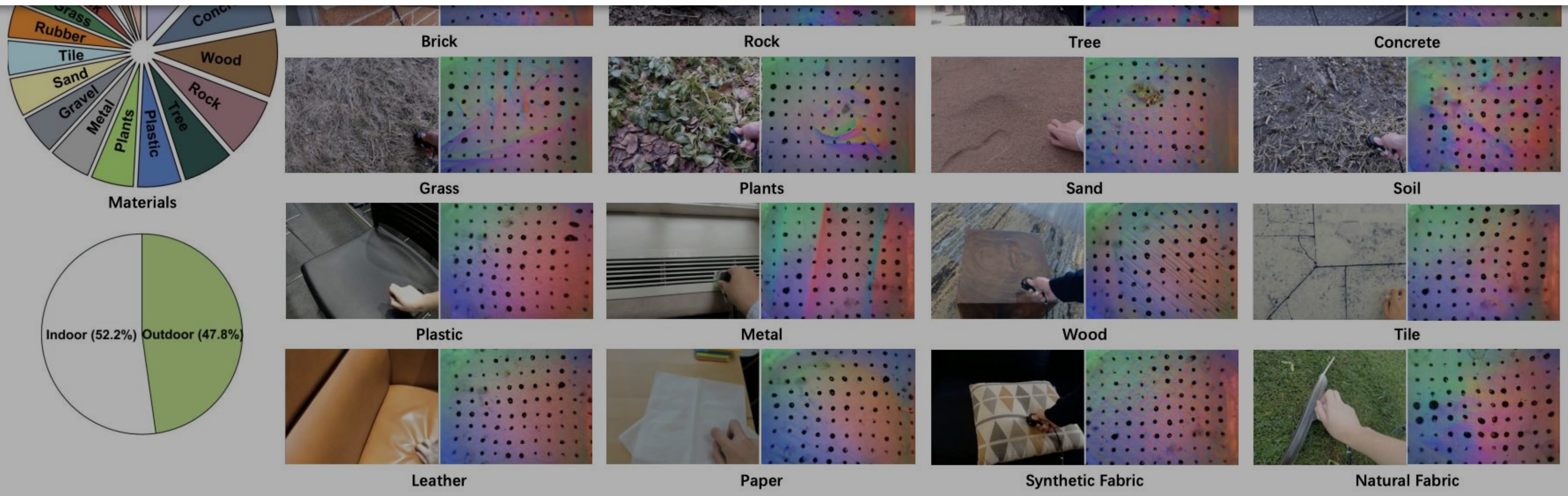
🗣️ Touch was not yet associated with open vocabulary descriptions




Picking-up garment and acquiring RGBD images
Rotating 360°
RGBD sensor
Acquiring images with Photometric Stereo
Acquiring tactile data
MACHINE LEARNING ALGORITHMS
Type: Towel
Material: Terrycloth
Pattern: Patternless



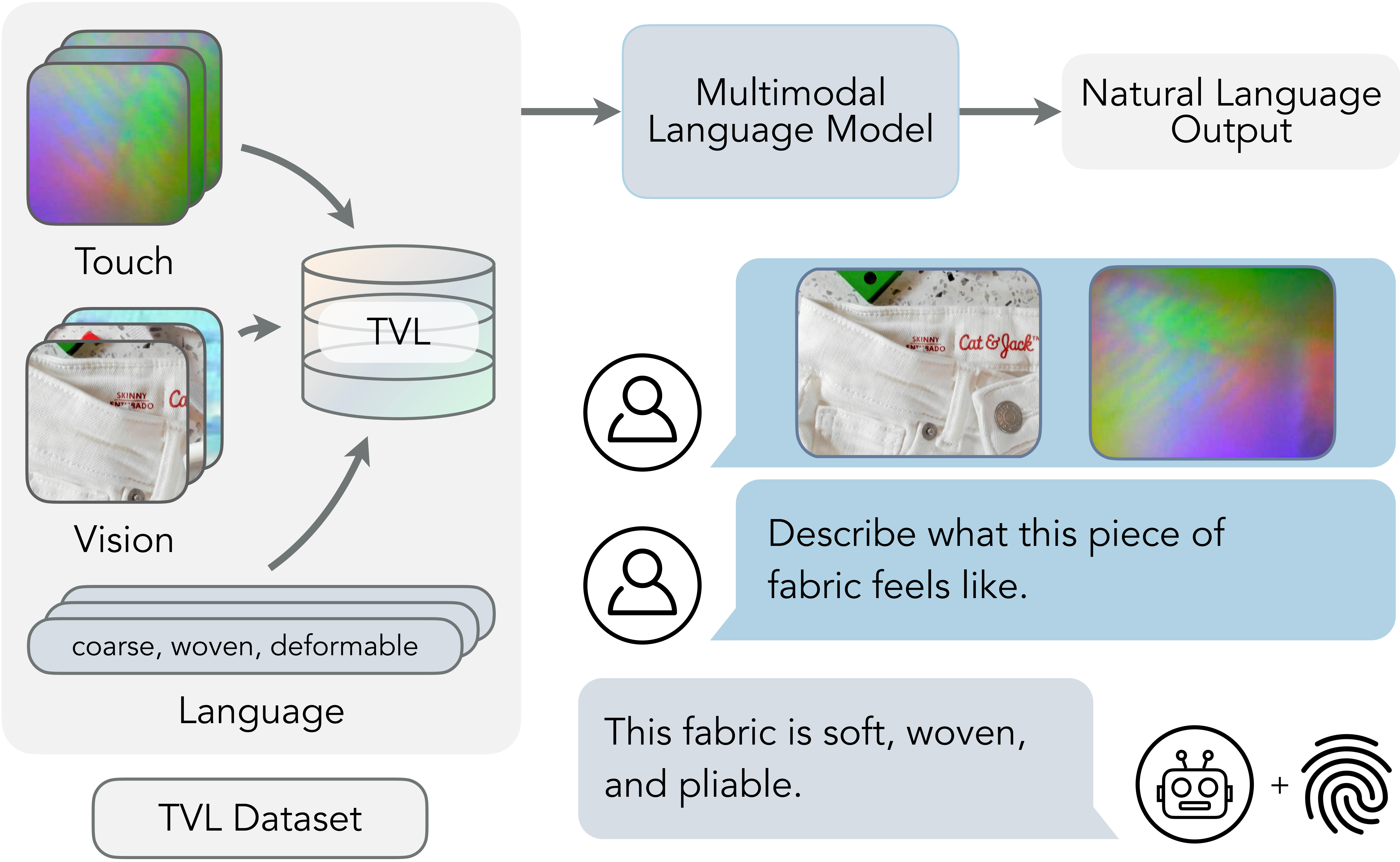
Touch, Vision, Language?



"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



Touch

Vision

TVL

coarse, woven, deformable

Language

TVL Dataset

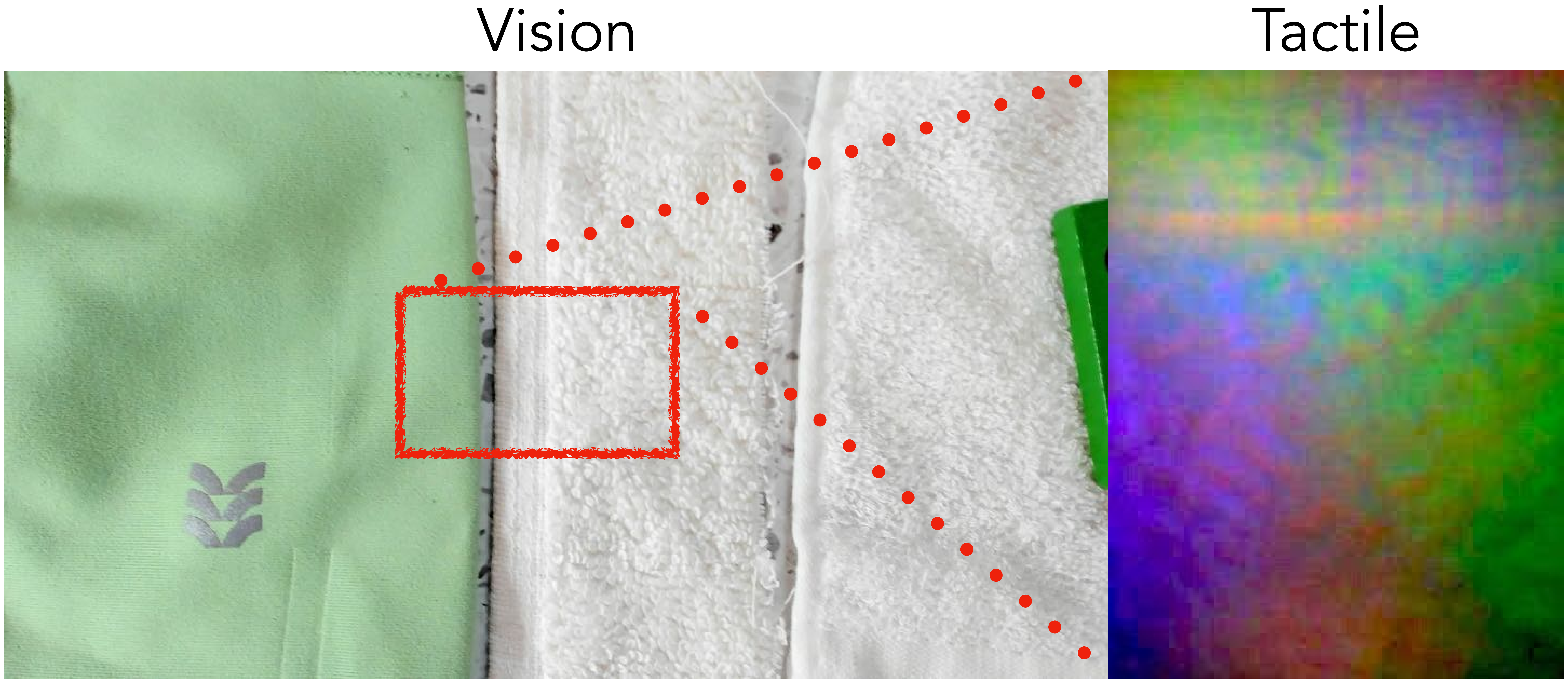
Multimodal Language Model

Natural Language Output

Describe what this piece of fabric feels like.

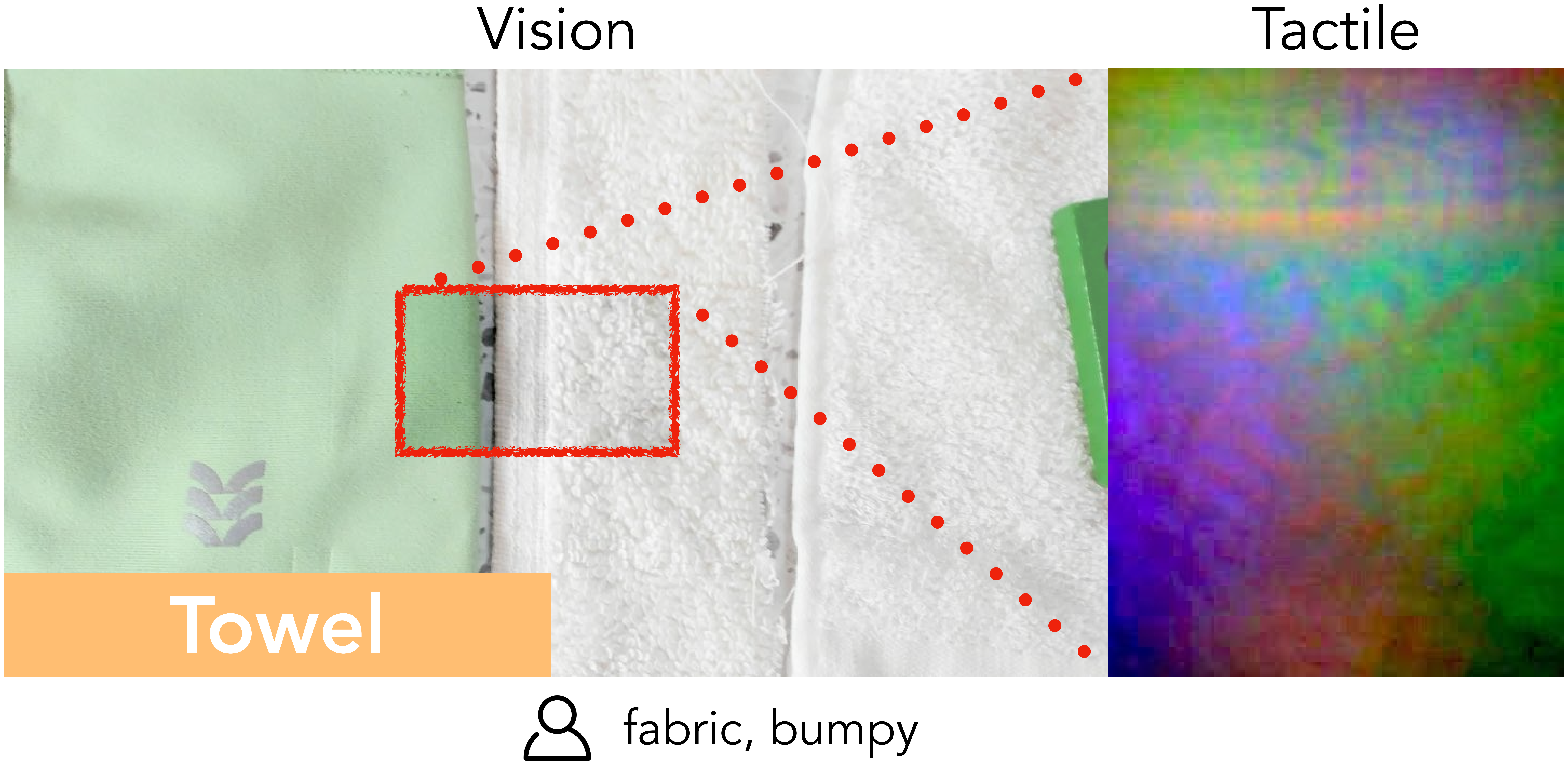
This fabric is soft, woven, and pliable.

TVL Dataset



[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.

TVL Dataset

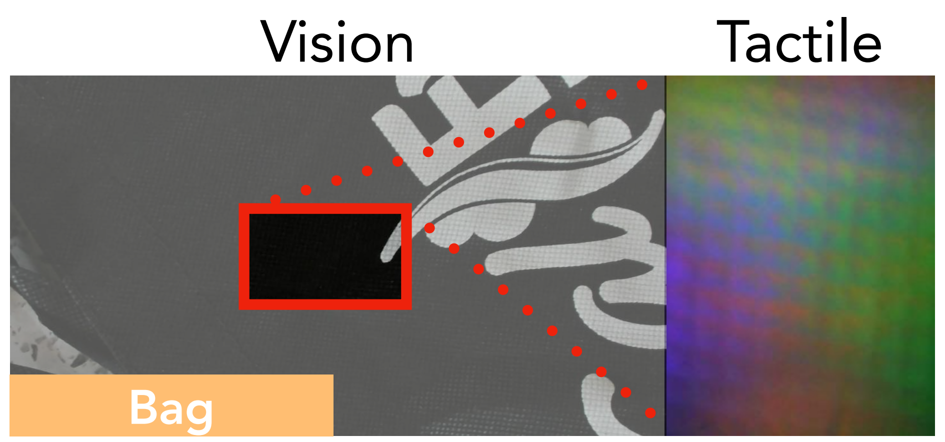


[1] Kerr, Justin et al. "Self-supervised visuo-tactile pretraining to locate and follow garment features." RSS 2023.
[2] Barnett, A.J. "400 Words to Describe Texture." 2023.

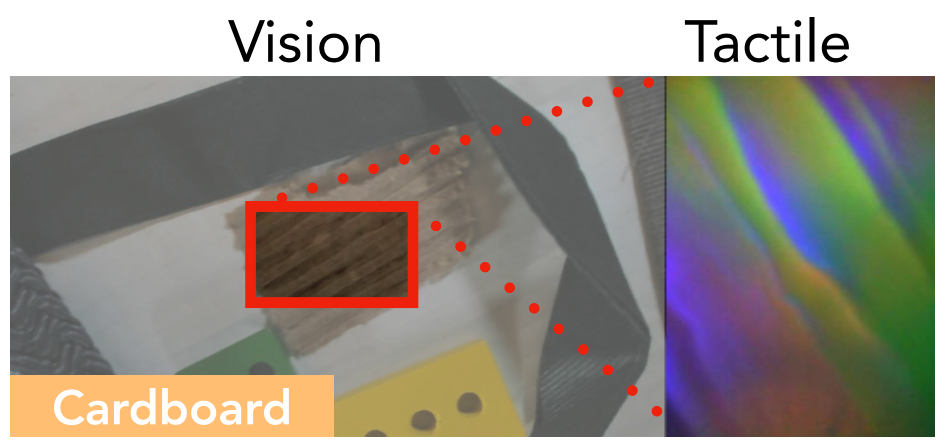
TVL Dataset

SSVTP [1]

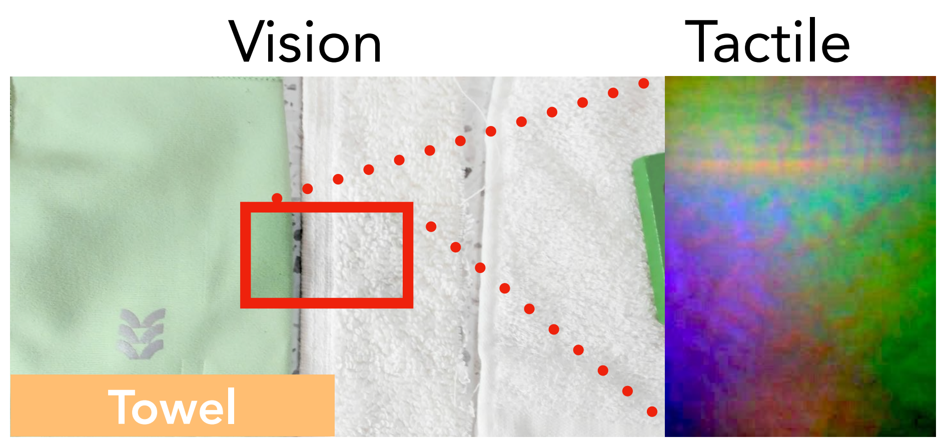
4.6K Human
Annotations



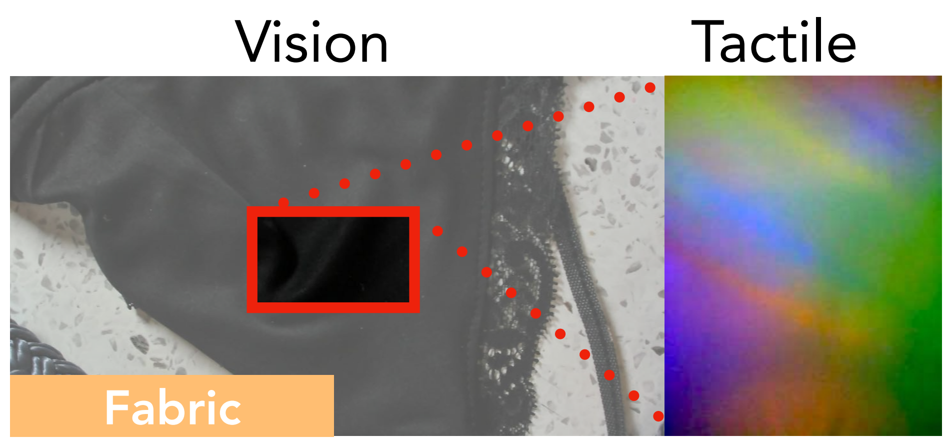
fabric, coarse



lined, cardboard, creased



fabric, bumpy



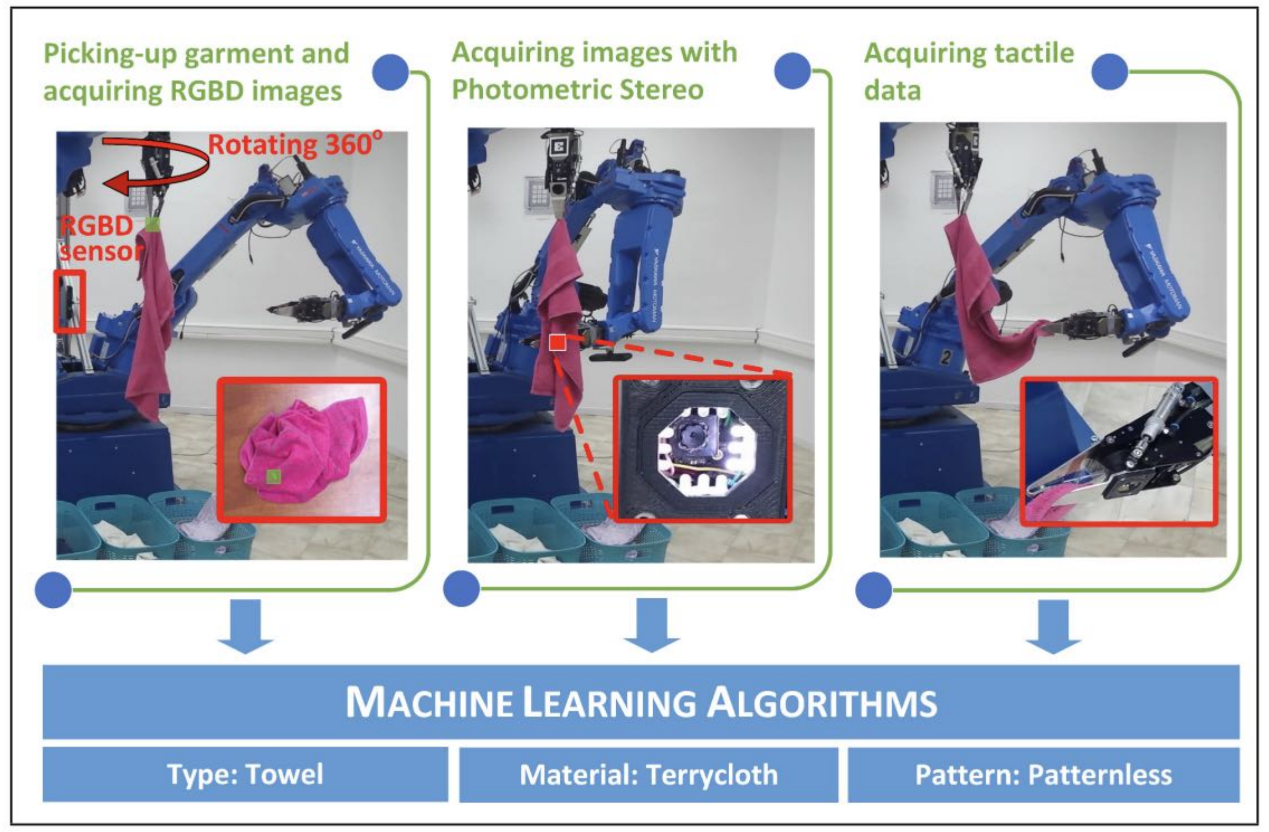
deformable, grainy

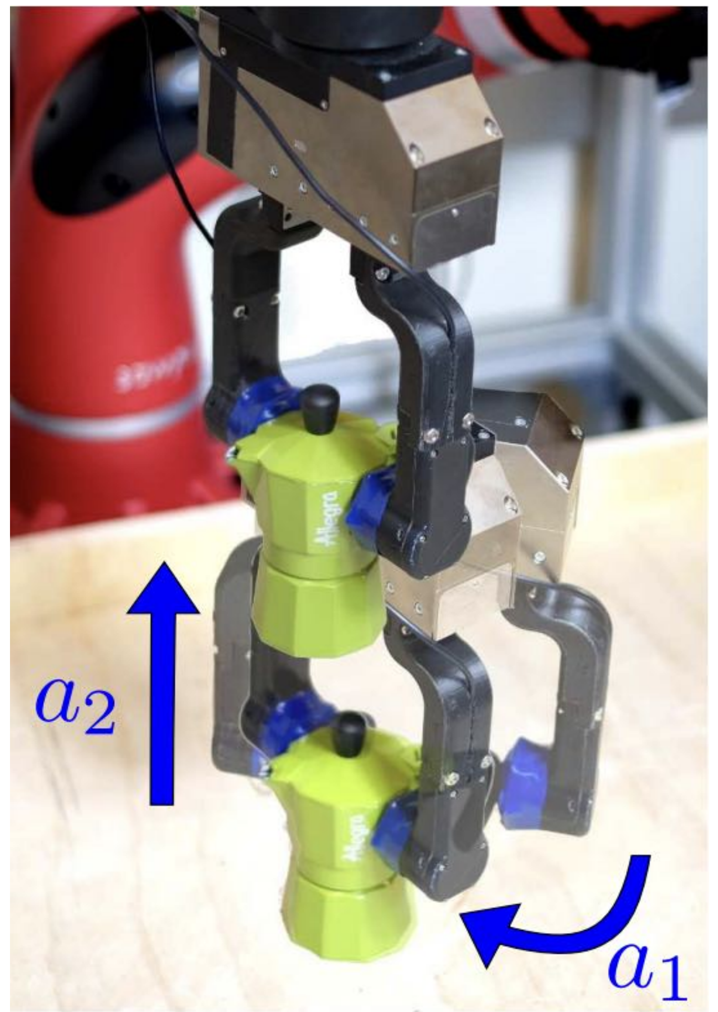
---

[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.

