

222 **A Additional Results**

223 **A.1 Ablations**

224 This section presents six ablation and sensitivity analyses shown in Table 2 examining the impact of
 225 model size and the proposed dataset on the encoder’s multi-modal classification performance. More
 226 ablations are included in the appendix.

227 **Model Sizes** (Table 2a) Performance varies significantly among different encoder sizes. ViT-Base has
 228 the highest validation accuracy but lags on the test set due to distribution shifts: the training labels
 229 from GPT-4V are less detailed and accurate compared to human-annotated test data. However, in
 230 tactile-vision classification on synchronized data, ViT-Base outperforms both of the smaller models.

231 **Disable Tactile-Text Loss** (Table 2b) resembles the setup in ImageBind [12], where data in all three
 232 modalities are considered but the tactile-text loss is omitted. Results suggest that using language to
 233 supervise the tactile encoder better aligns those two modalities.

234 **Data** (Tables 3c-f) We perform four sensitivity analyses on the different compositions of the dataset for
 235 training. We find that leveraging data from all three modalities improves tactile-language alignment.
 236 While adding not-in-contact data prevents the model from overfitting to the training set, its test set
 237 performance is comparable with having only in-contact data. We also experimented with prompting
 238 used in vanilla CLIP training [28], which brings marginal improvements in accuracy. Lastly, we
 239 separately train the model on SSVTP and HCT, and we find that the pseudo-labeled dataset can
 240 provide comparable performance with training on the entire dataset, which suggests that TVL’s
 241 tactile encoder can effectively leverage self-supervised learning to reduce the dependency on large,
 242 fully-labeled datasets while maintaining task performance.

243 **A.2 Performance Per Dataset**

244 In this section, we show a fine-grained breakdown of Table 1 of model performance on the TVU
 245 benchmark by showing the results per subset of the dataset. The performance of the models on the
 246 SSVTP subset is listed in Table 3 and the performance on the HCT subset is listed in Table 4. Results
 247 suggest that GPT-4V performs better on SSVTP, which is collected in a lab setting, than HCT, which
 248 is collected “in-the-wild”.

Model	Tac./Text % Acc.	Tac./Vis. % Acc.	Tactile- Text Loss	Tac./Text % Acc.	Tac./Vis. % Acc.	Modality	Tac./Text % Acc.	Tac./Vis. % Acc.
ViT-Tiny	36.7	79.5	Enabled	36.3	78.0	All	36.3	78.0
ViT-Small	36.3	78.0	Disabled	20.3	81.6	-Vision	29.9	1.0
ViT-Base	30.7	81.7				-Text	21.5	85.8

(a) **Model Architecture** used for transformer encoder backbone.

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

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

Contact	Tac./Text % Acc.	Tac./Vis. % Acc.	Prompting	Tac./Text % Acc.	Tac./Vis. % Acc.	Dataset	Tac./Text % Acc.	Tac./Vis. % Acc.
Contact	36.2	80.1	Baseline	36.3	78.0	SSVTP	19.2	8.0
+ 10% N.C.	36.3	78.0	+ Prompt	37.7	78.7	HCT	38.4	74.4
						TVL	36.3	78.0

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

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

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

Table 2: Ablations and Sensitivity Analysis for the TVL tactile encoder. We report top-1 and top-5 tactile-text and tactile-vision classification accuracy with ViT-Small. *baseline* indicates the default setting for training the TVL tactile encoder, which is the best-performing model on the *validation set* unless noted otherwise. **Bold** indicates the highest accuracy on the *test set*. Such discrepancy in performance is described in Appendix A.1.

249 A model that is trained with a large sample of only GPT-4V labels should achieve the same perfor-
 250 mance as GPT-4V. Our results in Table 4 suggest that training on a small dataset of human-labeled
 251 vision-touch **improves** the model’s tactile-visual understanding. This difference is statistically
 252 significant at $\alpha = 0.05$.

	Score (1-10)	p -value (d.f. = 401)
LLaVA-1.5 7B	3.64	2.32×10^{-3}
LLaVA-1.5 13B	3.55	1.30×10^{-3}
ViP-LLaVA 7B	2.72	4.45×10^{-8}
ViP-LLaVA 13B	4.10	3.76×10^{-2}
LLaMA-Adapter	2.56	7.826×10^{-6}
BLIP-2 Opt-6.7b	2.02	2.74×10^{-9}
InstructBLIP 7B	1.40	1.49×10^{-13}
InstructBLIP 13B	1.44	4.68×10^{-14}
GPT-4V	5.02	-
SSVTP-LLaMA	2.58	9.33×10^{-6}
TVL-LLaMA (ViT-Tiny)	6.09	2.65×10^{-2}
TVL-LLaMA (ViT-Small)	5.81	1.02×10^{-1}
TVL-LLaMA (ViT-Base)	6.16	1.67×10^{-2}

Table 3: TVL Benchmark Performance on SSVTP. We benchmarked TVL-LLaMA against existing VLMs and SSVTP-LLaMA, and show here the performance on only the SSVTP dataset. We report p -values from two-sided paired sample t -tests on each model’s scores against GPT-4V’s scores.

	Score (1-10)	p -value (d.f. = 401)
LLaVA-1.5 7B	3.55	8.49×10^{-8}
LLaVA-1.5 13B	3.63	1.74×10^{-7}
ViP-LLaVA 7B	3.44	4.10×10^{-11}
ViP-LLaVA 13B	3.76	1.57×10^{-5}
LLaMA-Adapter	3.08	2.05×10^{-13}
BLIP-2 Opt-6.7b	2.72	1.25×10^{-24}
InstructBLIP 7B	1.30	8.02×10^{-73}
InstructBLIP 13B	1.21	9.74×10^{-76}
GPT-4V	4.42	-
SSVTP-LLaMA	3.67	3.24×10^{-6}
TVL-LLaMA (ViT-Tiny)	4.79	5.79×10^{-4}
TVL-LLaMA (ViT-Small)	4.77	2.64×10^{-3}
TVL-LLaMA (ViT-Base)	4.89	6.82×10^{-5}

Table 4: TVL Benchmark Performance on HCT. We benchmarked TVL-LLaMA against existing VLMs and SSVTP-LLaMA, and show here the performance on only the HCT dataset. We report p -values from two-sided paired sample t -tests on each model’s scores against GPT-4V’s scores.

253 **A.3 Open Vocabulary Tactile Classification Full Result**

254 We present the result presented in ?? in Table 5 and Table 6 at different cosine similarity thresholds
 255 for synonyms. We find that while ViT-Small performs well on the SSVTP subset of the dataset, ViT-
 256 Tiny outperforms its larger counterparts (ViT-Small and ViT-Base) on the tactile-text classification
 257 task. However, for tactile-vision classification (Table 6), ViT-Base performs outperforms the smaller
 258 models. More insights are detailed in Appendix B.1.

Percentile		SSVTP		HCT		TVL	
		Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
0	ViT-Tiny	29.4%	71.7%	34.8%	70.1%	36.7%	70.3%
	ViT-Small	42.4%	76.1%	36.5%	68.0%	36.3%	66.4%
	ViT-Base	38.0%	69.6%	34.8%	65.6%	30.7%	63.6%
25	ViT-Tiny	3.3%	21.7%	7.2%	22.9%	4.6%	14.1%
	ViT-Small	10.9%	33.7%	9.1%	21.5%	6.7%	19.5%
	ViT-Base	8.7%	31.5%	5.9%	14.0%	4.4%	13.7%
50	ViT-Tiny	3.3%	19.6%	4.8%	17.8%	3.7%	11.8%
	ViT-Small	10.9%	32.6%	6.6%	15.3%	5.9%	11.0%
	ViT-Base	7.6%	28.3%	4.5%	9.8%	3.5%	11.0%
75	ViT-Tiny	3.3%	19.6%	4.1%	14.2%	3.7%	10.7%
	ViT-Small	10.9%	28.3%	3.5%	7.9%	3.4%	10.2%
	ViT-Base	7.6%	28.3%	3.5%	7.9%	3.4%	10.2%

Table 5: Effect of Model Architecture and Similarity Threshold ϕ on **Tactile-Text** Classification Accuracy. The similarity thresholds ϕ for each percentile are 0.636 (0th), 0.859 (25th), 0.893 (50th), and 0.921 (75th).

	SSVTP		HCT		TVL	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
ViT-Tiny	34.8%	70.7%	85.3%	99.0%	79.5%	95.7%
ViT-Small	28.3%	69.6%	84.4%	98.9%	78.0%	95.2%
ViT-Base	34.8%	66.3%	87.8%	99.7%	81.7%	95.7%

Table 6: Effect of Tactile Encoder Model Architecture on **Tactile-Vision** Classification.