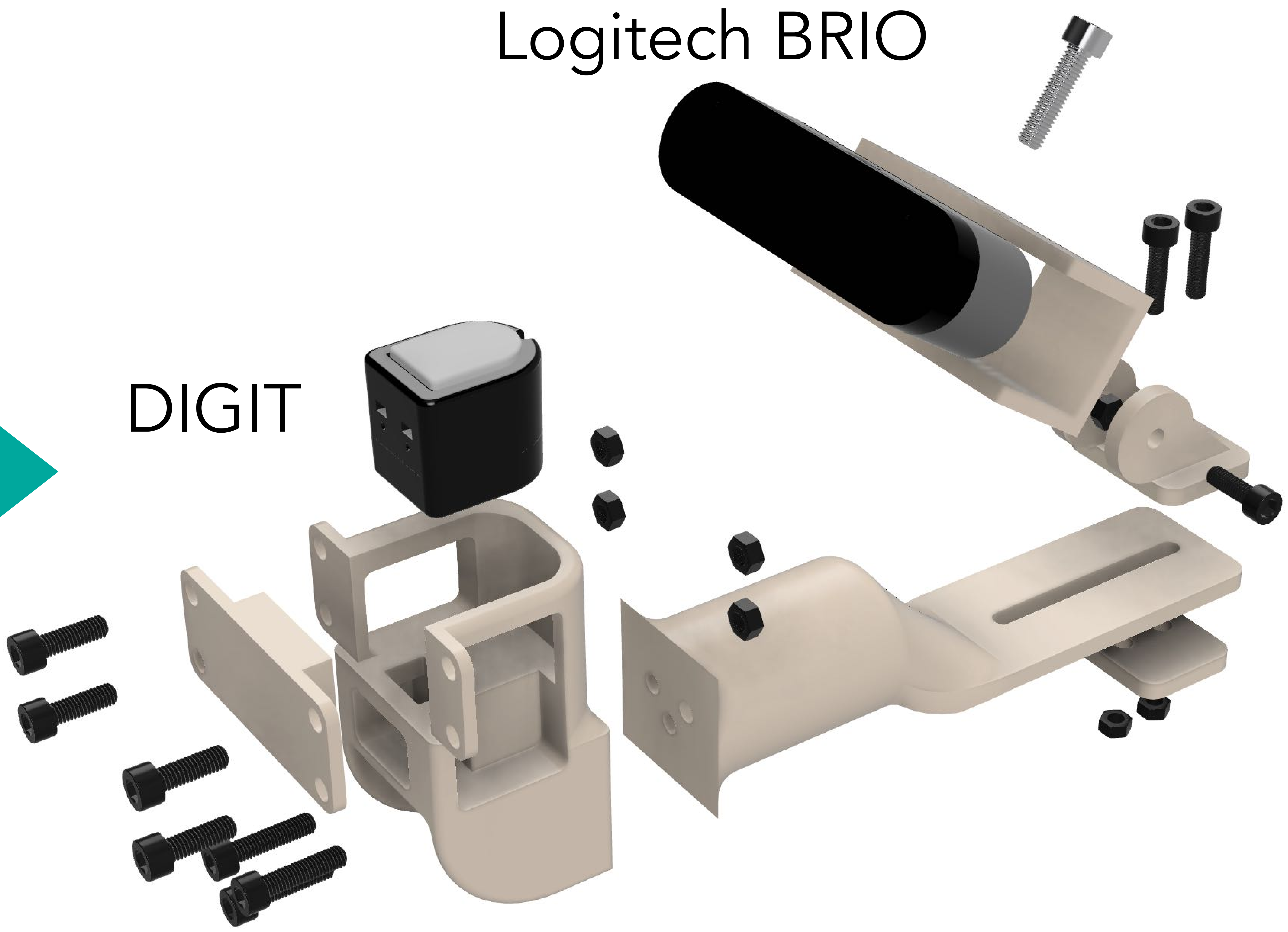
Data collection


Picking-up garment and acquiring RGBD images
Rotating 360°
RGBD sensor
Acquiring images with Photometric Stereo
Acquiring tactile data
Machine Learning Algorithms
Type: Towel
Material: Terrycloth
Pattern: Patternless



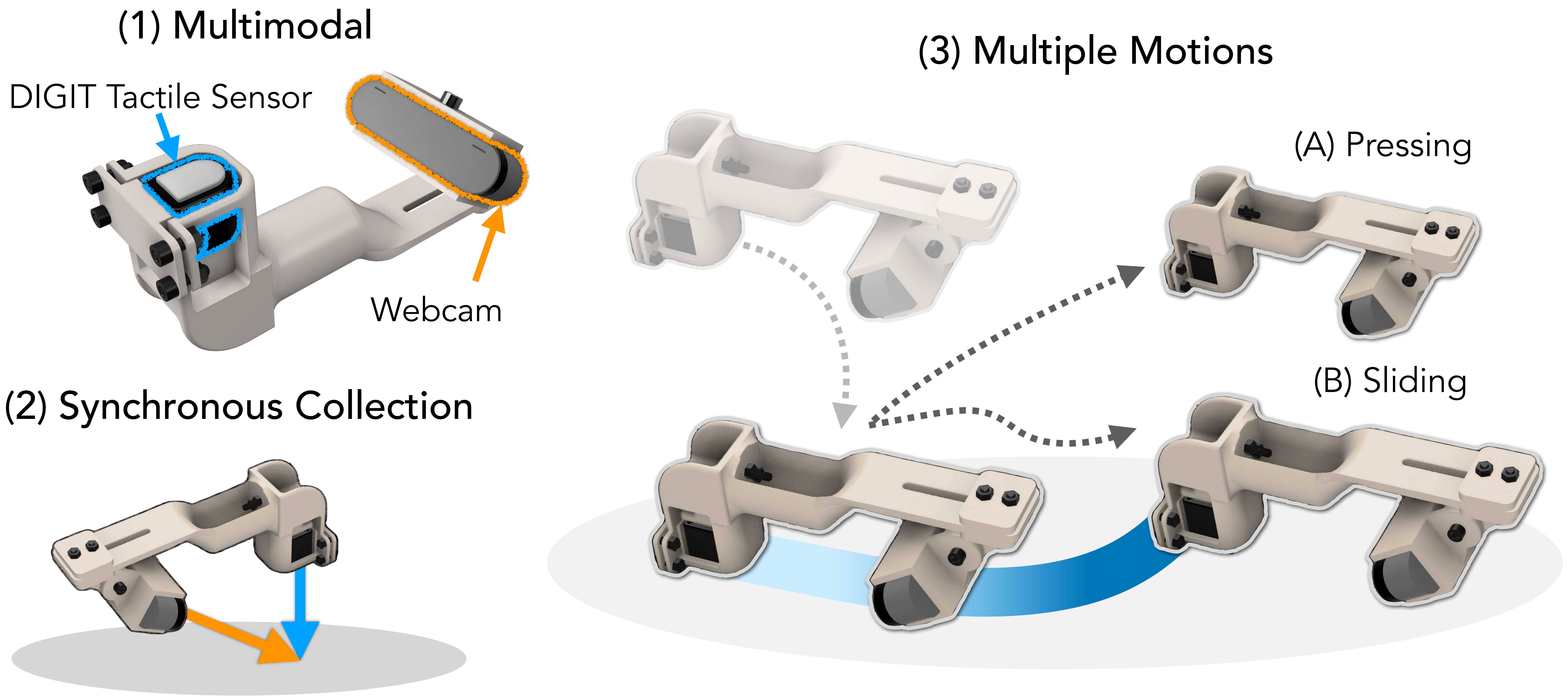
DIGIT Cover
DIGIT Housing
UR5 Mount
Webcam Holster

Controlled environments and objects



"In-the-wild" Device

Data collection



(1) Multimodal

DIGIT Tactile Sensor

Webcam

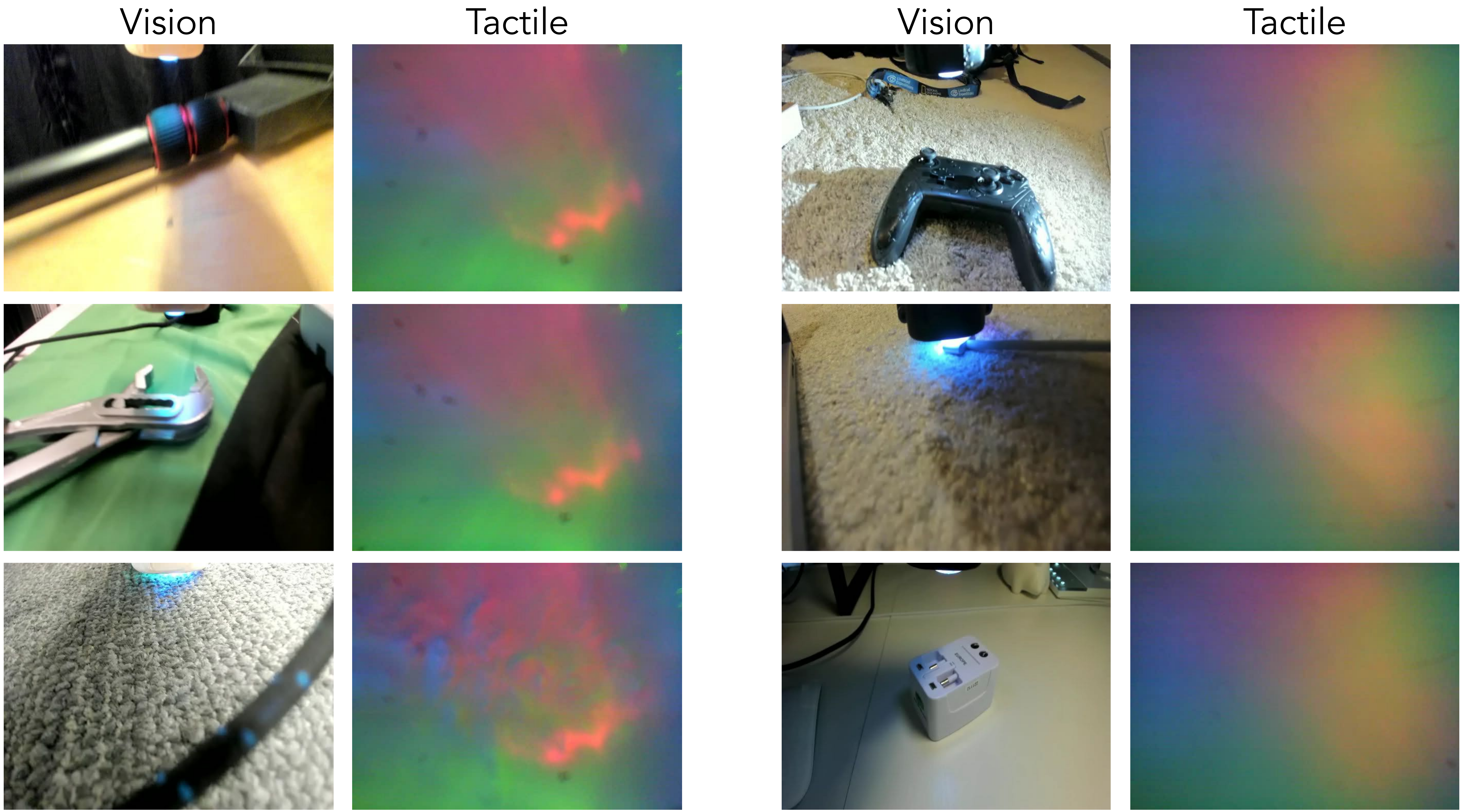
(2) Synchronous Collection

(3) Multiple Motions

(A) Pressing

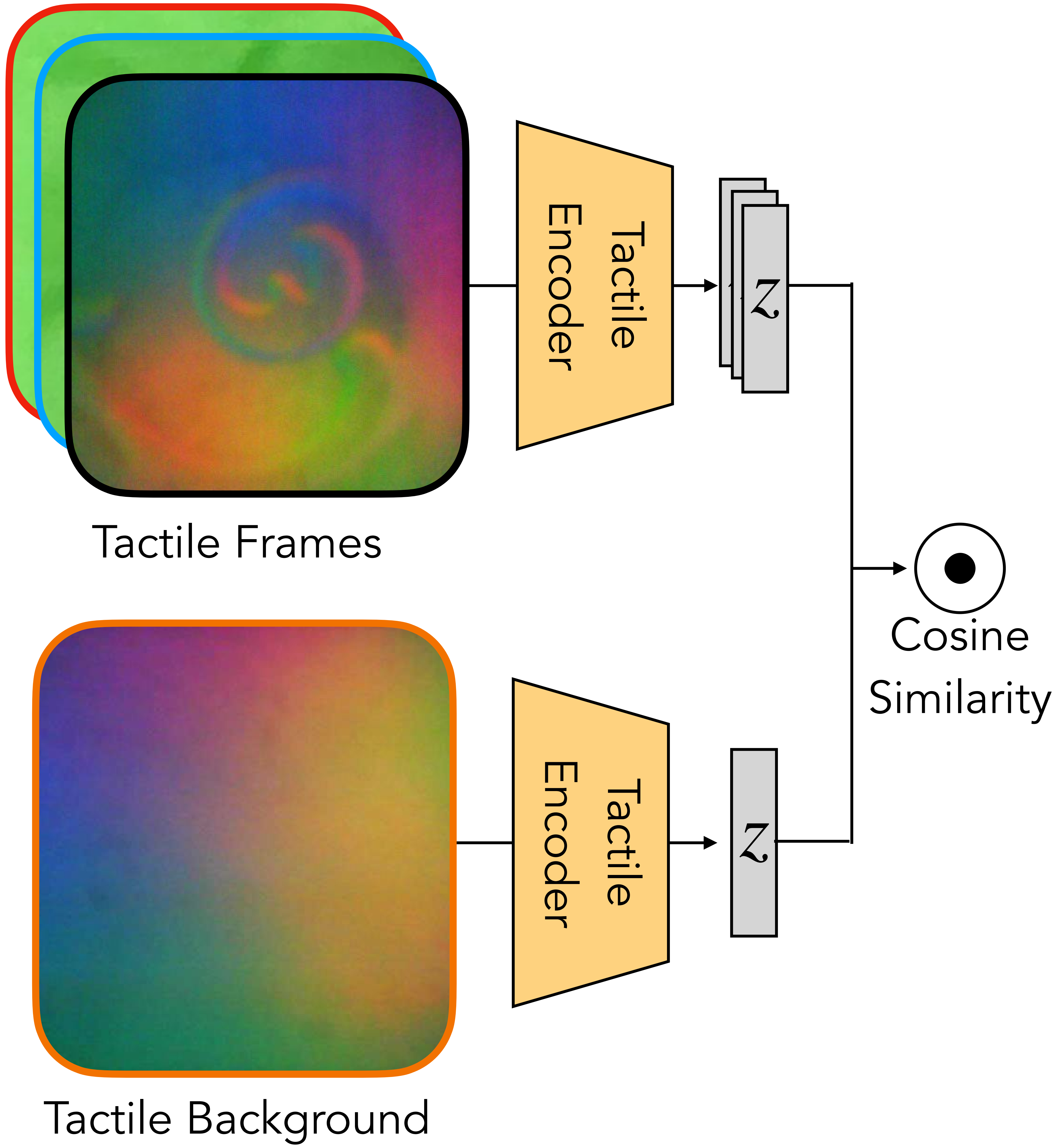
(B) Sliding

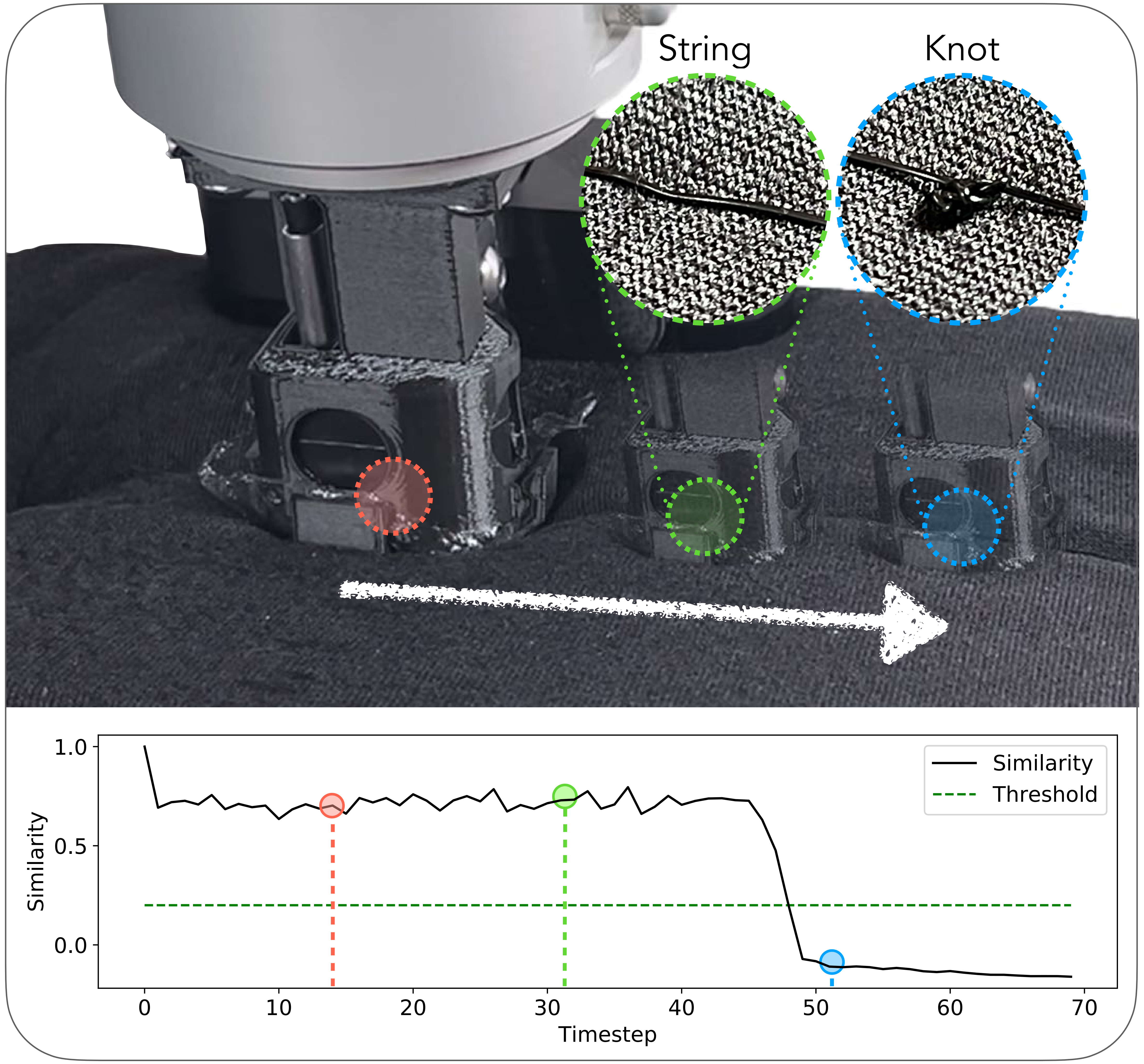
“In-the-wild” Data Collection



4.6K Human Annotations
Vision
Tactile
Bag
fabric, coarse
Vision
Tactile
Cardboard
lined, cardboard, creased
Vision
Tactile
Towel
fabric, bumpy
Vision
Tactile
Fabric
deformable, grainy
213K Human-Collected Image-Touch Pairs
Shoe
Glove
Shoe
Keyhole
Printed Part
Sandal
Outlet
Plaque
Screwdriver
Wrench
Pill Bottle
Light Switch

Data Preprocessing

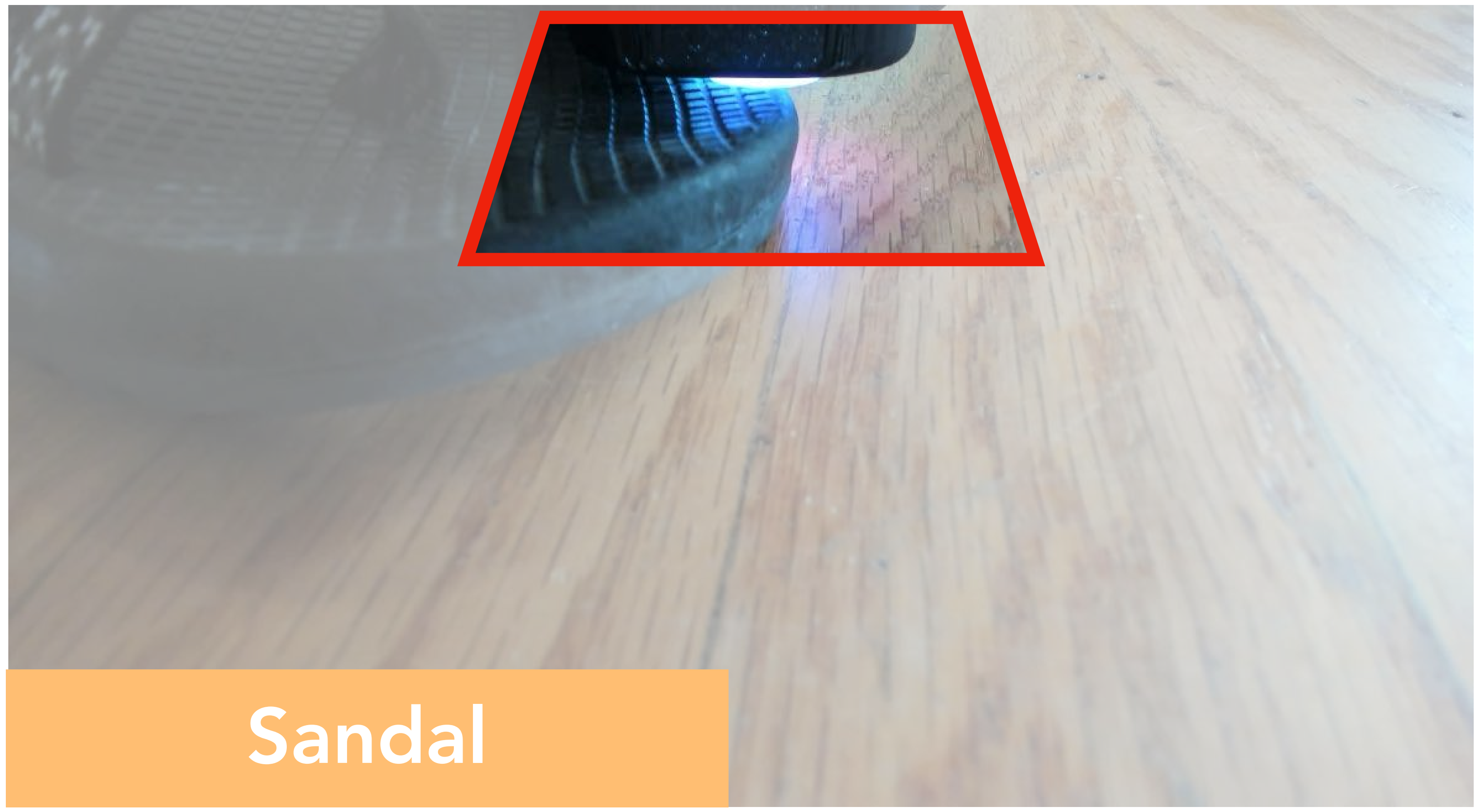




[1] Kerr, Justin et al. "Self-supervised visuo-tactile pretraining to locate and follow garment features." RSS 2023.

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."

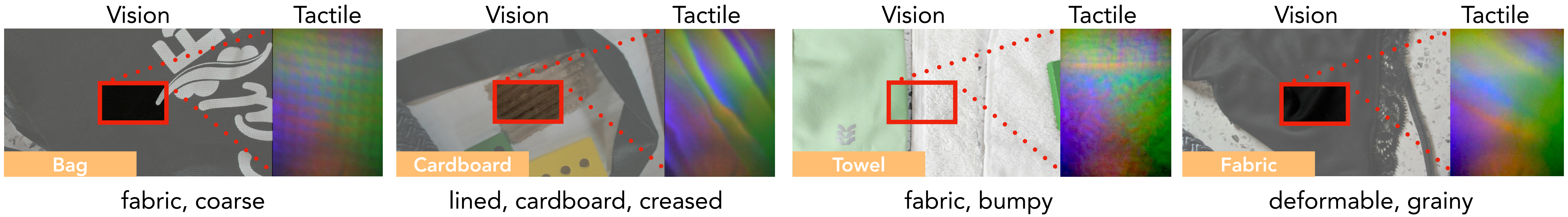
Task: Based on these images, describe the possible tactile feelings of the contact patch using sensory adjectives. Limit your response up to five adjectives, separated by commas.

GPT-4V [1] → "textured, firm, worn, cool"

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

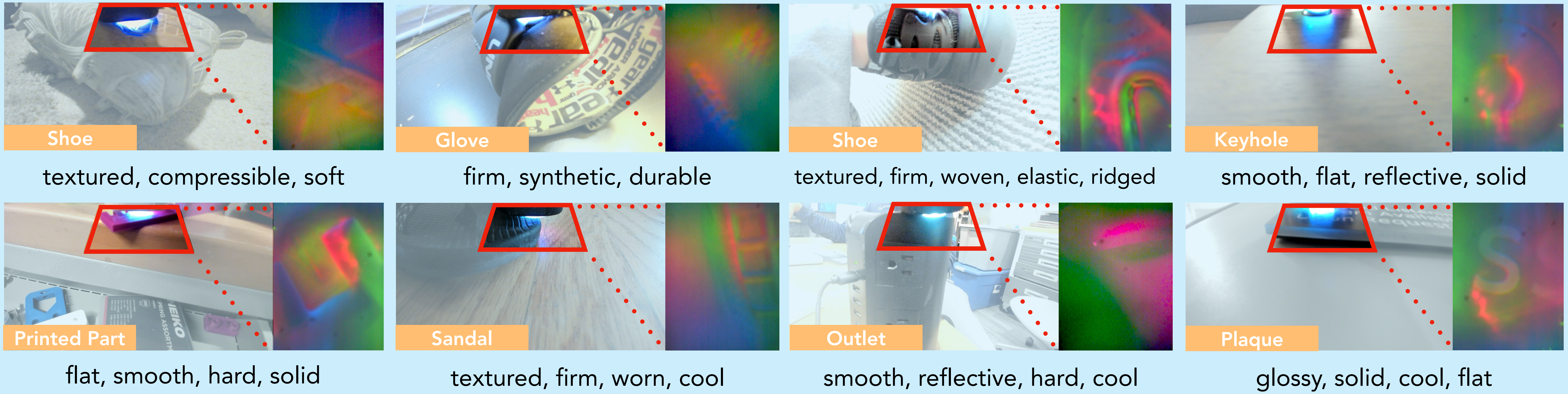
Human + VLM Pseudo-labels

4.6K Human Annotations

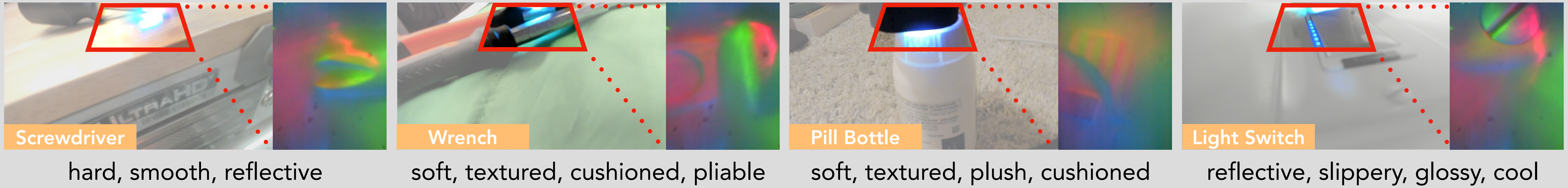


39K Pseudo-Labels with GPT-4V

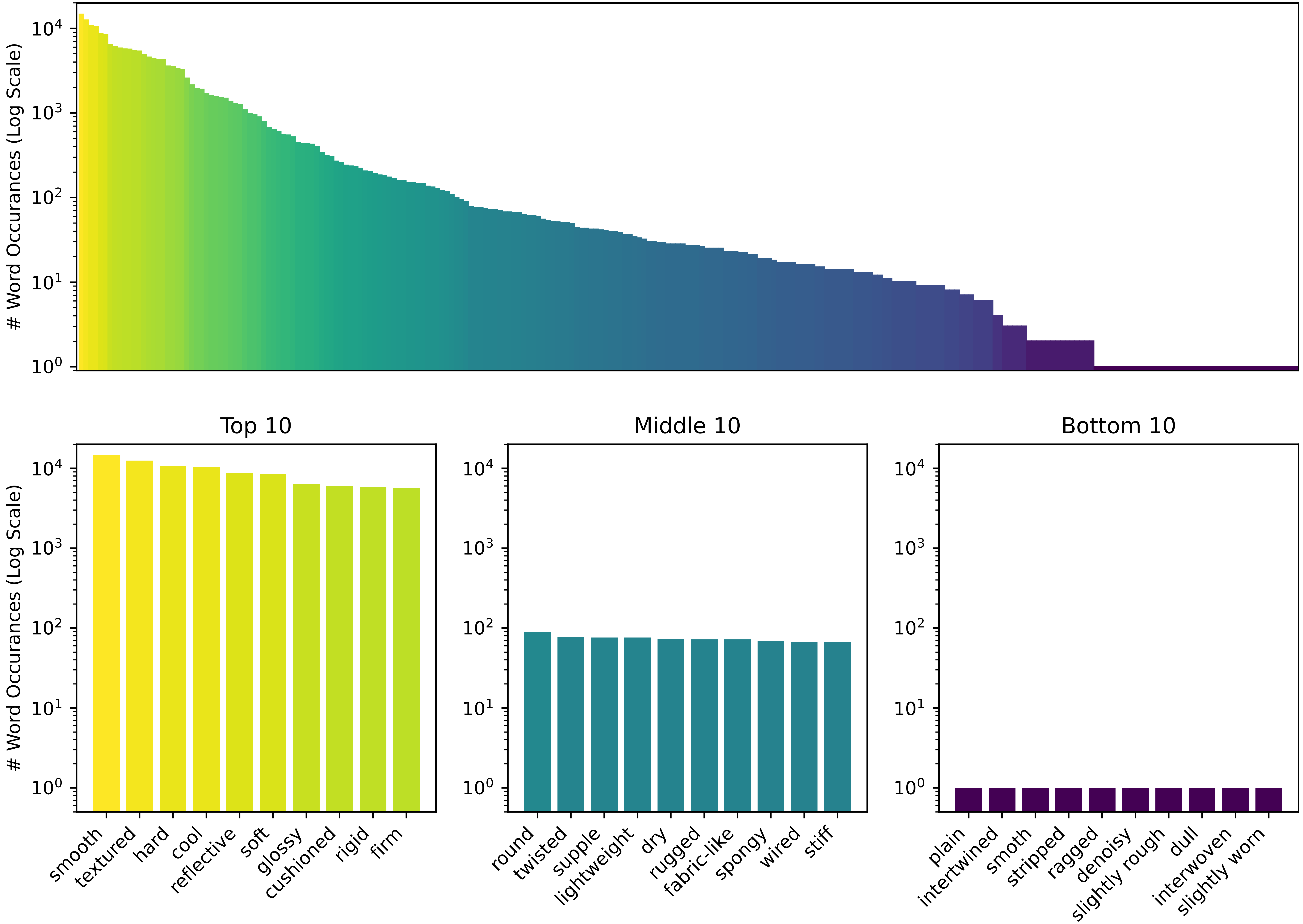
Correctly Labeled


Shoe
textured, compressible, soft
Glove
firm, synthetic, durable
Shoe
textured, firm, woven, elastic, ridged
Keyhole
smooth, flat, reflective, solid
Printed Part
flat, smooth, hard, solid
Sandal
textured, firm, worn, cool
Outlet
smooth, reflective, hard, cool
Plaque
glossy, solid, cool, flat

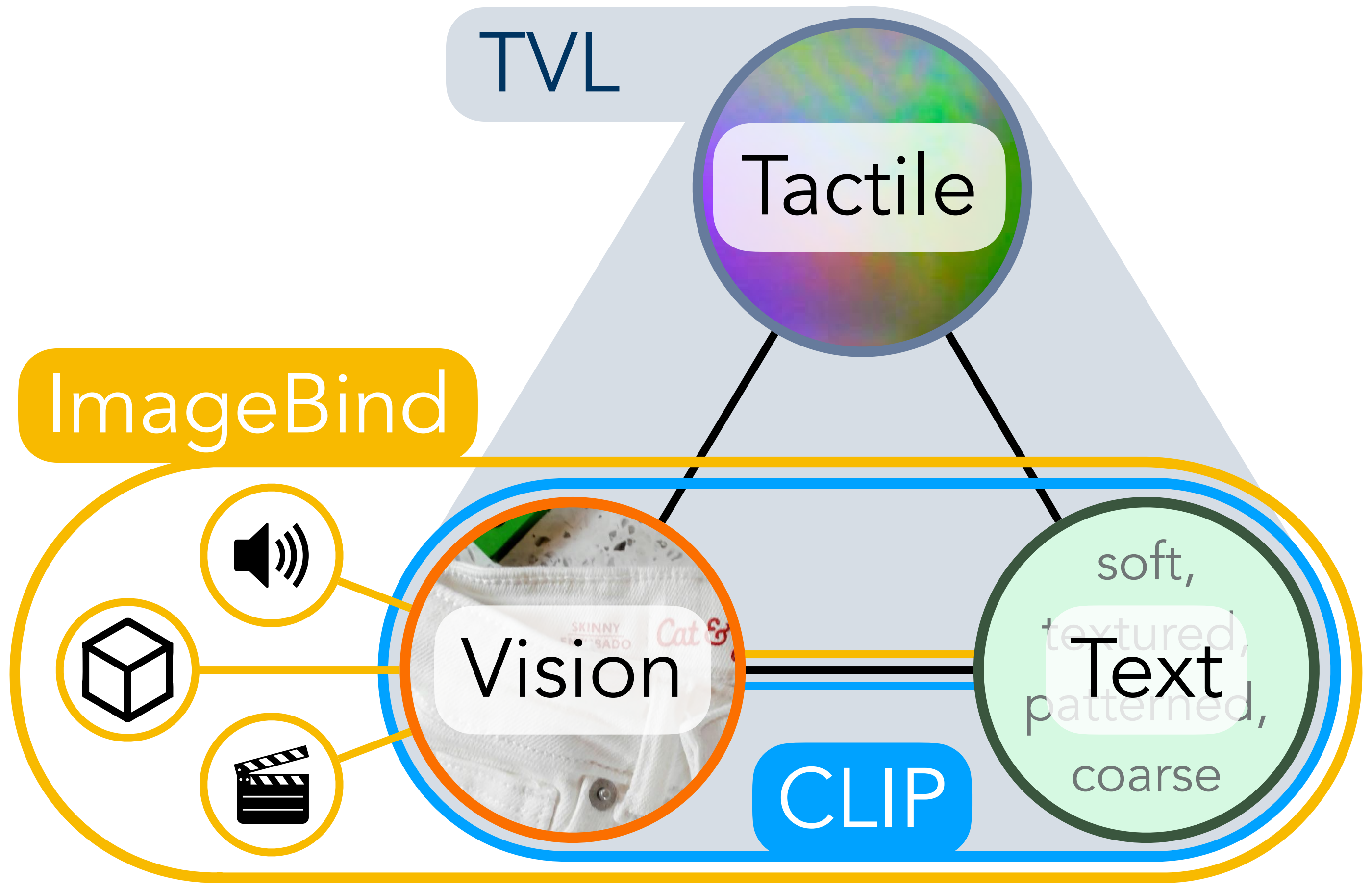
Mislabeled


Screwdriver
hard, smooth, reflective
Wrench
soft, textured, cushioned, pliable
Pill Bottle
soft, textured, plush, cushioned
Light Switch
reflective, slippery, glossy, cool

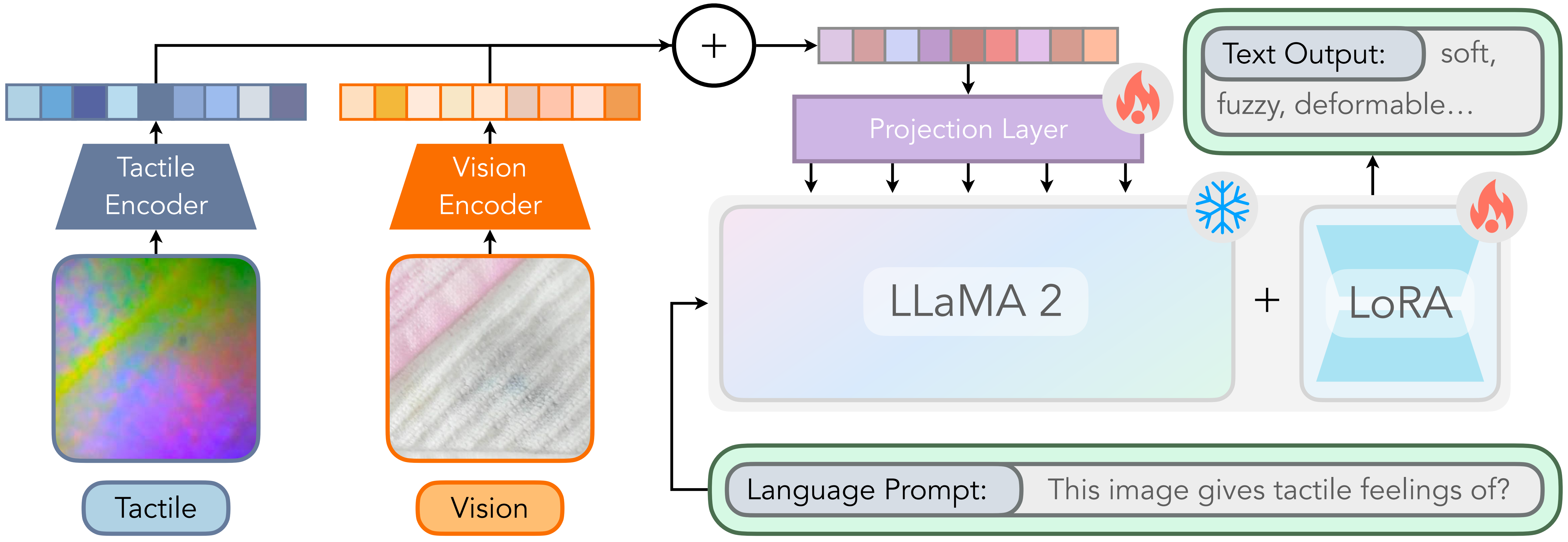
Distribution of Tactile Descriptors



TVL Models



TVL-Tactile Encoder

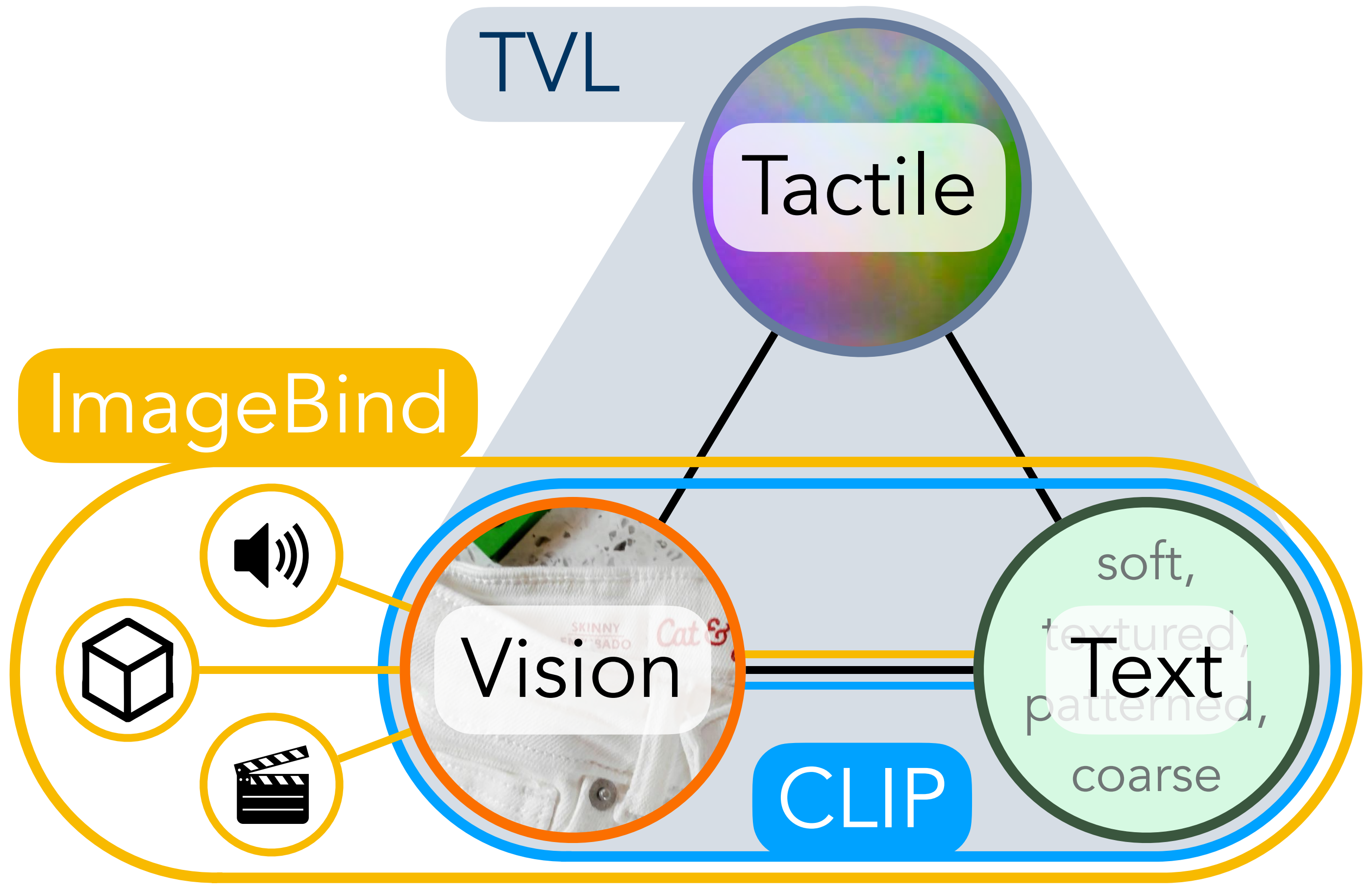


Text Output: soft, fuzzy, deformable…

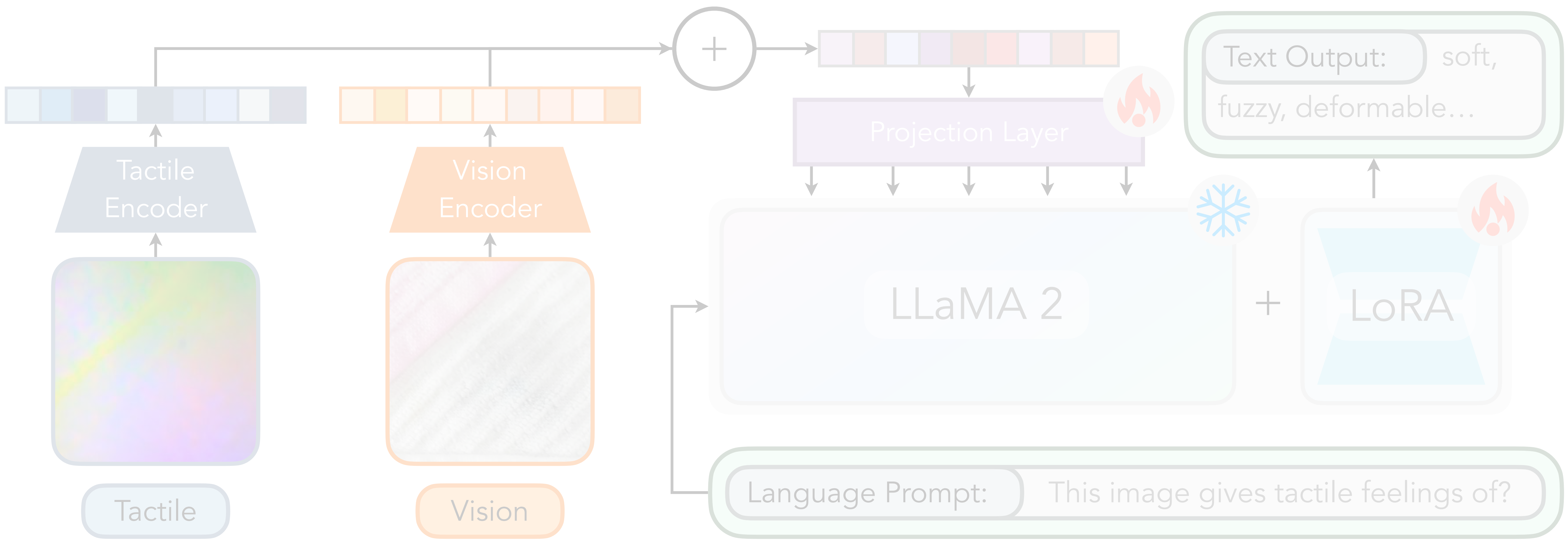
Language Prompt: This image gives tactile feelings of?

TVL-LLaMA

TVL-Tactile Encoder

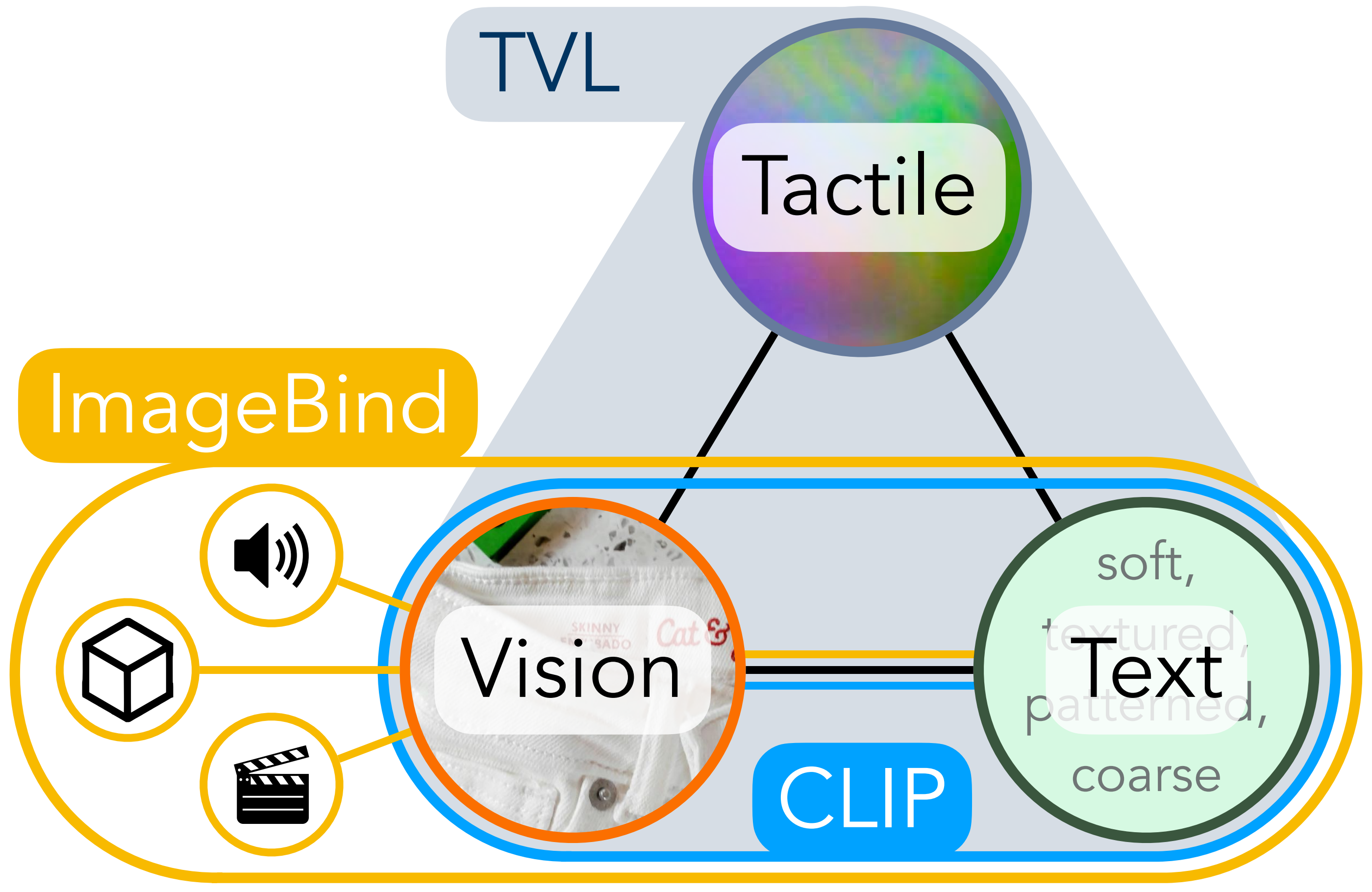


TVL-Tactile Encoder

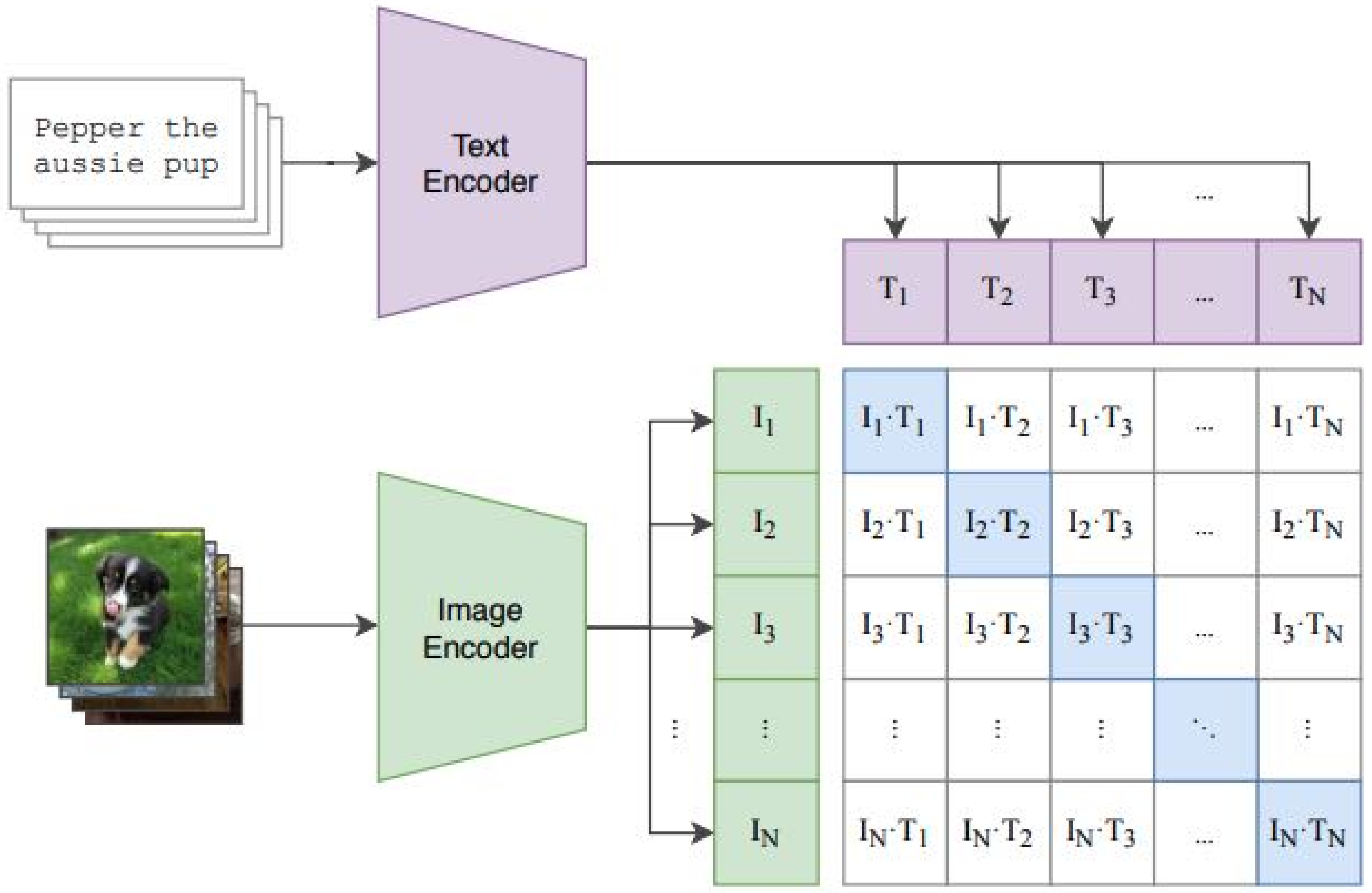

Tactile Encoder
Tactile
Vision Encoder
Vision
+
Projection Layer
LLaMA 2
+
LoRA
Text Output: soft, fuzzy, deformable…
Language Prompt: This image gives tactile feelings of?