259 A.4 Additional Open Vocabulary Downstream Tasks

260 In the tactile classification experiment in ??, the results suggest that the model can classify tactile
261 inputs by the texture of surfaces. In this section, we add an experiment to perform *object category*
262 *classifications*. For simplicity of this test, we perform binary classification of whether the touched
263 surface is “fabric” or “plastic” (to answer the question of “identifying the object category”). Note that
264 since the model binds to the CLIP latent space, we carry out the experiment in a zero-shot manner.
265 We relabelled 50 instances in the test set with 20 as fabric and 30 as plastic. We then fed “fabric” and
266 “plastic” into the CLIP text encoder to extract the latent to perform cosine-similarity calculation with
267 the tactile latent extracted from the tactile observations. On this specific test, the ViT-Small version
268 of the TVL tactile encoder achieved 82% classification accuracy. We hope future works can explore
269 other potential downstream applications of the dataset and the learned tactile representations.

270 B Training Details and Hyperparameters

271 In this section, we offer more insights and details of the training process and the particular hyperpa-
272 rameters.

273 B.1 Overfitting to Pseudo-labels

274 A core obstacle with leveraging pseudo-labels generated by GPT-4V (gpt-4-vision-preview) is that
275 the logits are not provided for us to build uncertain estimates for the generated labels, which is usually
276 required for prior works in computer vision that leverages pseudo-labels for model prediction (*e.g.*
277 Lee et al. [19], Sohn et al. [30], Wang et al. [35]). This makes pseudo-labels noisy and challenging
278 to fit for ViT-Small on the contact only dataset, even when 4K human labels are introduced (see
279 Figure 3).

280 In 3.1, we address this problem by letting 10% of the data be in contact. We sample 10% of the
281 data uniformly at random without replacement at the start of the training. This prevents the model
282 from overfitting on all three model sizes: (ViT-Tiny, ViT-Small, and ViT-Base). However, since the
283 test set is all labeled by human annotators, the distribution shift leads to worse tactile-image, and
284 tactile-language classification performance (observed in ??). As an ablation study, we also finetuned
285 the ViT-Small trained only on in-contact data for tactile language generation. The test set performance

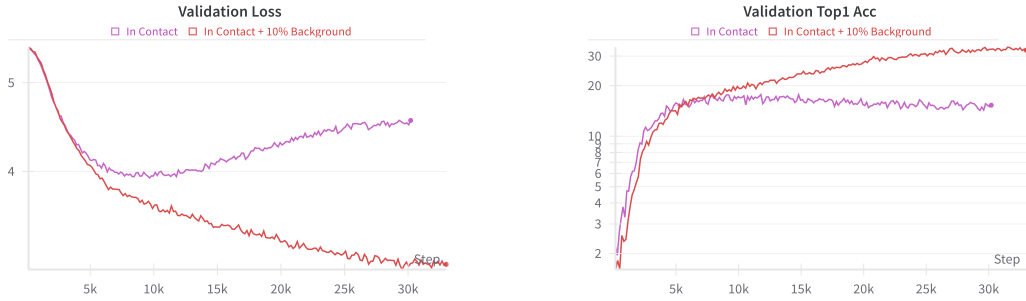


Figure 3: Overfitting is significant when all data is in contact. When 10% not in contact data is added, the overfitting issue is addressed.

286 is 4.81, only very marginally lower than that obtained by the ViT-Small trained with not-in-contact
 287 data (4.89). Future works can look into how to scale with noisy inputs or leverage existing works on
 learning from a teacher model that does not give uncertain estimates.

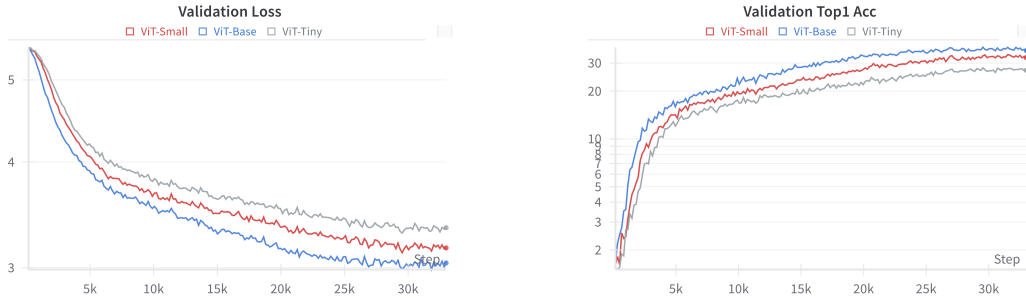


Figure 4: While we find that the model scales on the dataset, the test set performance does not align with the validation set performance. One potential cause of this is distribution shift: the validation set uses pseudo-labels generated by GPT-4V, while the test set is human-labeled.