TVL-LLaMA

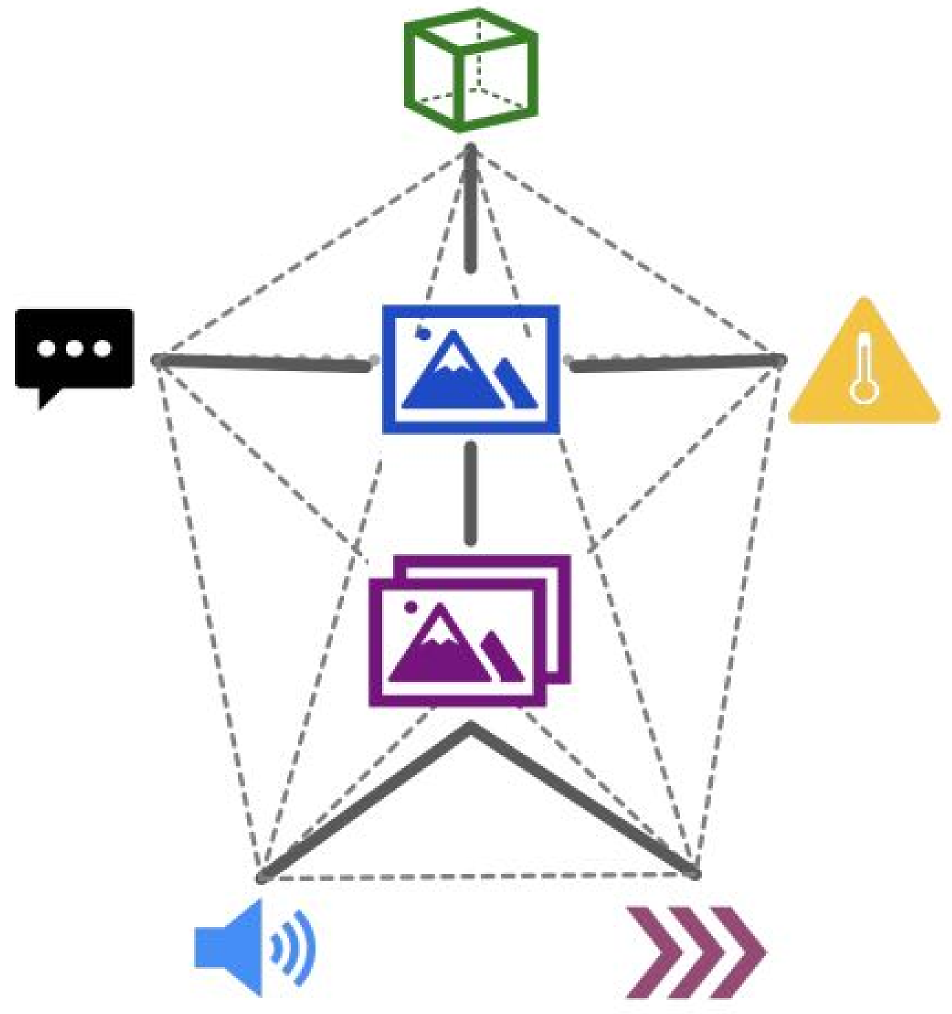
TVL-Tactile Encoder



TVL-Tactile Encoder



CLIP [1]



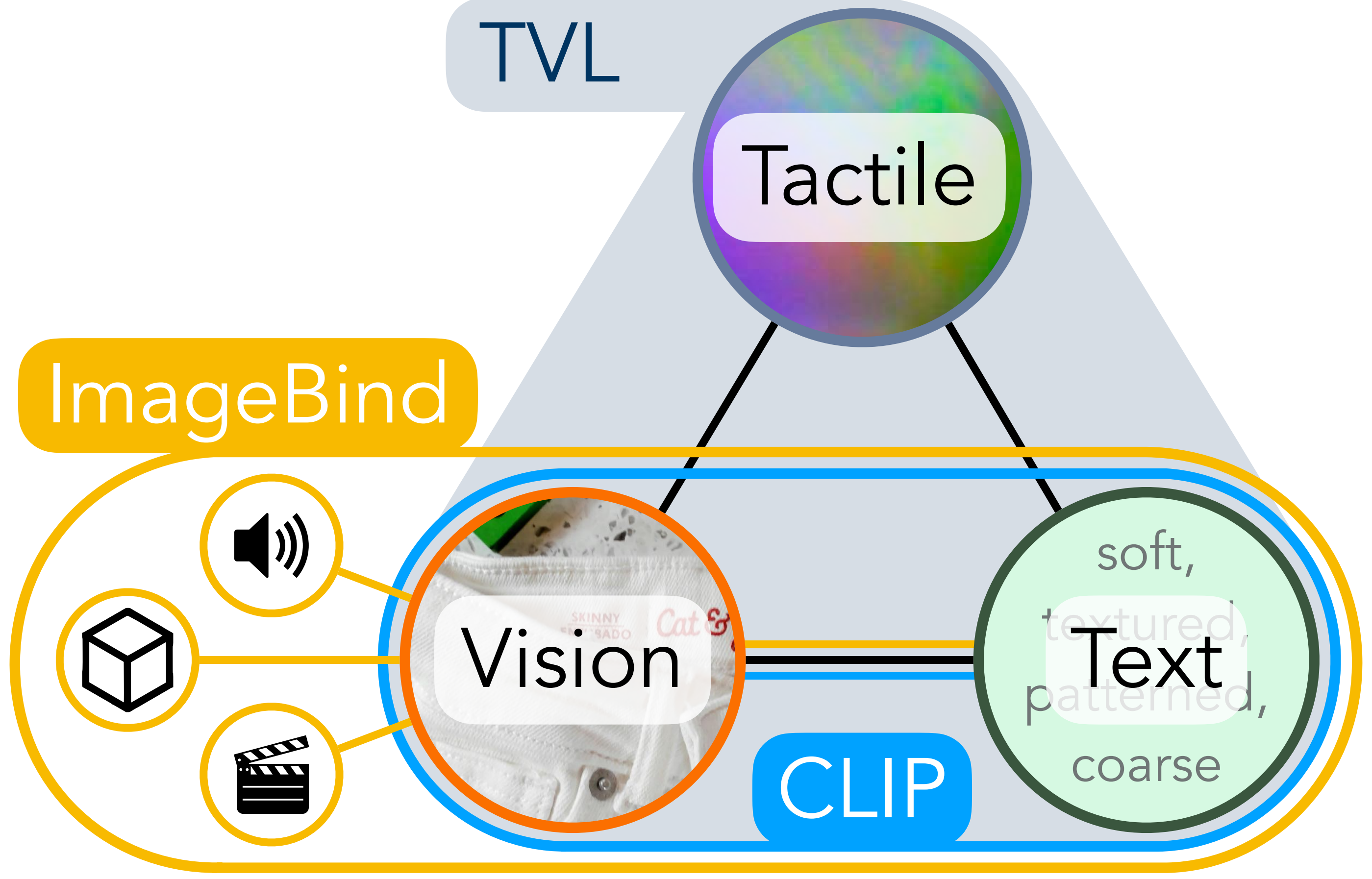
ImageBind [2]

---

[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.

TVL-Tactile Encoder



TVL

Tactile

ImageBind

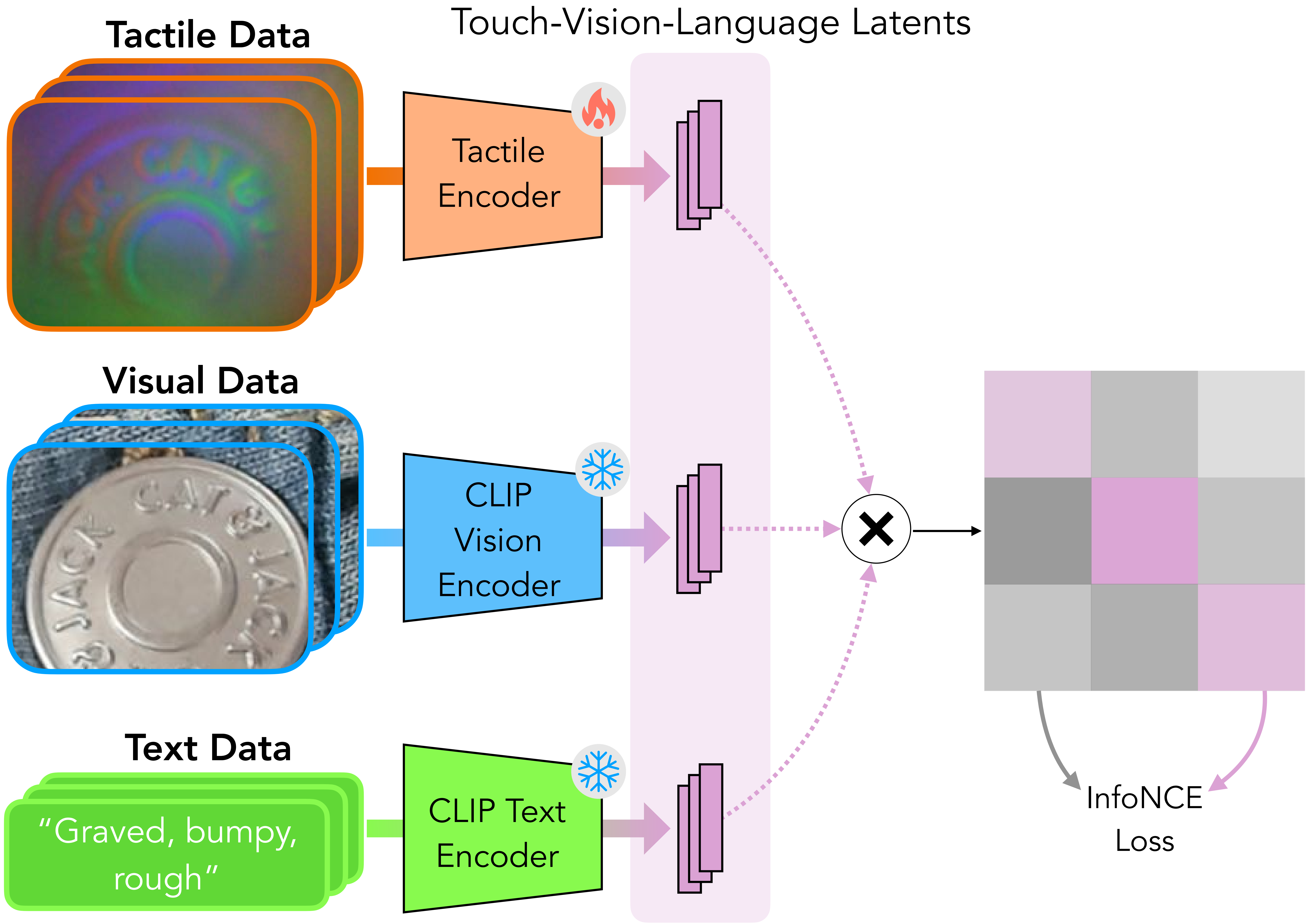
Vision

Text

soft, textured, patterned, coarse

CLIP

TVL-Tactile Encoder



Tactile Data

Touch-Vision-Language Latents

Tactile Encoder

Visual Data

CLIP Vision Encoder

Text Data

"Graved, bumpy, rough"

CLIP Text Encoder

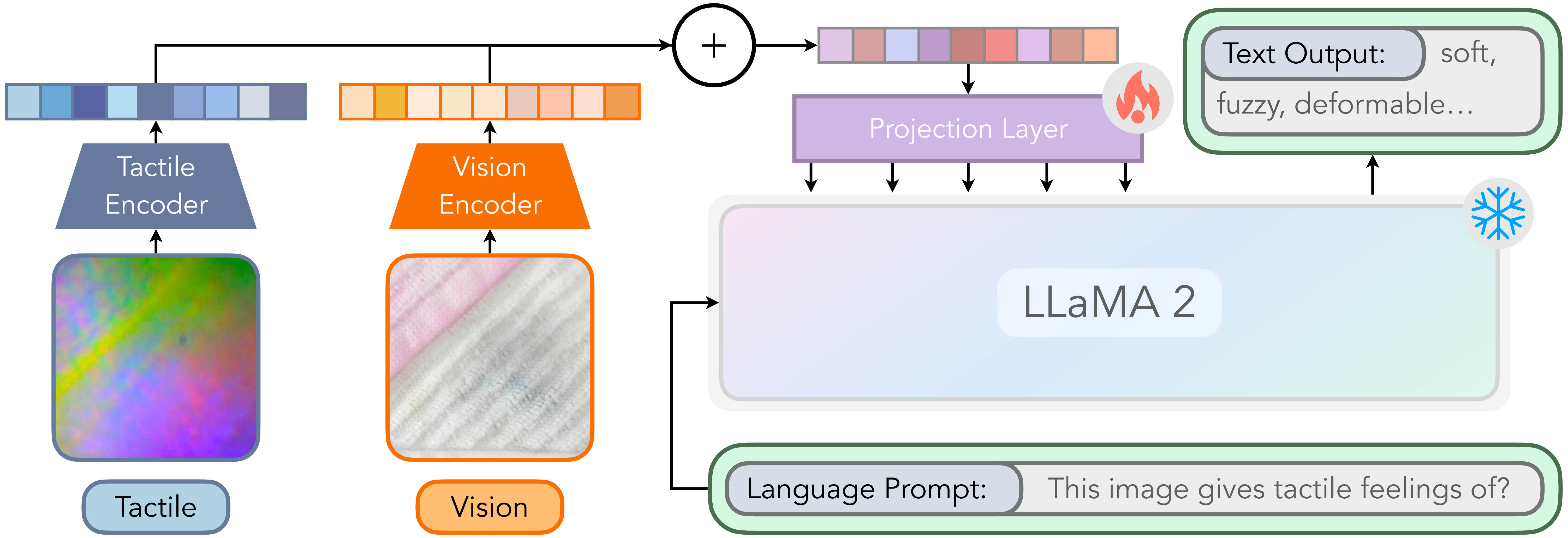
InfoNCE Loss

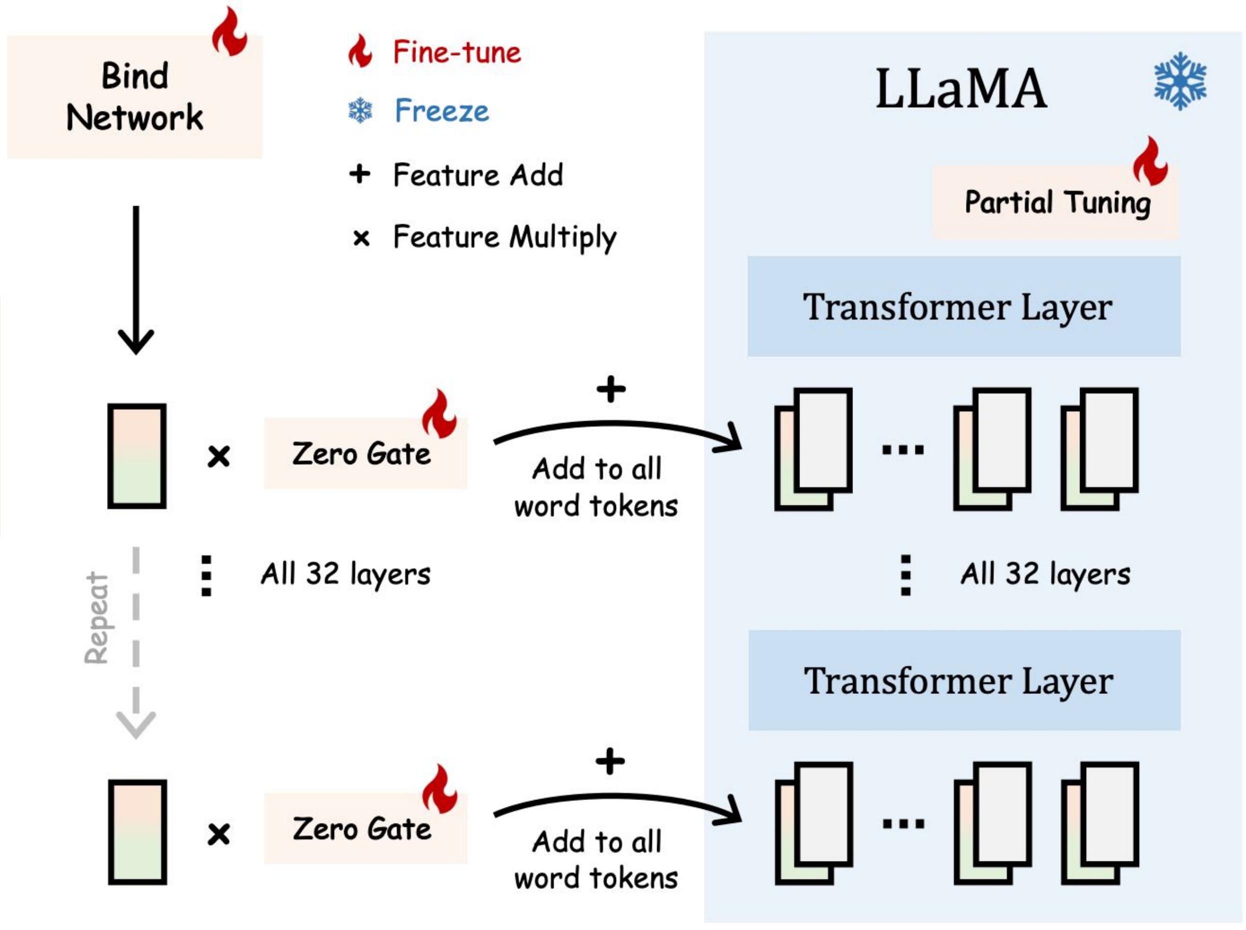
TVL-LLaMA

Text Output: soft, fuzzy, deformable…
Projection Layer
Tactile Encoder
Vision Encoder
LLaMA 2
+
LoRA
Tactile
Vision
Language Prompt: This image gives tactile feelings of?

[1] Han, Jiaming et al. "ImageBind-LLM: Multi-modality Instruction Tuning." arXiv 2023.

TVL-LLaMA


Tactile Encoder
Vision Encoder
+
Projection Layer
Text Output: soft, fuzzy, deformable…
LLaMA 2
Language Prompt: This image gives tactile feelings of?
Tactile
Vision



ImageBind-LLM [1]

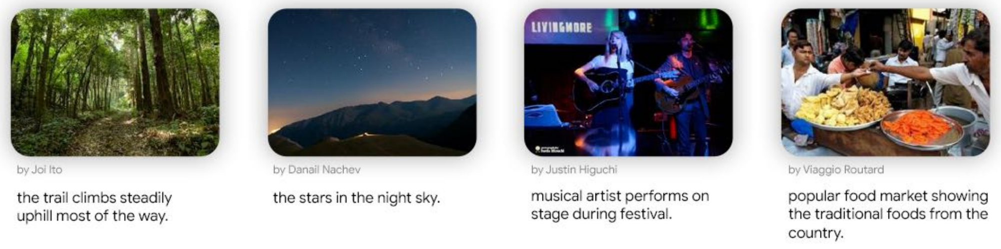
Learnable Gate

$$T^j = T_I \cdot g_{zero} + T_W^j$$

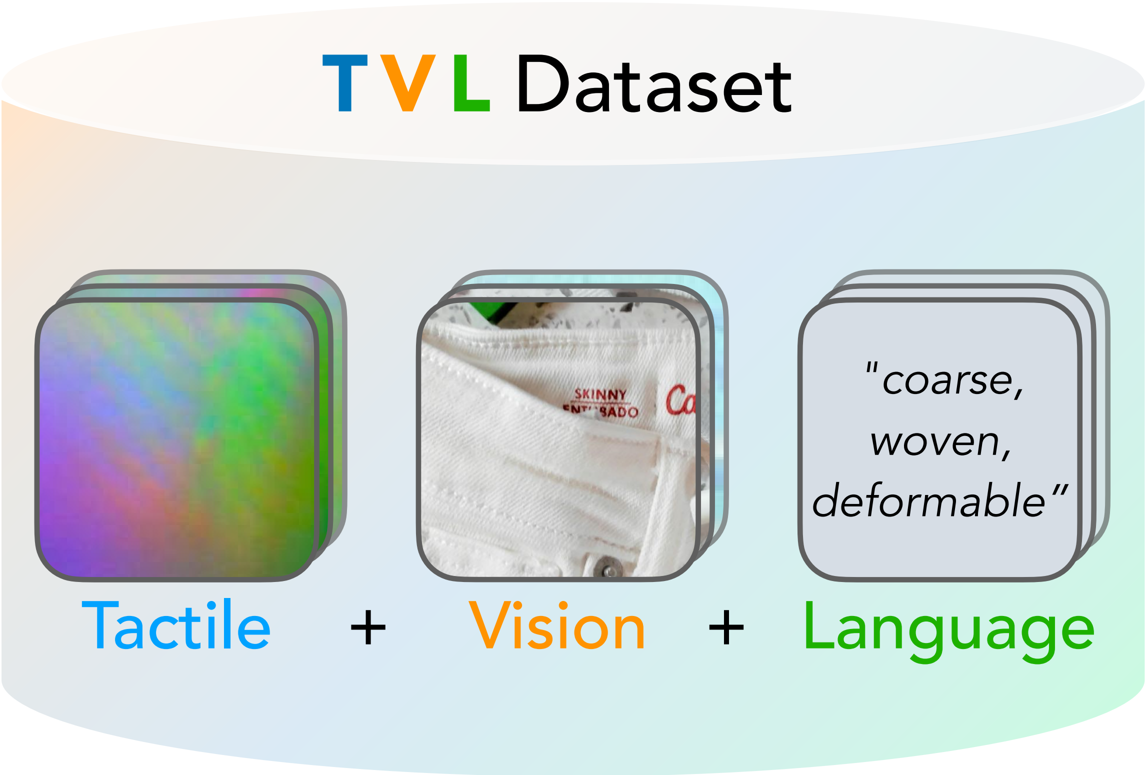
Tactile and Image Features

Word Token

Pretraining


the trail climbs steadily uphill most of the way.
the stars in the night sky.
musical artist performs on stage during festival.
popular food market showing the traditional foods from the country.

CC3M [2]

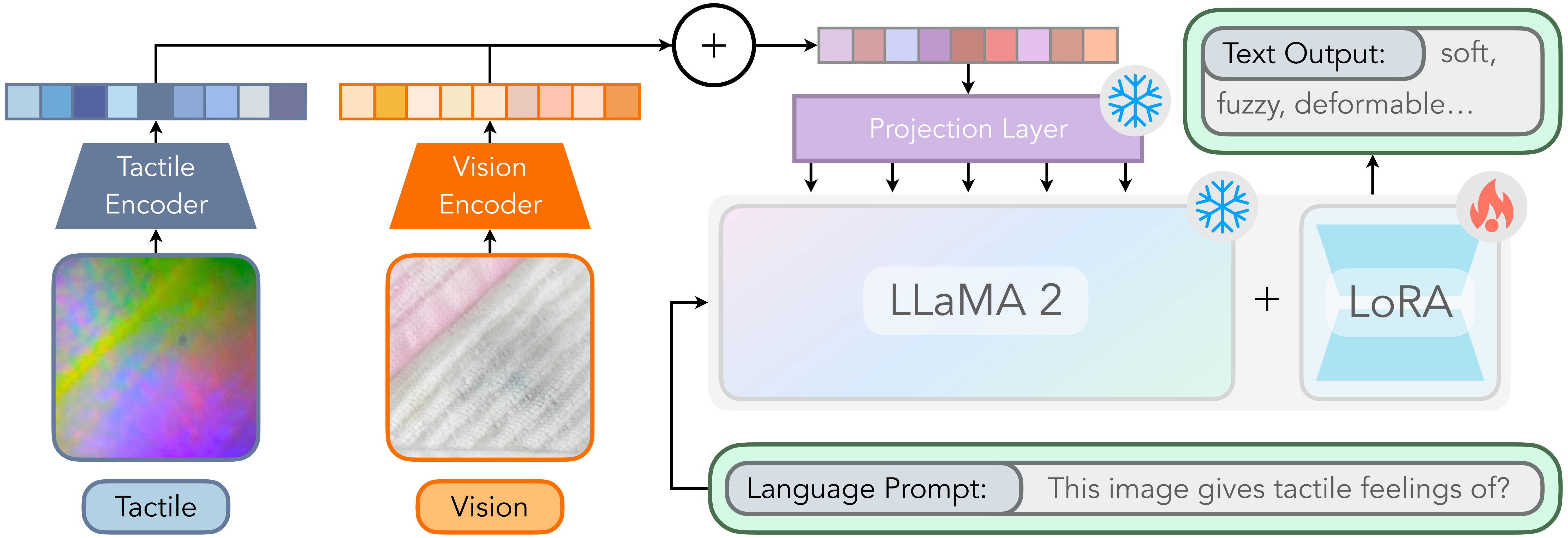


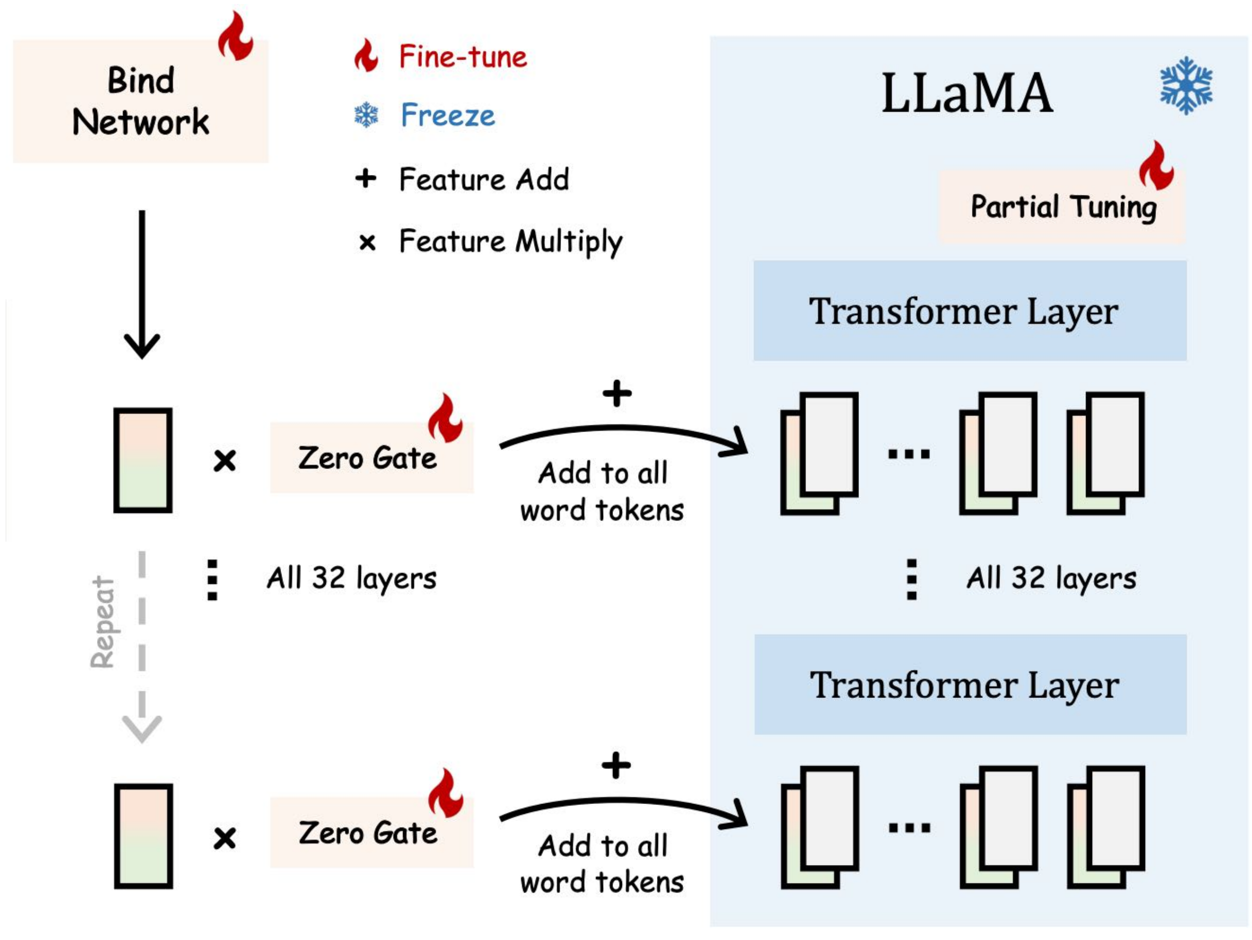
---

[1] Han, Jiaming et al. "ImageBind-LLM: Multi-modality Instruction Tuning." arXiv 2023.

[2] Sharma, Piyush et al. "Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning." ACL 2018.

TVL-LLaMA


Tactile Encoder
Vision Encoder
Tactile
Vision
Projection Layer
LLaMA 2
+
LoRA
Text Output: soft, fuzzy, deformable…
Language Prompt: This image gives tactile feelings of?



ImageBind-LLM [1]

Learnable Gate

$$T^j = T_I \cdot g_{zero} + T^j_W$$

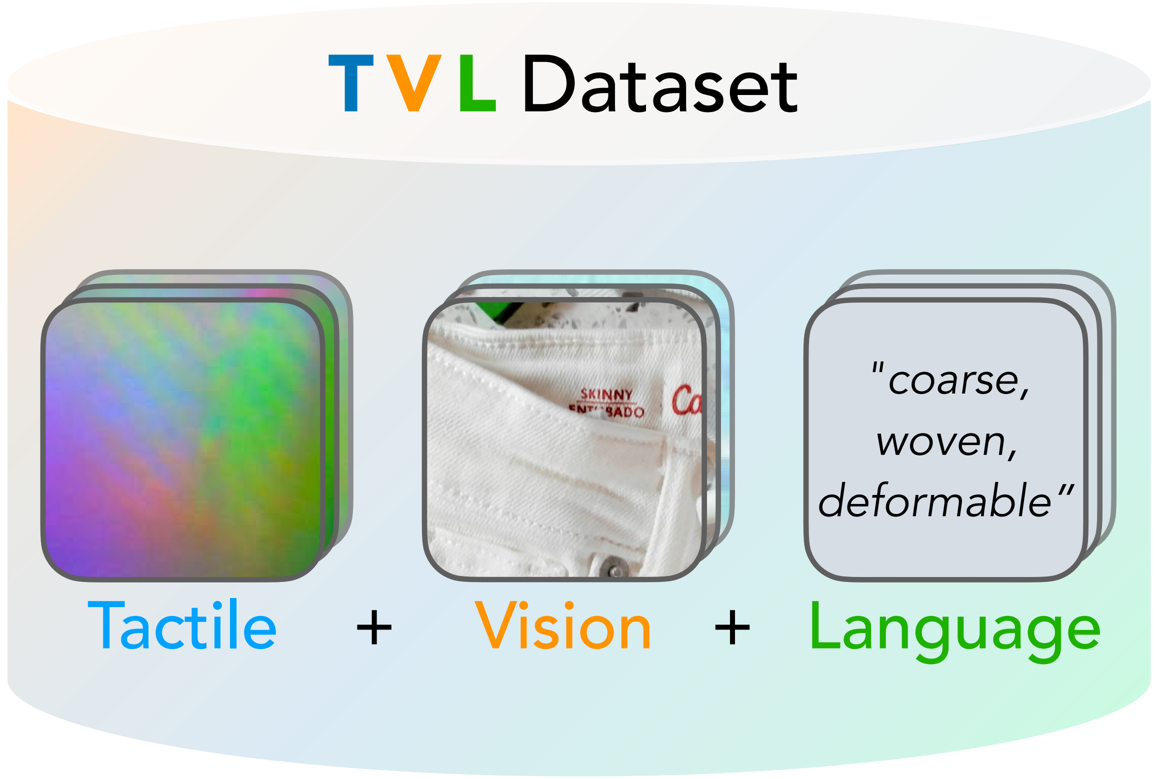
Tactile and Image Features

Word Token

Finetuning

Alpaca [1]

LLaVA-Instruct-150K [3]



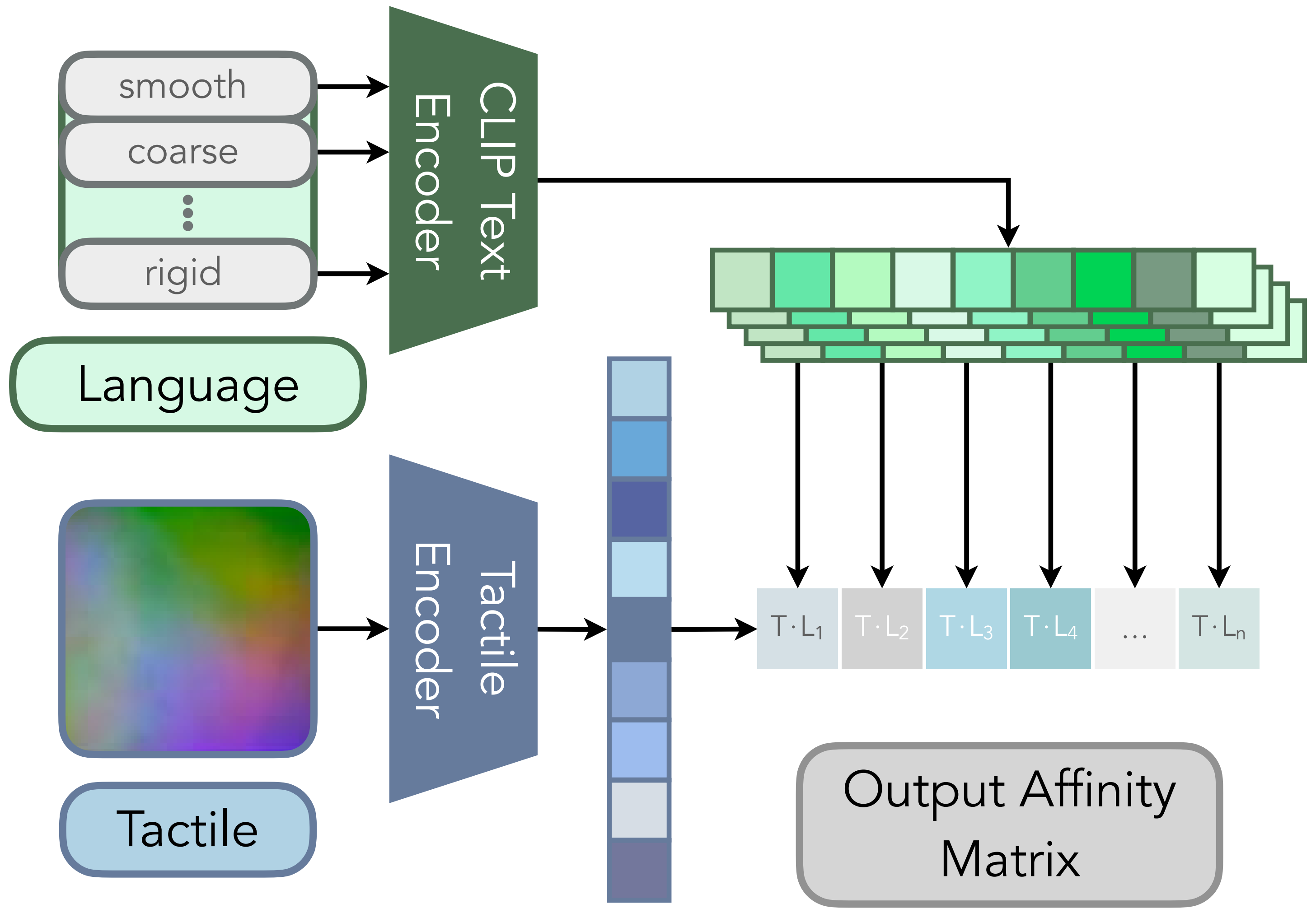
[1] Han, Jiaming et al. "ImageBind-LLM: Multi-modality Instruction Tuning." arXiv 2023.
[2] Taori, Rohan et al. "Stanford Alpaca: An Instruction-following LLaMA model." GitHub 2023.
[3] Liu, Haotian et al. "Visual Instruction Tuning." NeurIPS 2023.

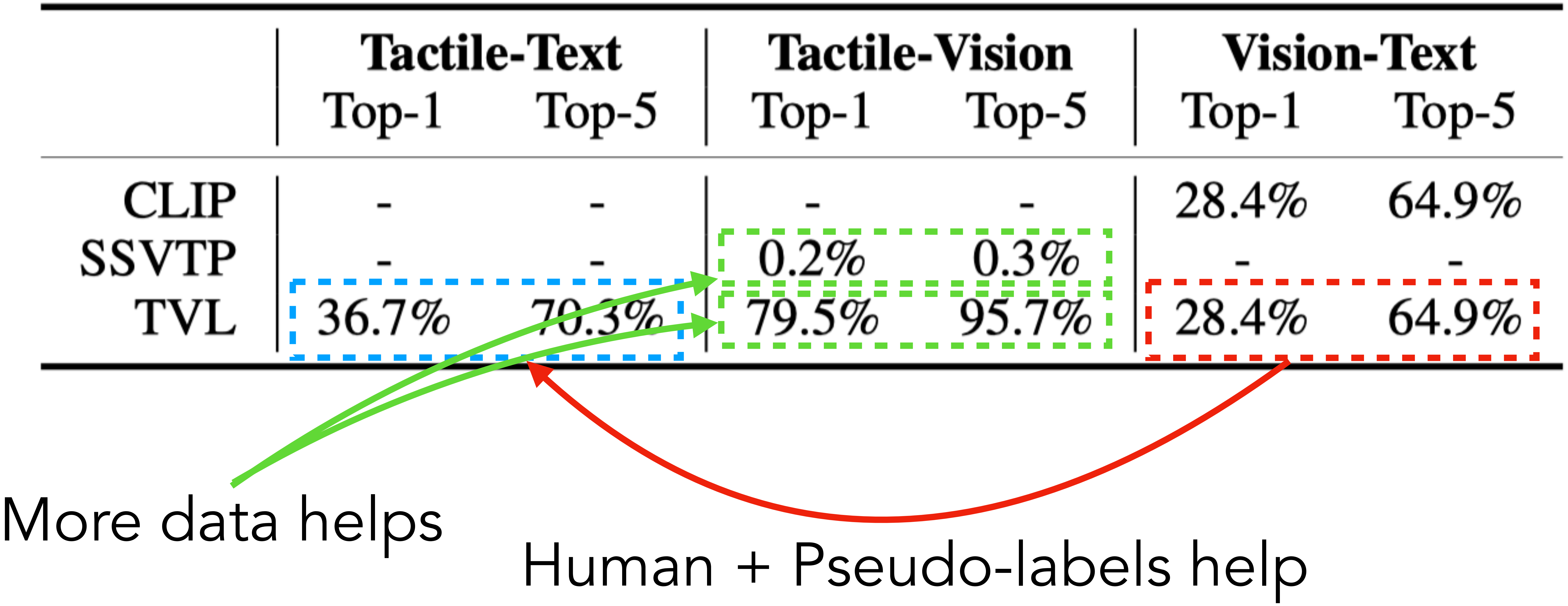
TVL-Benchmark

smooth
coarse
rigid
Language
CLIP Text Encoder
Tactile
Tactile Encoder
$T \cdot L_1$
$T \cdot L_2$
$T \cdot L_3$
$T \cdot L_4$
...
$T \cdot L_n$
Output Affinity Matrix
+
+
+
Language Model
Tactile
Vision
Language

Human Labels:	"bumpy, plush, soft, cushioned"	Score: GT
GPT-4V:	"textured, soft, plush, fibrous, cushioned"	Score: 9/10
TVL-LLaMA:	"soft, plush, textured, cushioned, fibrous"	Score: 7.5/10
Human Labels:	"flat, lined, hard"	Score: GT
GPT-4V:	"Textured, flexible, soft, rubbery, woven"	Score: 2.5/10
TVL-LLaMA:	"flat, hard."	Score: 7/10
Human Labels:	"fibrous, textured, uneven, pliable"	Score: GT
GPT-4V:	"smooth, reflective, hard, cool, sleek"	Score: 1/10
TVL-LLaMA:	"smooth, glossy, hard, cool, sleek."	Score: 1/10

TVL-Benchmark



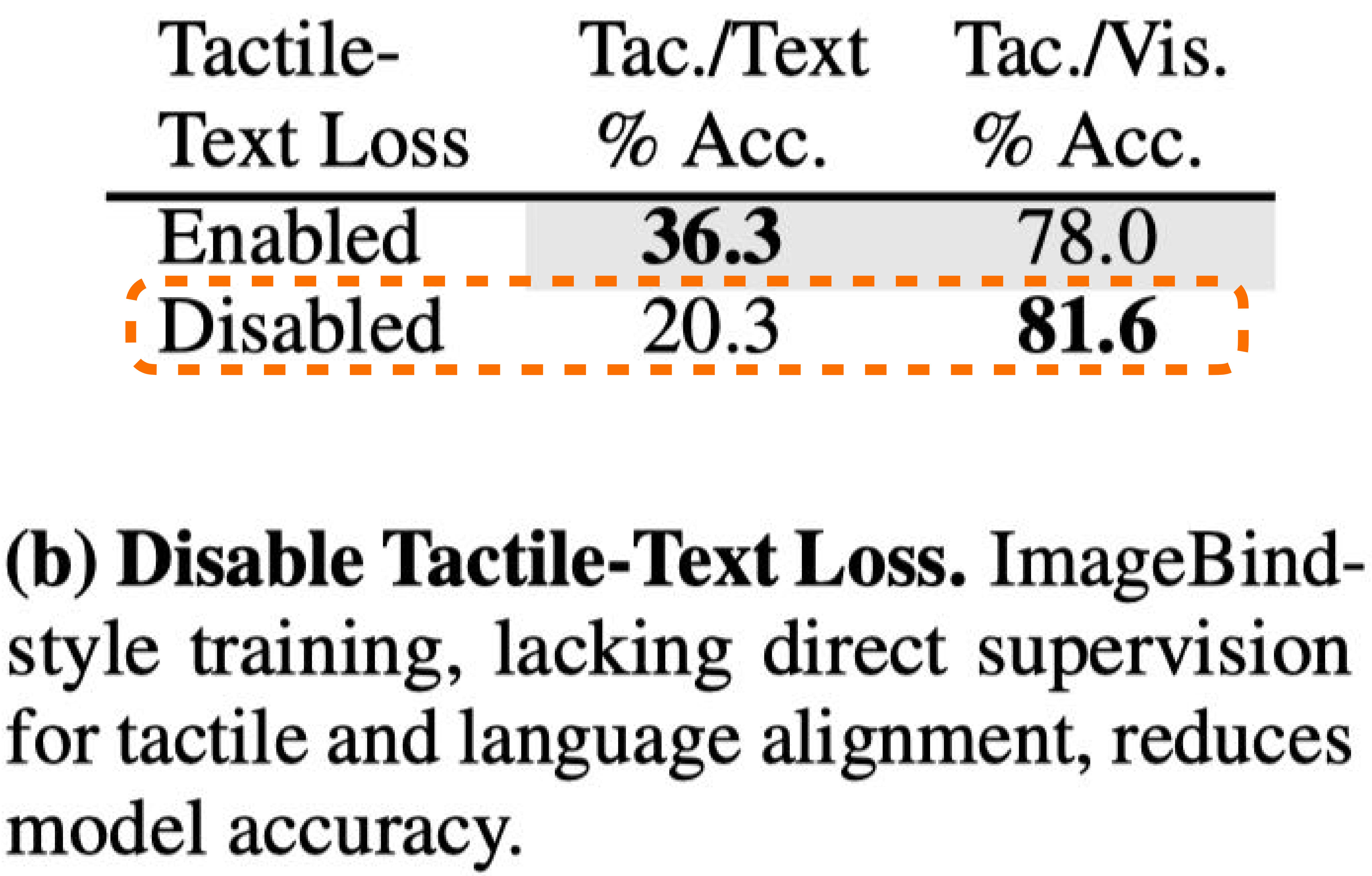


	Tactile-Text		Tactile-Vision		Vision-Text	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
CLIP	-	-	-	-	28.4%	64.9%
SSVTP	-	-	0.2%	0.3%	-	-
TVL	36.7%	70.3%	79.5%	95.7%	28.4%	64.9%