288

289 B.2 Ablation: Background Subtraction

290 While we find that naively performing contrastive learning amongst tactile, vision, and language
 291 works for zero-shot classification, to further facilitate generalization across different tactile sensors
 292 used in data collection, a solution is to leverage the still background of tactile sensors (*i.e.* the readings
 293 from the sensor when it is not in contact). We preprocess the tactile observation by performing
 294 background subtraction, and normalize the input observations based on the post-processed dataset
 295 statistics. Empirically, we find that this method, when used jointly with not-in-contact data, improves
 classification accuracy and the downstream TVL-LLaMA’s performance (Table 7).

	Tac./Text % Acc	Tac./Vis % Acc	TVL Score
In-Contact Frames	36.2	80.1	4.81
+10% No-Contact	36.3	78.0	4.89
+ Background Subtract	42.3	78.9	5.06

Table 7: Effect of no-contact data and background subtraction during ViT-Small tactile encoder training on classification accuracy and performance on the TVL benchmark.

296

297 **B.3 Ablation: (Zero-shot) Single Modality For Generation (Out of Distribution)**

298 Because we naively average the tactile latent and the image latent during the training of TVL-LLaMA,
 299 as a zero-shot experiment to see consistency between vision and tactile embeddings, we can at *test*
 300 time arbitrarily drop one of the vision or tactile modalities. We report the results in Table 8. While a
 301 larger encoder may be more expressive, we find that a larger tactile encoder results in worse zero-shot
 302 performance in this experimental setting, which aligns with Table 2a. Interestingly, background
 subtraction (in Appendix B.2) improves the zero-shot performance on tactile.

	Zero-Shot Tactile	Zero-Shot Vision	Tactile & Vision
TVL-LLaMA (ViT-Tiny)	4.56	4.66	4.94
TVL-LLaMA (ViT-Small)	3.50	4.81	4.89
TVL-LLaMA (ViT-Base)	2.80	4.85	5.03
TVL-LLaMA (ViT-Small) + Background Subtract	4.52	-	5.06

Table 8: Dropping one modality (out-of-distribution) zero shot experiments

303

304 **B.4 Ablation: Finetuning v.s. Freezing the Language Model**

305 We add the experiment of just freezing the language model without LoRA fine-tuning. Interestingly,
 306 on the HCT test set, the frozen LLM with the trained encoders gives a score of 4.92, resulting in a
 307 marginal improvement compared to the score of a fine-tuned LLM of 4.89 (Table 1). This suggests
 308 that the vision and tactile modalities are already well aligned to the language space and further
 309 fine-tuning is unnecessary.

310 **B.5 Preprocessing**

311 The tactile observation is first zero-padded to have equal width and height, optionally background
 312 subtracted, normalized by the calculated data statistics, and resized the inputs to 224x224. The key
 313 differences with SSVTP are 1) the input is resized to 128x128, and 2) SSVTP does not perform
 314 normalization or background subtraction. The image observation follows the same center cropping
 315 procedure as SSVTP on the SSVTP dataset. On HCT, instead of the center crop, we start the crop
 316 from the top of the image but maintain the crop size. Note that this procedure is kept consistent when
 317 generating pseudo-labels from GPT-4V. Different from SSVTP, we use the statistics provided by
 318 OpenCLIP to normalize the post-crop observations. The specific statistics are provided in Table 9
 and Table 10.

Tactile Statistics	Mean	Std.
	0.292	0.188
With Background	0.297	0.195
	0.291	0.219
	-0.008	0.045
Background Subtracted	-0.019	0.044
	-0.018	0.053

Table 9: Tactile Normalization Statistics

319

320 **B.6 TVL Tactile Encoder Hyperparameters**

321 All of ViT-Tiny, ViT-Small, and ViT-Base share the same hyperparameters (see Table 11). All
 322 experiments are run on a single NVIDIA A100 GPU.