More data helps

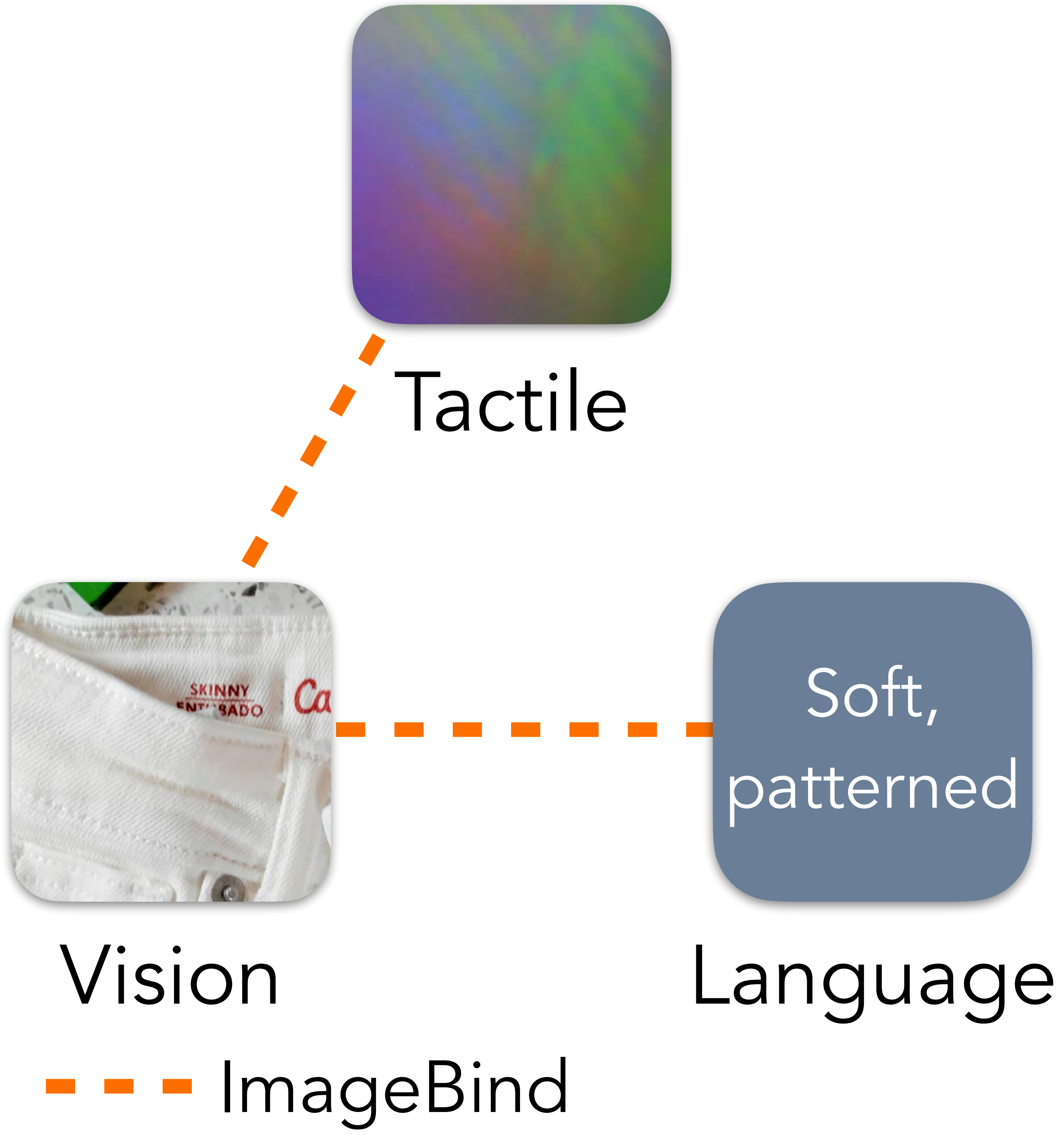
Human + Pseudo-labels help

TVL-Benchmark

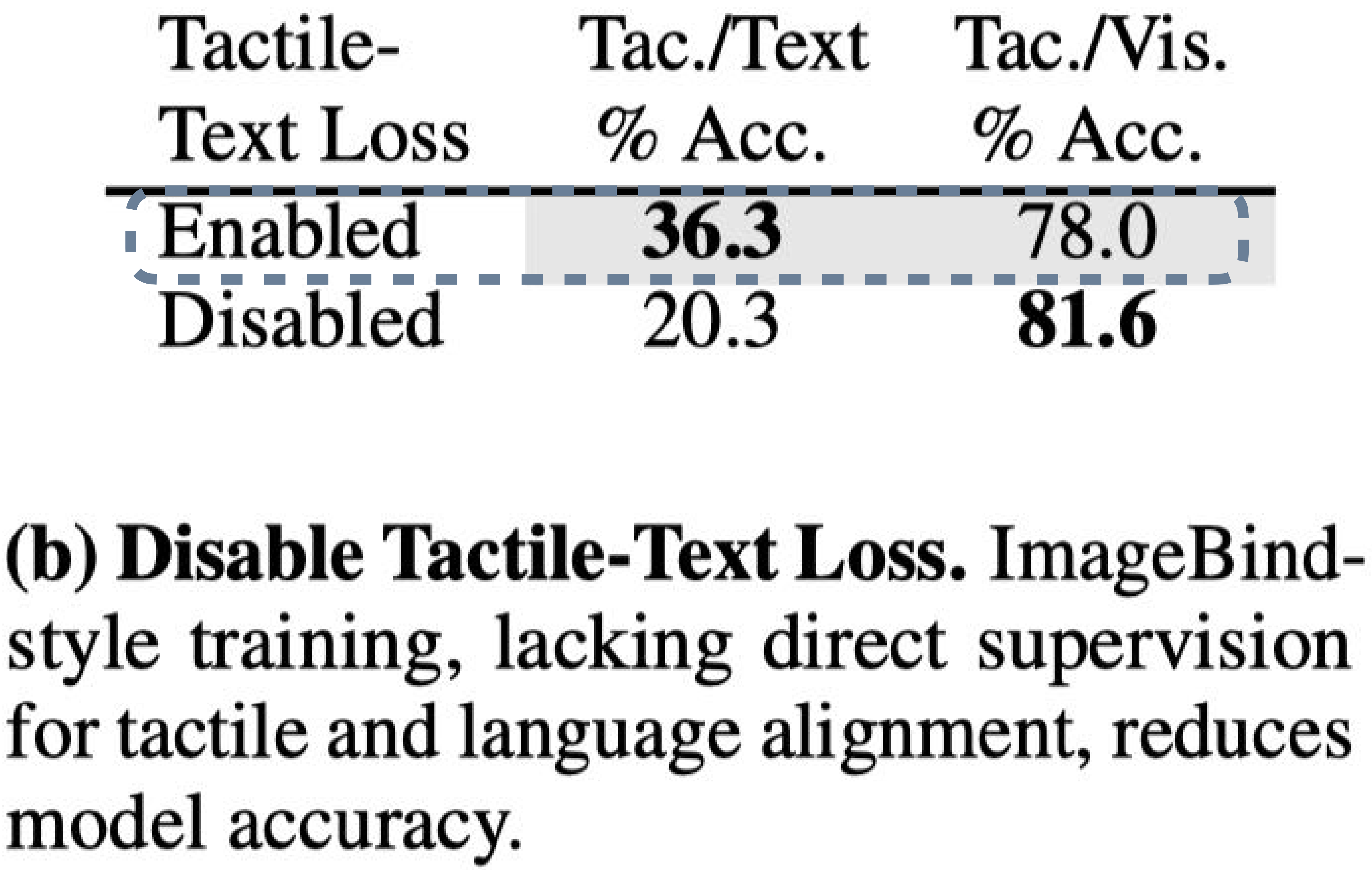


Tactile-Text Loss	Tac./Text % Acc.	Tac./Vis. % Acc.
Enabled	**36.3**	78.0
Disabled	20.3	**81.6**

(b) Disable Tactile-Text Loss. ImageBind-style training, lacking direct supervision for tactile and language alignment, reduces model accuracy.

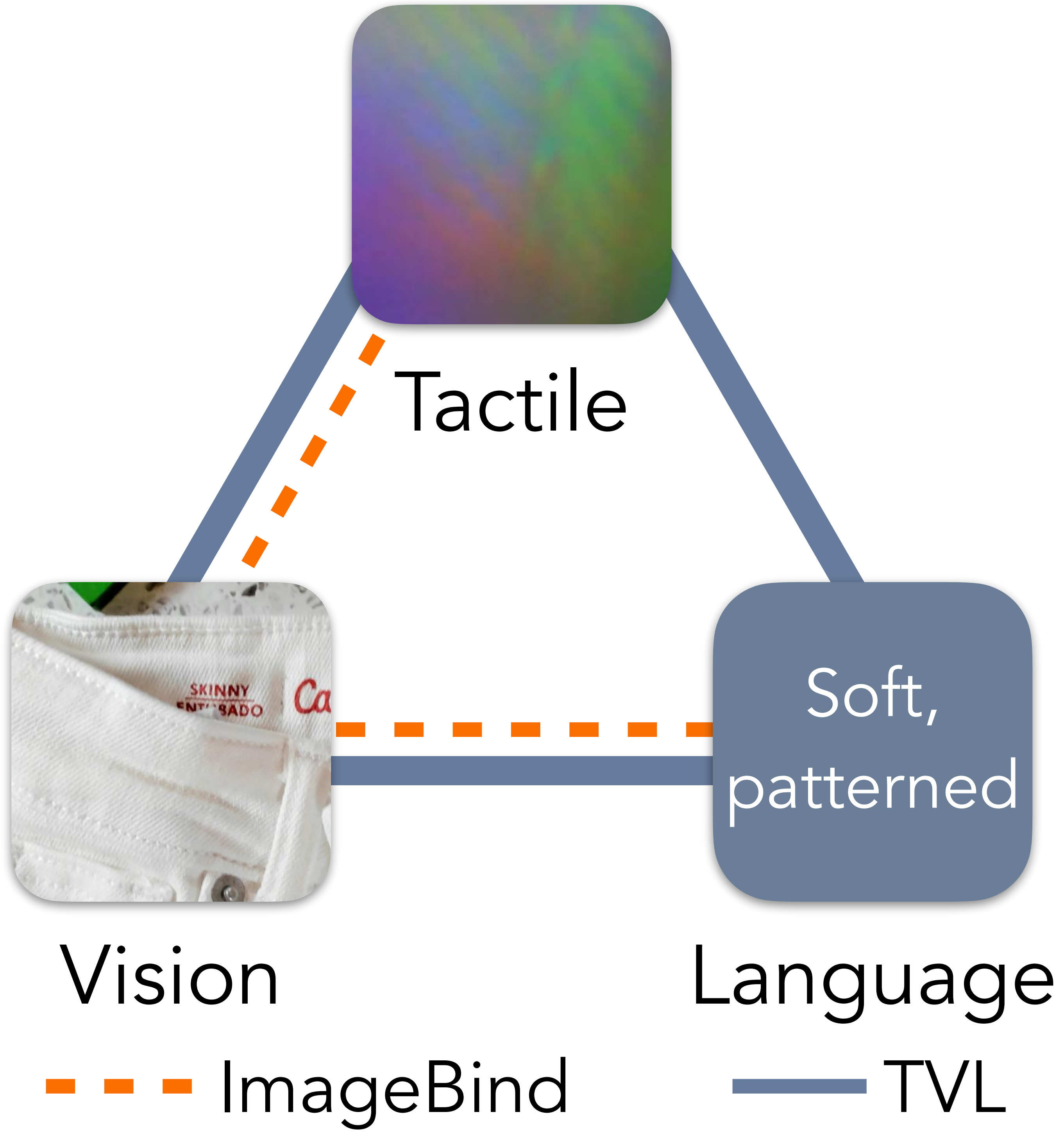


TVL-Benchmark

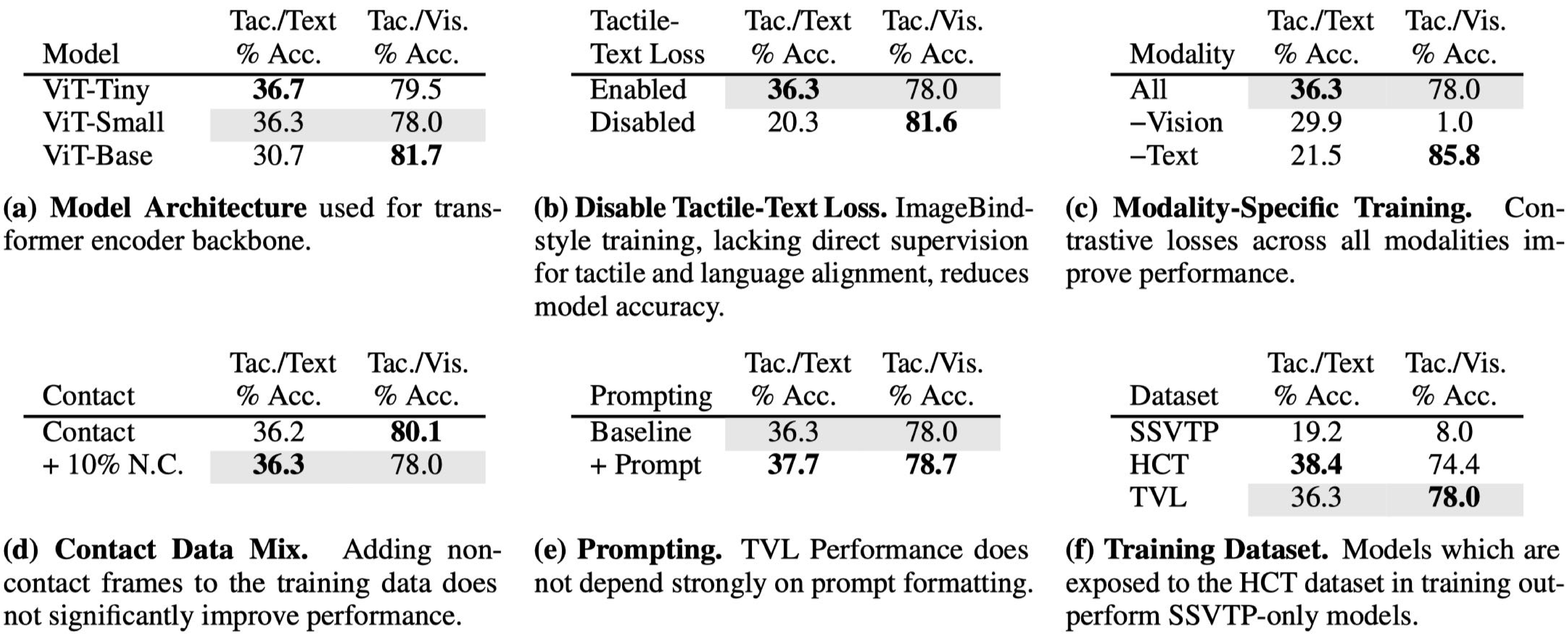


Tactile-Text Loss	Tac./Text % Acc.	Tac./Vis. % Acc.
Enabled	**36.3**	78.0
Disabled	20.3	**81.6**

(b) Disable Tactile-Text Loss. ImageBind-style training, lacking direct supervision for tactile and language alignment, reduces model accuracy.



TVL-Benchmark



Model	Tac./Text % Acc.	Tac./Vis. % Acc.
ViT-Tiny	**36.7**	79.5
ViT-Small	36.3	78.0
ViT-Base	30.7	**81.7**

(a) Model Architecture used for transformer encoder backbone.

Tactile-Text Loss	Tac./Text % Acc.	Tac./Vis. % Acc.
Enabled	**36.3**	78.0
Disabled	20.3	**81.6**

(b) Disable Tactile-Text Loss. ImageBind-style training, lacking direct supervision for tactile and language alignment, reduces model accuracy.

Modality	Tac./Text % Acc.	Tac./Vis. % Acc.
All	**36.3**	78.0
−Vision	29.9	1.0
−Text	21.5	**85.8**

(c) Modality-Specific Training. Contrastive losses across all modalities improve performance.

Contact	Tac./Text % Acc.	Tac./Vis. % Acc.
Contact	36.2	**80.1**
+ 10% N.C.	**36.3**	78.0

(d) Contact Data Mix. Adding non-contact frames to the training data does not significantly improve performance.

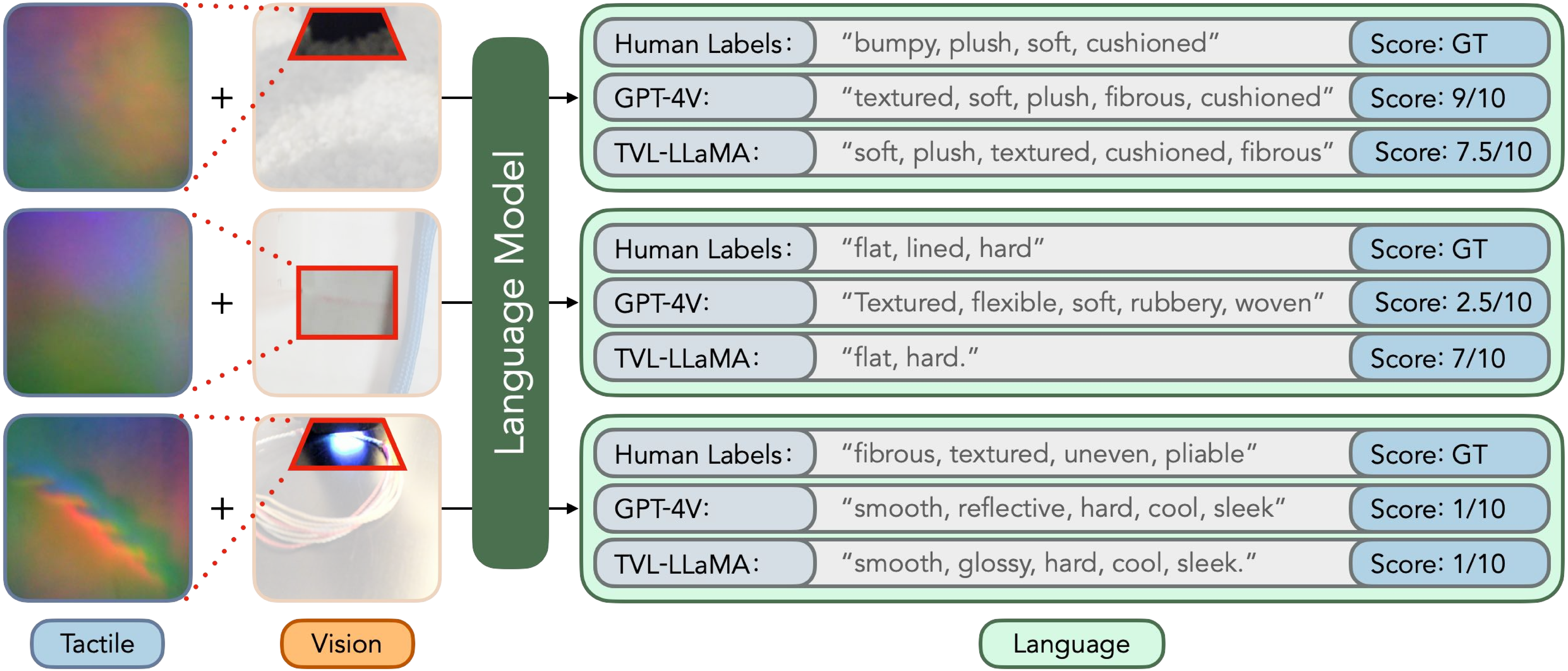
Prompting	Tac./Text % Acc.	Tac./Vis. % Acc.
Baseline	36.3	78.0
+ Prompt	**37.7**	**78.7**

(e) Prompting. TVL Performance does not depend strongly on prompt formatting.

Dataset	Tac./Text % Acc.	Tac./Vis. % Acc.
SSVTP	19.2	8.0
HCT	**38.4**	74.4
TVL	36.3	**78.0**

(f) Training Dataset. Models which are exposed to the HCT dataset in training outperform SSVTP-only models.

TVL-Benchmark



Human Labels: "bumpy, plush, soft, cushioned" Score: GT
GPT-4V: "textured, soft, plush, fibrous, cushioned" Score: 9/10
TVL-LLaMA: "soft, plush, textured, cushioned, fibrous" Score: 7.5/10

Human Labels: "flat, lined, hard" Score: GT
GPT-4V: "Textured, flexible, soft, rubbery, woven" Score: 2.5/10
TVL-LLaMA: "flat, hard." Score: 7/10

Human Labels: "fibrous, textured, uneven, pliable" Score: GT
GPT-4V: "smooth, reflective, hard, cool, sleek" Score: 1/10
TVL-LLaMA: "smooth, glossy, hard, cool, sleek." Score: 1/10

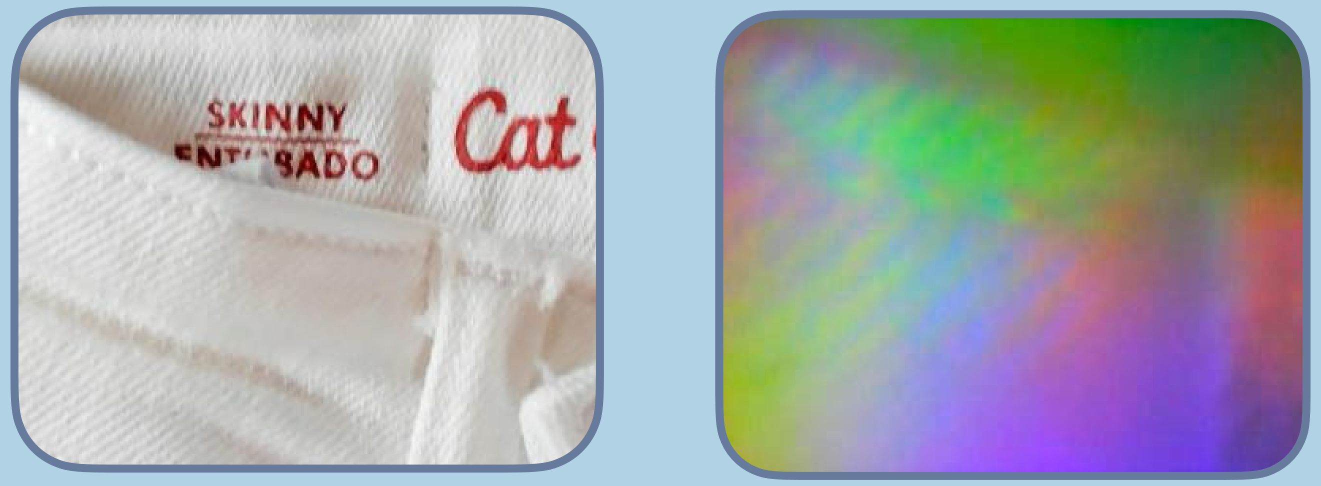
TVL-Benchmark

Vision Language Models

This image gives tactile feelings of?

"Soft, smooth"

Touch Vision Language Models



This image gives tactile feelings of?

"Soft, woven, and pliable"

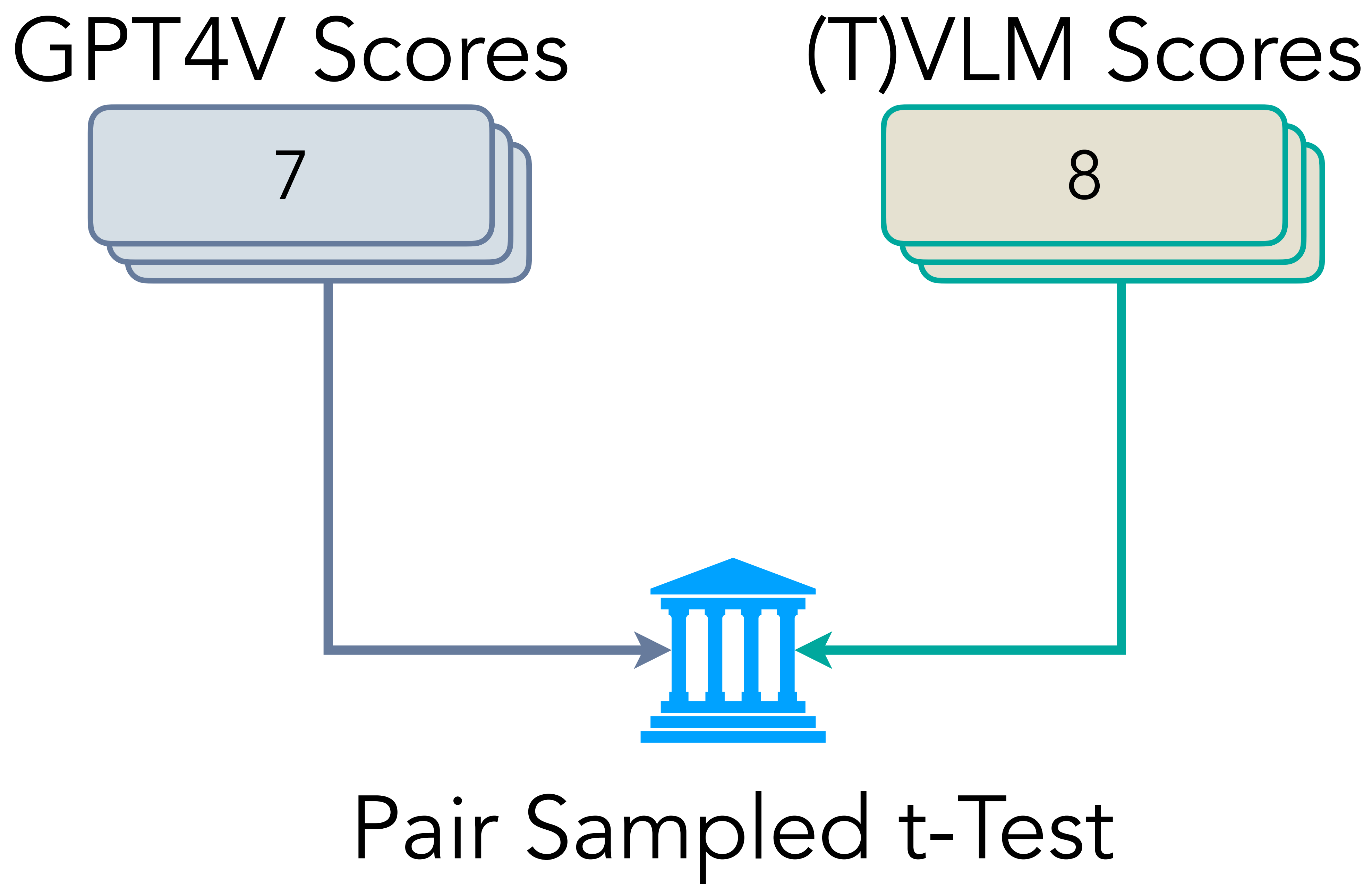
+

TVL-Benchmark

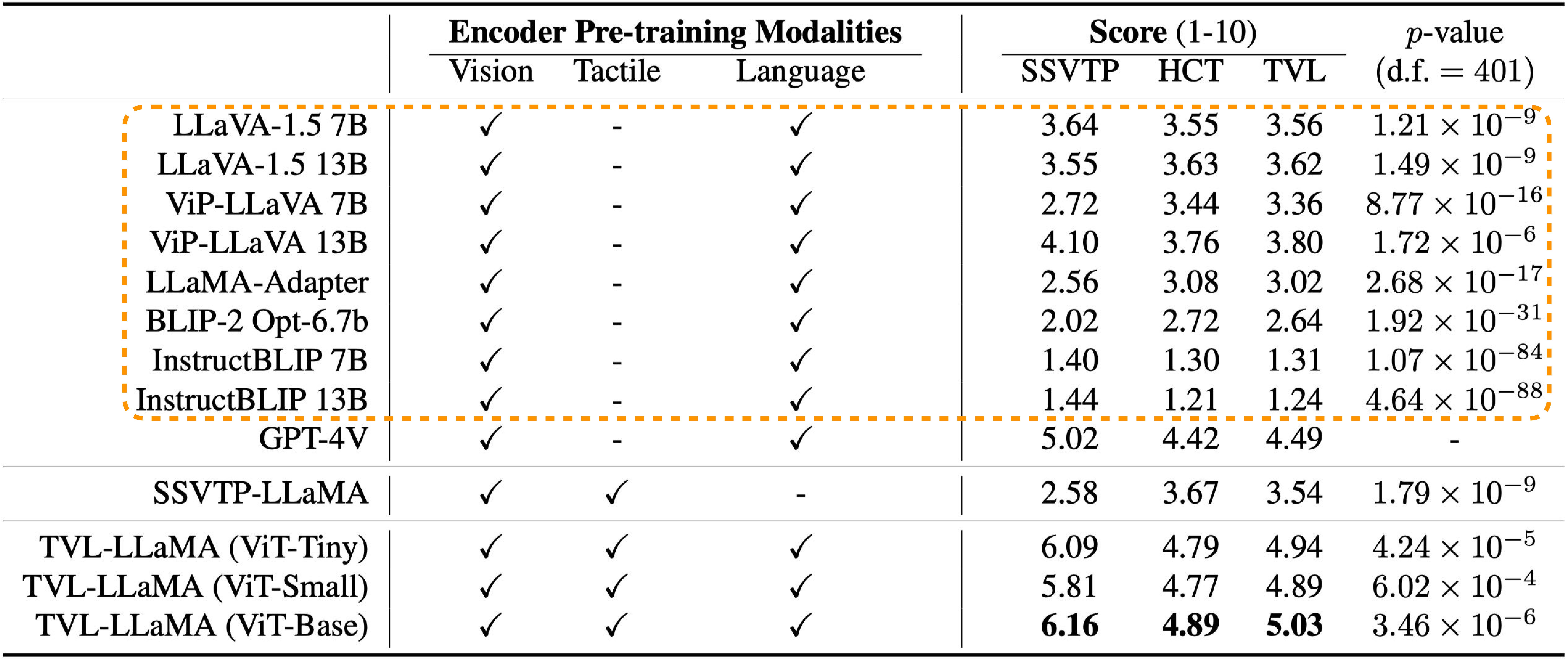
"Soft, smooth"

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 …

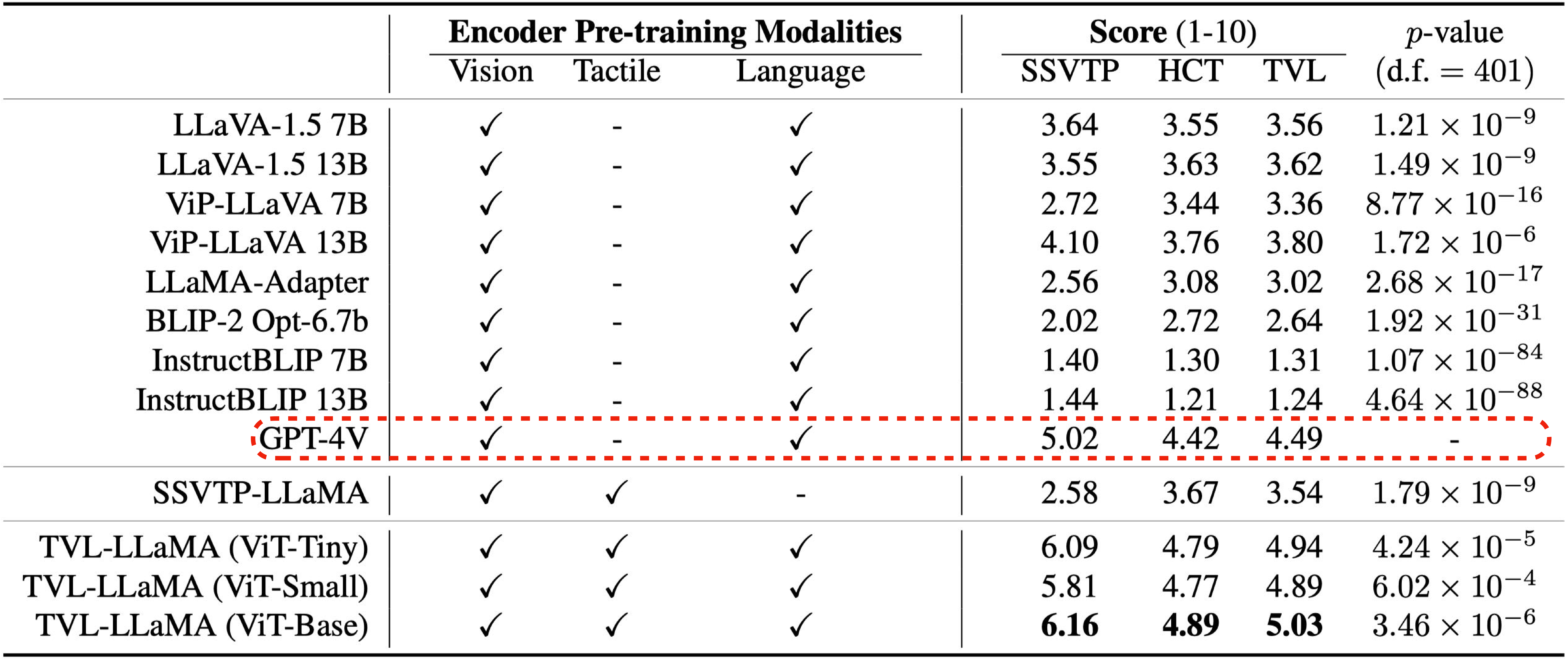

GPT4V Scores
7
(T)VLM Scores
8
Pair Sampled t-Test

TVL-Benchmark



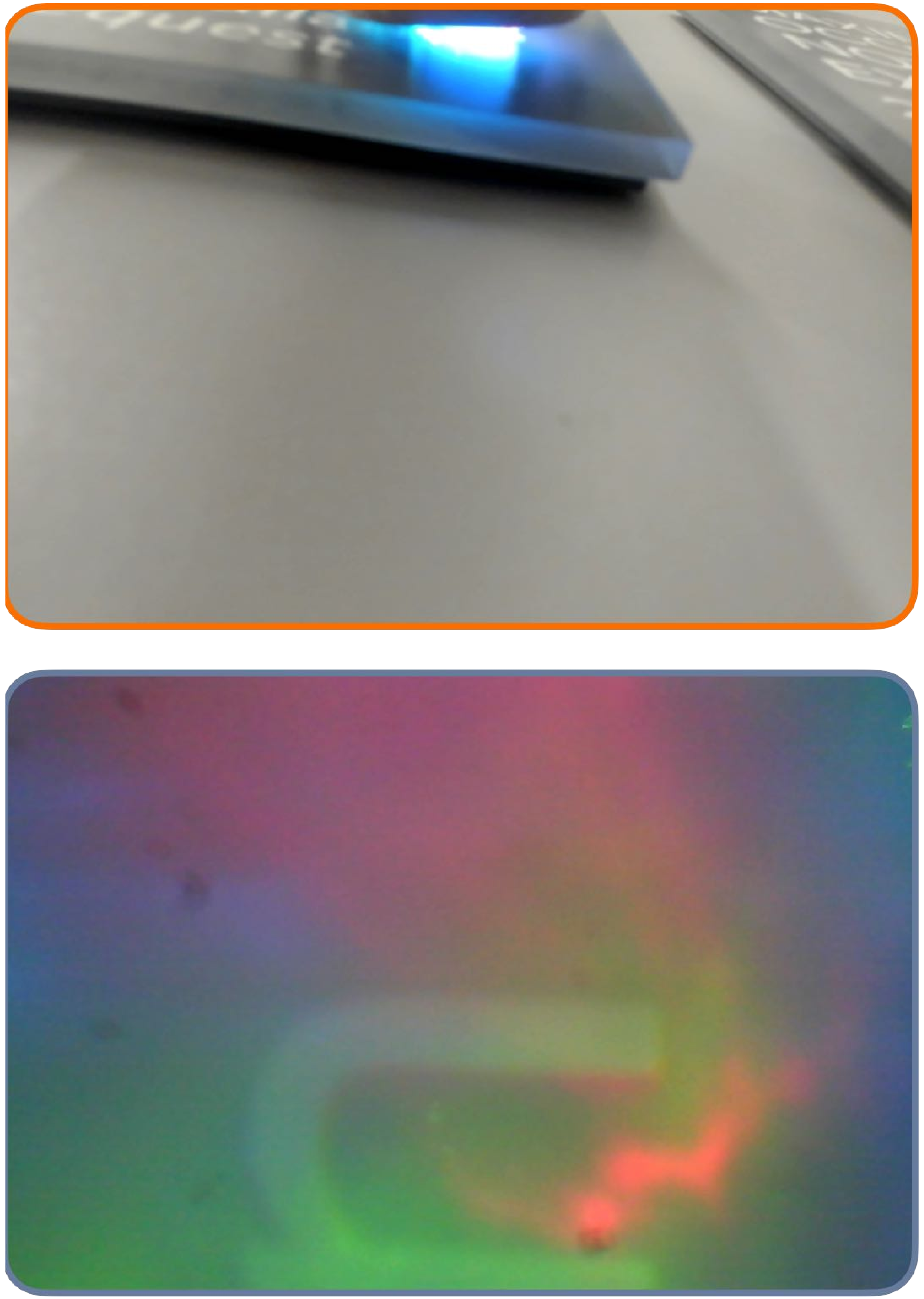
	Encoder Pre-training Modalities			**Score** (1-10)			p-value
	Vision	Tactile	Language	SSVTP	HCT	TVL	(d.f. = 401)
LLaVA-1.5 7B	✓	-	✓	3.64	3.55	3.56	1.21×10^{-9}
LLaVA-1.5 13B	✓	-	✓	3.55	3.63	3.62	1.49×10^{-9}
ViP-LLaVA 7B	✓	-	✓	2.72	3.44	3.36	8.77×10^{-16}
ViP-LLaVA 13B	✓	-	✓	4.10	3.76	3.80	1.72×10^{-6}
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}

TVL-Benchmark



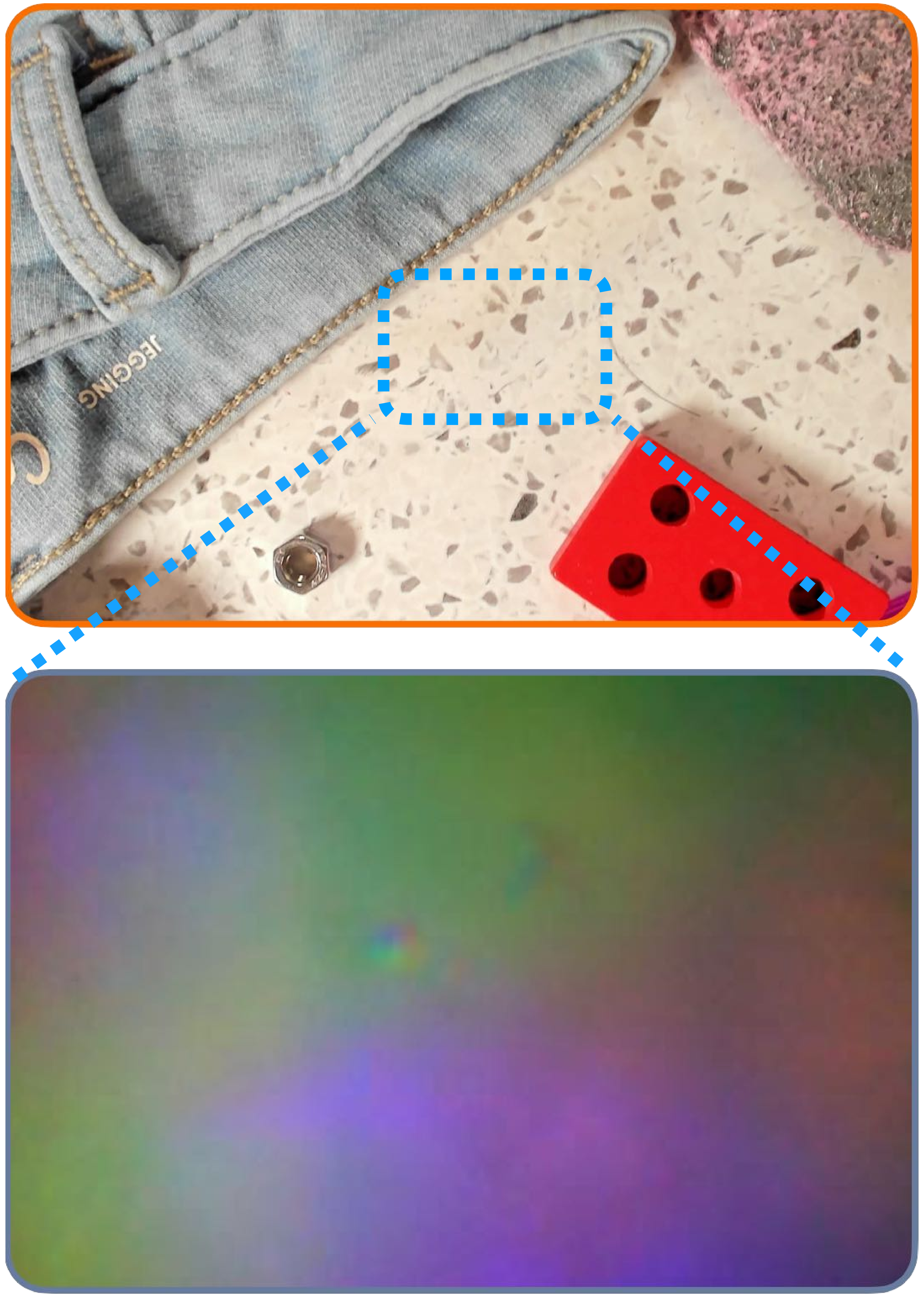
	Encoder Pre-training Modalities			**Score** (1-10)			p-value
	Vision	Tactile	Language	SSVTP	HCT	TVL	(d.f. $= 401$)
LLaVA-1.5 7B	✓	-	✓	3.64	3.55	3.56	1.21×10^{-9}
LLaVA-1.5 13B	✓	-	✓	3.55	3.63	3.62	1.49×10^{-9}
ViP-LLaVA 7B	✓	-	✓	2.72	3.44	3.36	8.77×10^{-16}
ViP-LLaVA 13B	✓	-	✓	4.10	3.76	3.80	1.72×10^{-6}
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



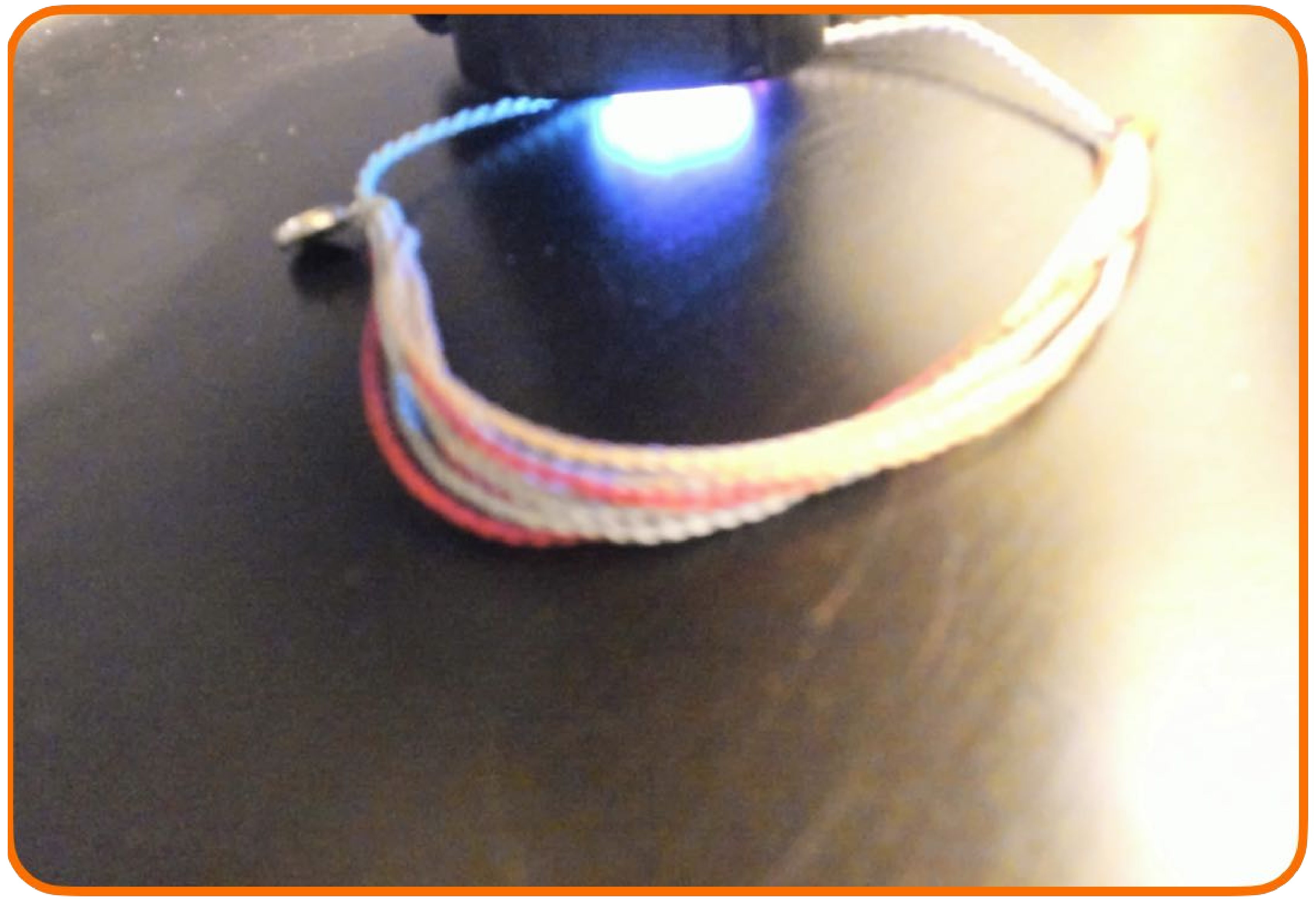
Model	Description	Score (1-10)
Human Labels	smooth, hard, sleek	GT
TVL-LLaMA	reflective, cool, sleek, smooth.	8
GPT-4V	smooth, reflective, hard, cool, sleek	8
SSVTP-LLaMA	smooth, reflective, hard, cool, sleek.	8.5
LLaVA 1.5 7B:	Slick, cold, smooth, hard, reflective.	8.5
LLaVA 1.5 13B	The surface is smooth, 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

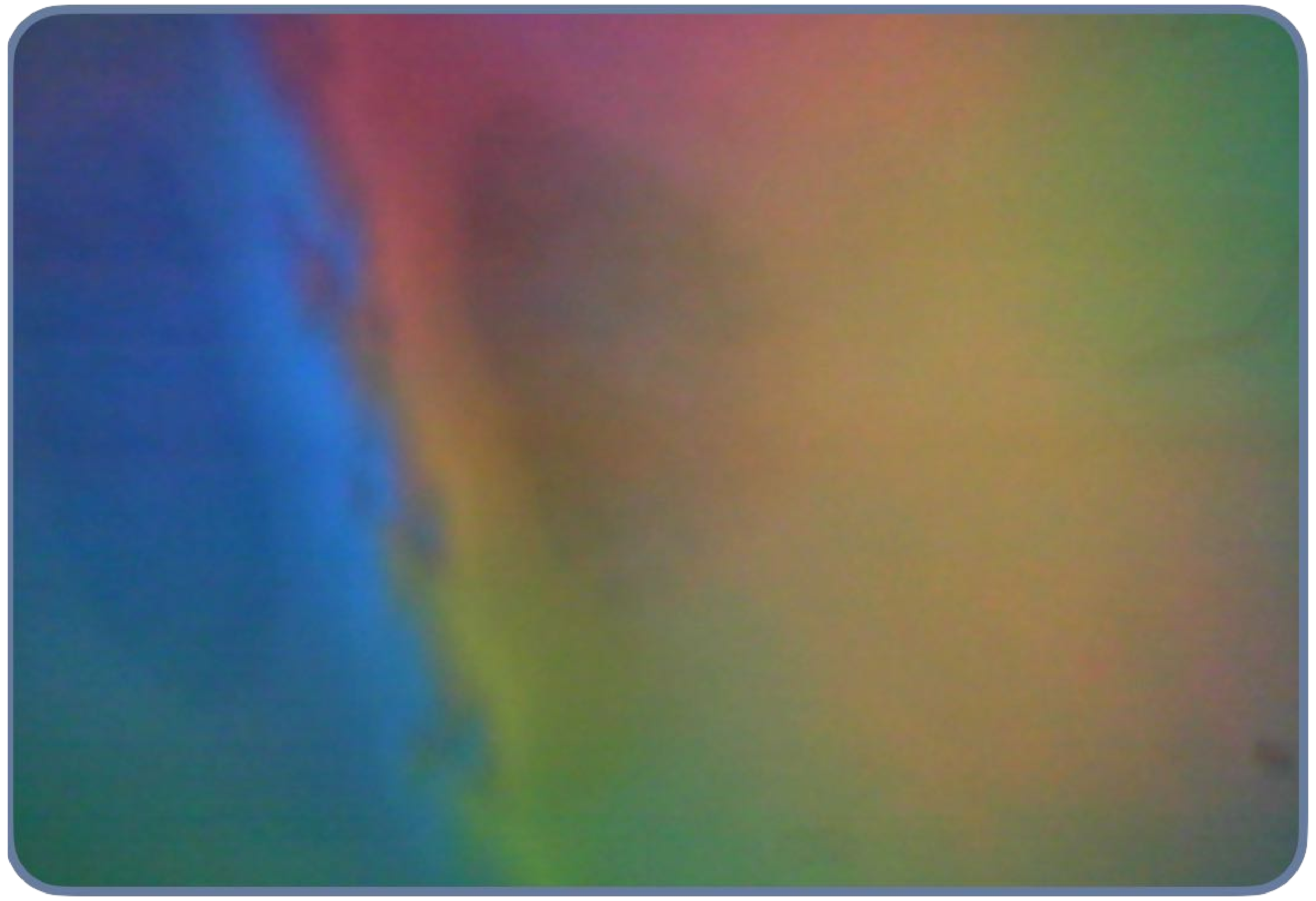
Examples



Model	Description	Score (1-10)
Human Labels	hard, flat, splotchy, grainy	GT
TVL-LLaMA	flat, hard.	6.5
GPT-4V	Textured, solid, cool, rough, uneven	4
SSVTP-LLaMA	smooth, reflective, hard, cool, sleek.	2
LLaVA 1.5 7B:	Slick, rough, cold, hard, glassy.	5
LLaVA 1.5 13B	For a white, speckled surface, the description might be "pebble, textured, matte, porous, spotty."	6.5
ViP-LLaVA 7B	Shiny, cool, smooth, hard, unyielding.	2
ViP-LLaVA 13B	Smooth, hard, cool, sturdy, firm.	4
LLaMA-Adapter	1. Slick, hard, unyielding, cold, glossy.	6.5
BLIP-2 OPT-6.7	the patch is soft, velvety, smooth and slippery	1
InstructBLIP 7B	some dice and a blue pair of jeans on the table	1
InstructBLIP 13B	a close up of a person's jeans and a dice 2	1

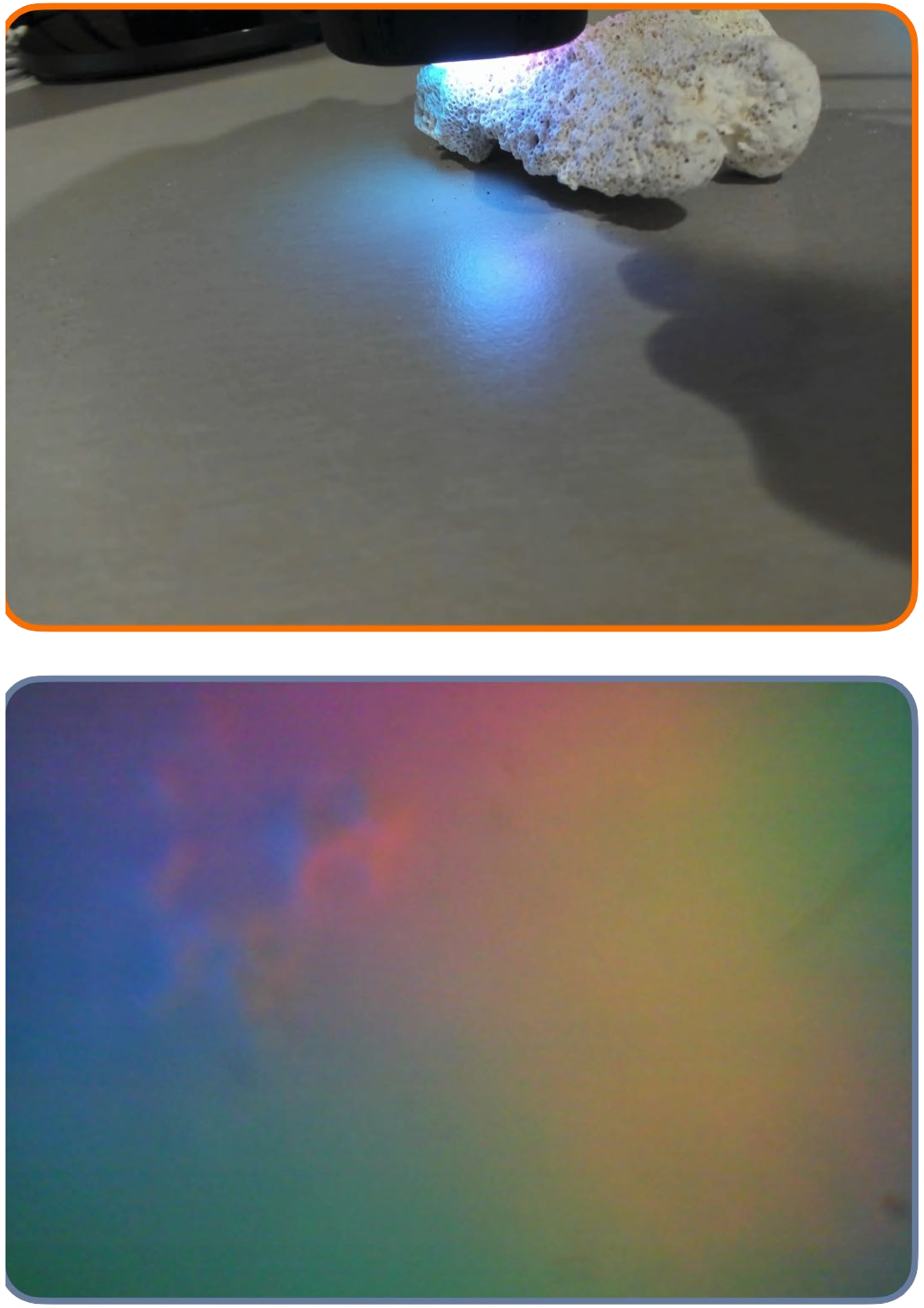
Examples





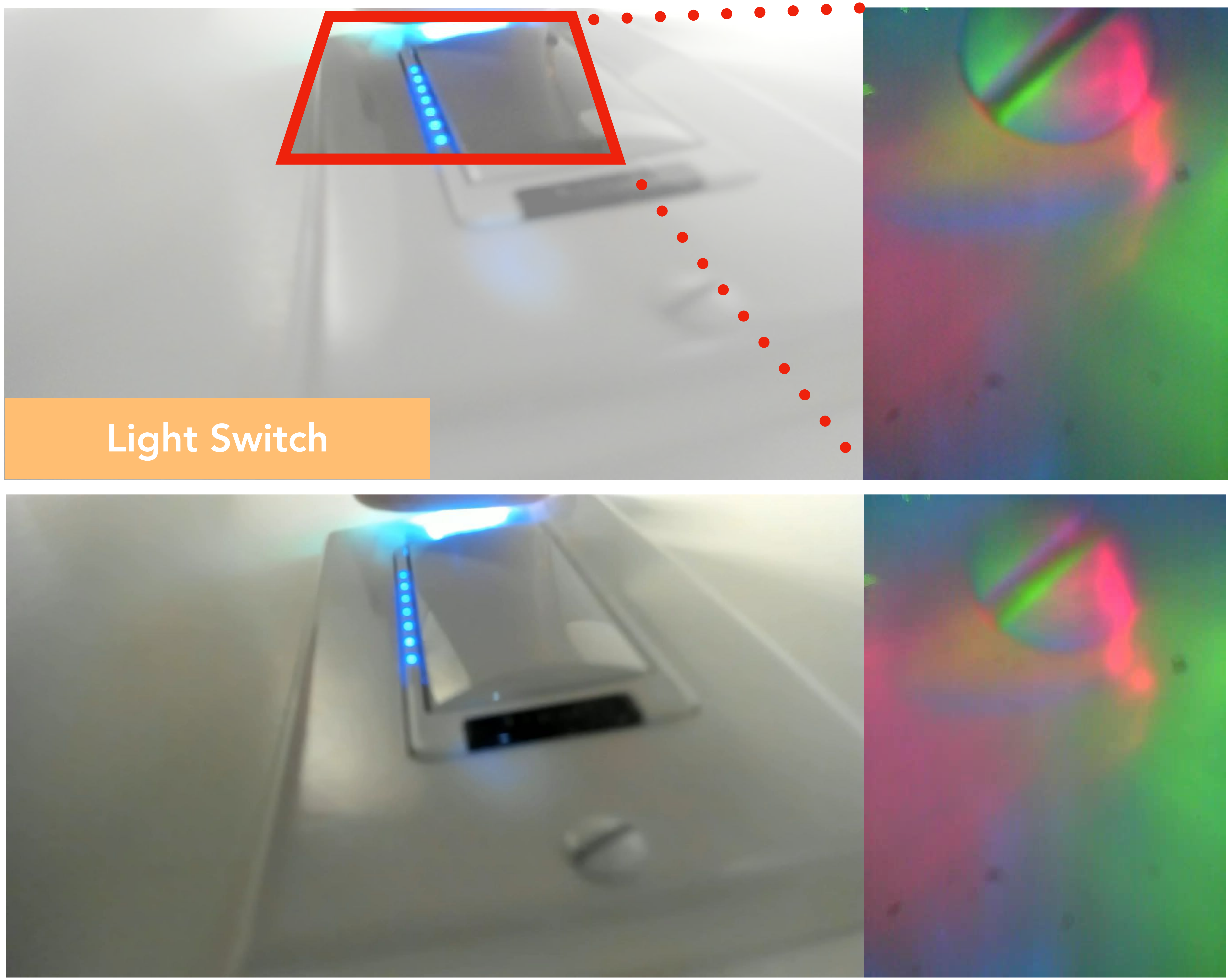
Model	Description	Score (1-10)
Human Labels	fibrous, textured, uneven, pliable, patterned	GT
TVL-LLaMA	woven, flexible, soft, glossy.	6.5
GPT-4V	smooth, reflective, hard, cool, glossy	1
SSVTP-LLaMA	smooth, reflective, hard, cool, sleek.	1
LLaVA 1.5 7B:	Cool, gritty, textured, adhesive, reflective	6.5
LLaVA 1.5 13B	For the patch where the sensor makes contact, the possible feeling could be described as "smooth, soft, textured, slightly warm, and secure."	2
ViP-LLaVA 7B	Slick, smooth, shiny, glossy, reflective	1
ViP-LLaVA 13B	Shiny, reflective, firm, smooth, cool	1
LLaMA-Adapter	1. Slick, hard, unyielding, cold, glossy.	1
BLIP-2 OPT-6.7	The sensor is shiny and there is a smooth surface.	2
InstructBLIP 7B	A multicolored bracelet sitting on top of a black table	1
InstructBLIP 13B	A wrist band with a light is on a table 81-1	1

Examples



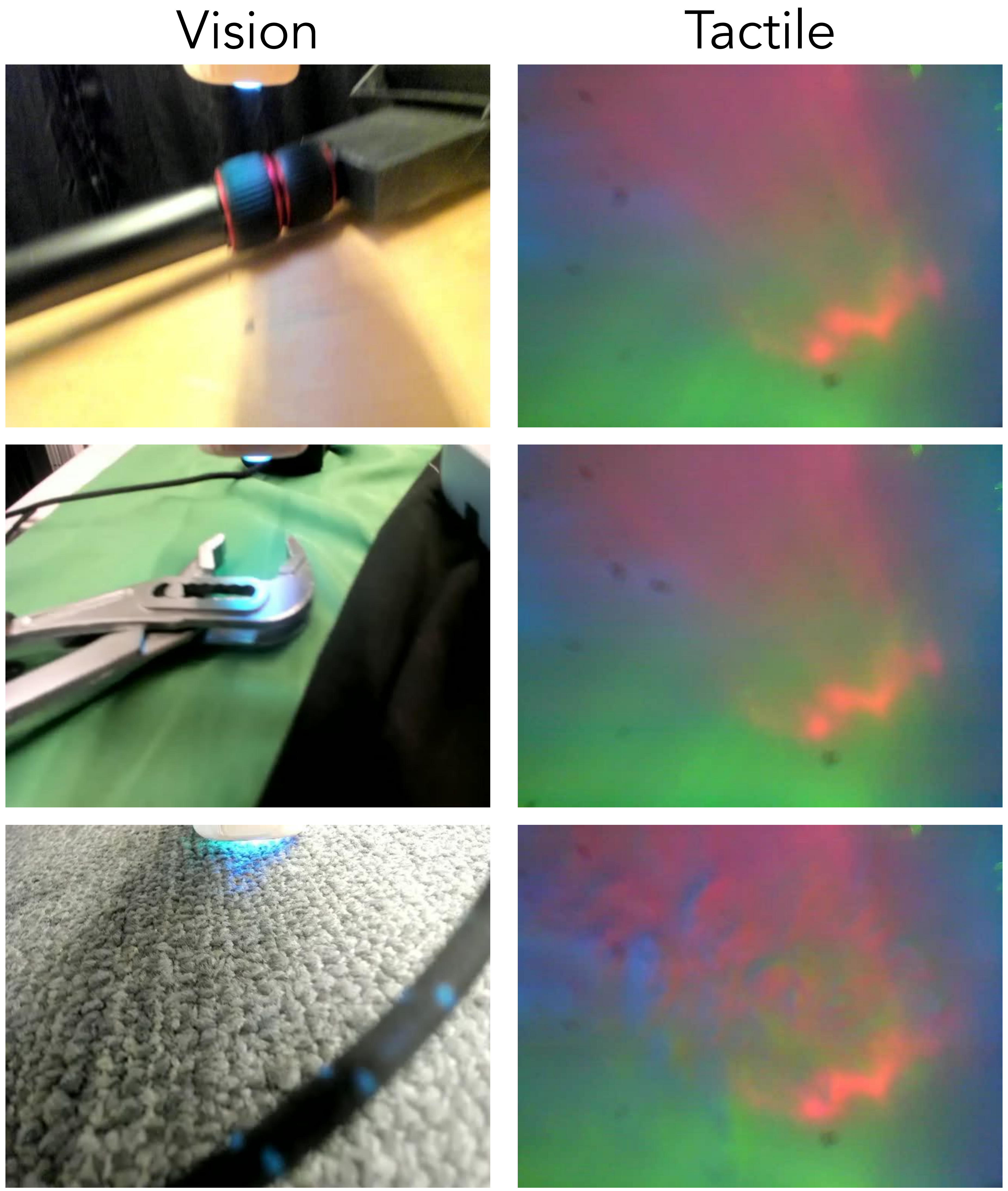
Model	Description	Score (1-10)
Human Labels	uneven, grainy, bumpy, hard	GT
TVL-LLaMA	smooth, matte, firm, cool, flat.	1
GPT-4V	smooth, cool, hard, matte, solid	2
SSVTP-LLaMA	smooth, reflective, hard, cool, sleek.	2
LLaVA 1.5 7B:	Shiny, silky, smooth, slippery, polished.	1
LLaVA 1.5 13B	Glossy, smooth, cool, hard, and non-porous are some of the adjectives that could be used to describe the surface where the tactile sensor makes contact.	2
ViP-LLaVA 7B	Slick, gritty, textured, sharp, rough	5.5
ViP-LLaVA 13B	Slick, gritty, textured, sharp, rough	1
LLaMA-Adapter	1. Slick 2. Cold 3. Hard 4. Unyielding 5. Glossy.	4
BLIP-2 OPT-6.7	The first word that came to my mind was "sharp" because I can see the pointy edge. I also think it's very comfortable because it's	2
InstructBLIP 7B	some rocks are being exposed to some shining light	2
InstructBLIP 13B	a black light is shown shining on a piece of coral 60924	1

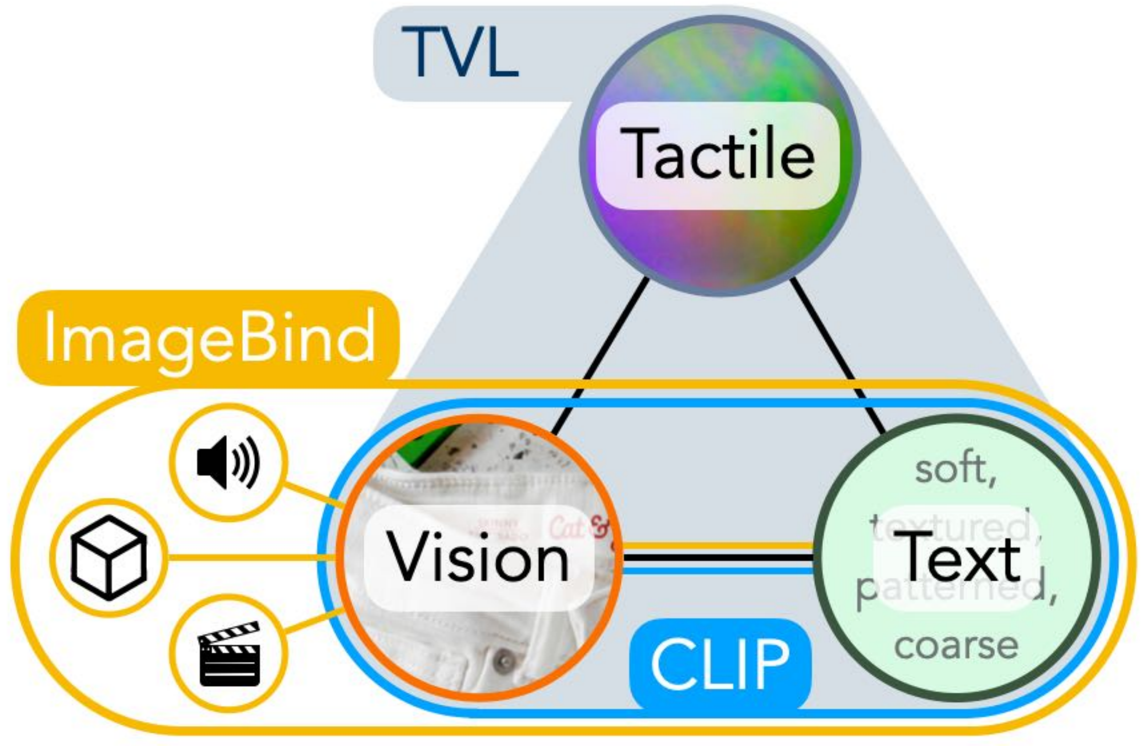
Limitations and Future Works



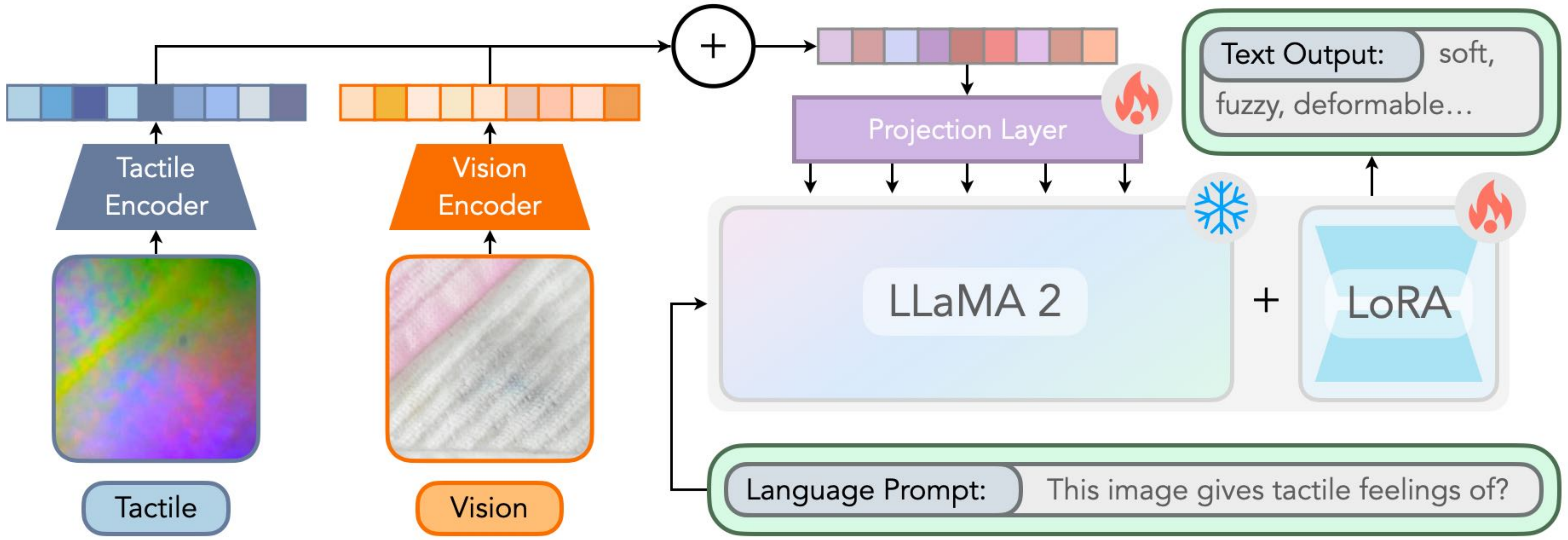
1) Occlusion

Limitations and Future Works





TVL-Tactile Encoder


Tactile Encoder
Vision Encoder
Tactile
Vision
+
Projection Layer
Text Output: soft, fuzzy, deformable…
LLaMA 2
+
LoRA
Language Prompt: This image gives tactile feelings of?

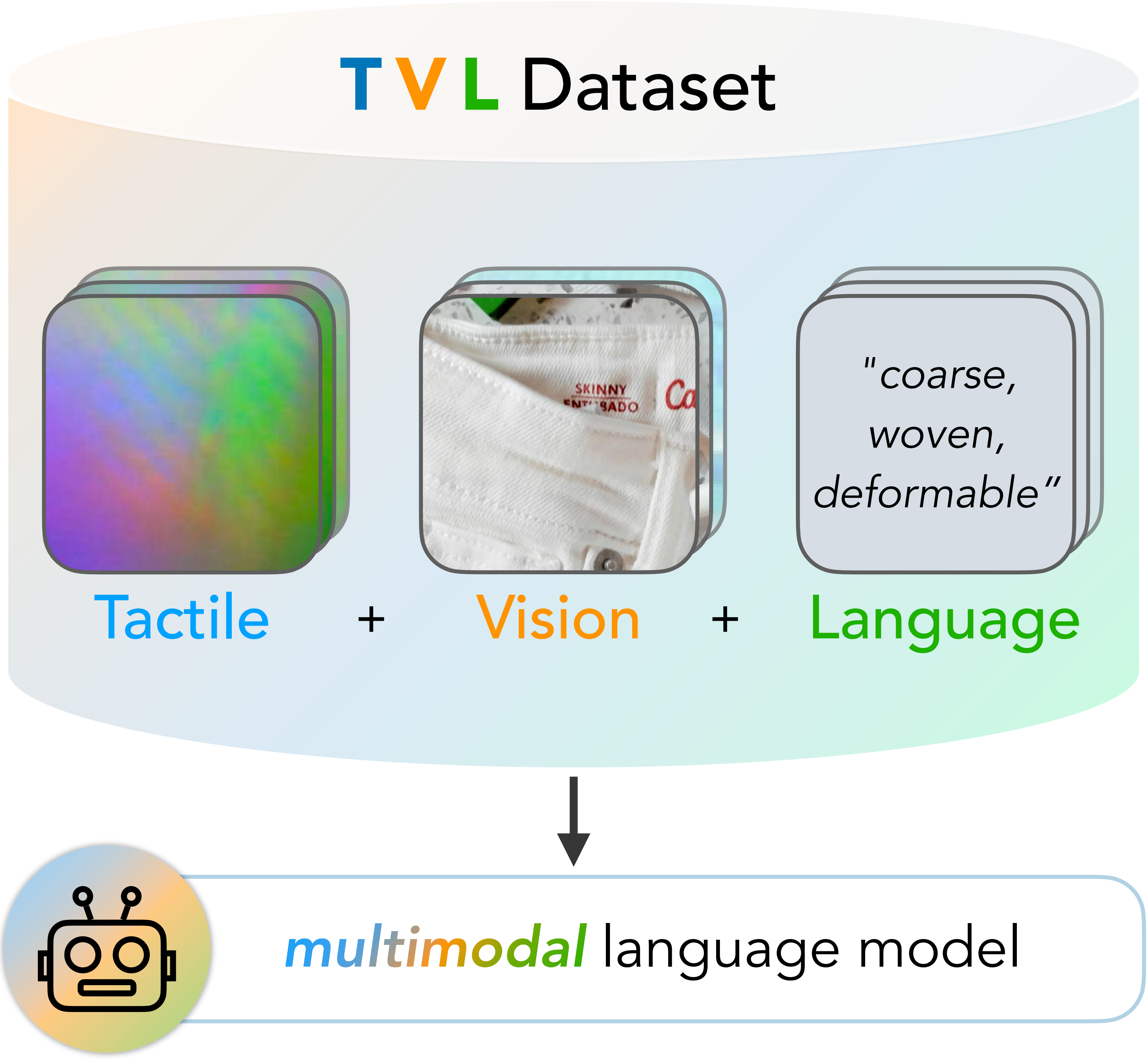
TVL-LLaMA

2) Does not model time/dynamics

A Touch, Vision, and 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

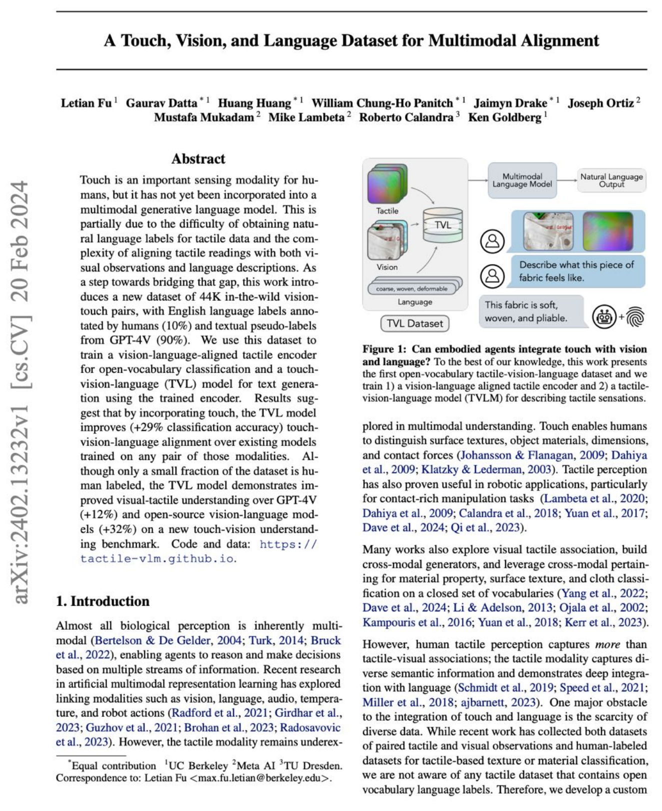
Berkeley UNIVERSITY OF CALIFORNIA
Meta
TECHNISCHE UNIVERSITÄT DRESDEN


TVL Dataset
"coarse, woven, deformable"
Tactile + Vision + Language
multimodal language model

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 Fu[1] Gaurav Datta[*1] Huang Huang[*1] William Chung-Ho Panitch[*1] Jaimyn Drake[*1] Joseph Ortiz[2]
Mustafa Mukadam[2] Mike Lambeta[2] Roberto Calandra[3] Ken Goldberg[1]

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

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 those 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 (Bertelson & De Gelder, 2004; Turk, 2014; Bruck 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; Girdhar et al., 2023; Guzhov et al., 2021; Brohan et al., 2023; Radosavovic 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; Dahiya et al., 2009; Klatzky & Lederman, 2003). Tactile perception has also proven useful in robotic applications, particularly for contact-rich manipulation tasks (Lambeta et al., 2020; Dahiya 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 pertaining 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; Kampouris 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; ajbarnett, 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

Tactile
Vision
TVL
coarse, woven, deformable
Language
TVL Dataset
Multimodal Language Model
Natural Language Output
Describe what this piece of fabric feels like.
This fabric is soft, woven, and pliable.

Figure 1: Can embodied agents integrate touch with vision and language? To the best of our knowledge, this work presents the first open-vocabulary tactile-vision-language dataset and we train 1) a vision-language aligned tactile encoder and 2) a tactile-vision-language model (TVLM) for describing tactile sensations.

---

*Equal contribution [1]UC Berkeley [2]Meta AI [3]TU Dresden. Correspondence to: Letian Fu <max.fu.letian@berkeley.edu>.

1

+