Image Statistics	Mean	Std.
	0.481	0.269
OpenCLIP Statistics	0.458	0.261
	0.408	0.276

Table 10: RGB Normalization Statistics

Config	Value
optimizer	AdamW [25]
base learning rate	1.5e-4
learning rate schedule	cosine decay [24]
batch size	256
weight decay	0.05
optimizer momentum	$\beta_1, \beta_2 = 0.9, 0.95$ [5]
warm up epoch [13]	10
total epochs	200
RGB Augmentation	RandomHorizontalFlip, ColorJitter, RandomGrayscale, GaussianBlur
Tactile Augmentation	(Optional) Background Subtraction

Table 11: Encoder Pretraining Hyperparameters

323 B.7 TVL-LLaMA Hyperparameters

324 We follow the hyperparameter setup in ImageBind-LLM [15]. Since the original experiments were
 325 conducted on 8 NVIDIA A100 GPUs, we use gradient accumulation of 2 for both pre-training and
 326 finetuning the model to fit the model on 4 NVIDIA A100 GPUs so that the batch size is maintained.
 327 We use the same data augmentation as in the encoder pretraining (Table 11).

328 C Dataset

329 C.1 Hardware

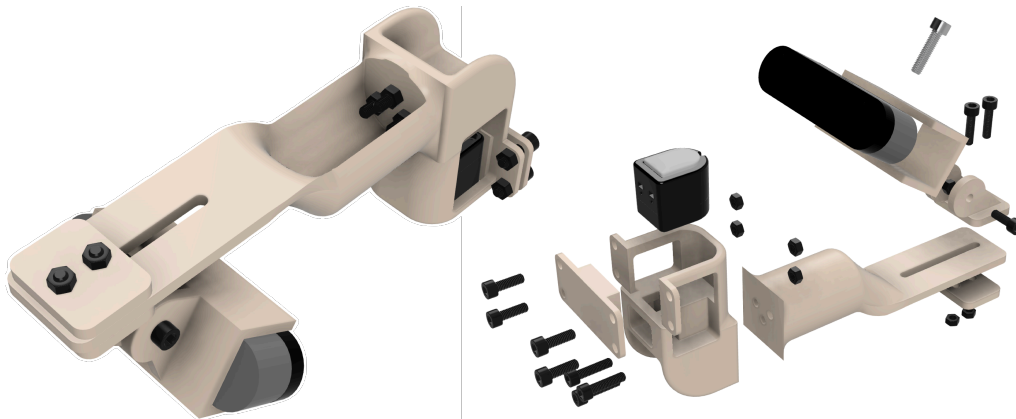


Figure 5: Alternative perspectives of the sensor holder CAD model: face-down view (left) and exploded view (right).

330 We design and 3D print a set of handheld, low-cost data collection devices for human subjects to
 331 carry around and collect data. As shown in Fig. 5, the hardware consists of a DIGIT tactile sensor and
 332 a Logitech BRIO camera, which are connected via USB to a portable computing device, such as a
 333 laptop. The angle and distance between the tactile sensor and the camera are adjustable, allowing the
 334 user to collect data from a variety of viewing angles and ranges. To ensure the utility of our dataset

335 for multimodal training, we always set the relative positions such that the tactile sensor and its point
336 of contact with the object of interest are in view of the camera during each trajectory. The handle
337 design was conceptualized in Autodesk Fusion 360 and printed on a Bambu Lab P1P 3D FDM printer.
338 CAD files will be open-sourced.

339 C.2 List of Prompts for Tactile Language Generation

340 When finetuning our language model for tactile language generation, we formulate it as a visual
341 instruction tuning problem [23]. We randomly select from the following set of semantically similar
342 prompts as the question and treat the set of human labels as the answer. This serves to increase the
343 diversity of data seen during training.

```
344 This image gives tactile feelings of
345 This image evokes a sense of
346 This visual representation imparts a tactile sensation of
347 This picture conveys a touchable quality of
348 This image communicates a palpable feeling of
349 This graphic suggests a tactile experience of
350 This artwork manifests a tangible sensation of
351 This visual elicits a haptic impression of
352 This depiction gives rise to a tactile perception of
353 This illustration induces a touch-sensitive feeling of
354 This photo brings forth a tactile awareness of
355 This image arouses a tactile familiarity of
356 This snapshot renders a tactile essence of
357 This visual stimulates a touch-based sensation of
358 This portrayal invokes a tactile resonance of
359 This image delivers a touch-oriented impression of
360 This visual medium offers a tactile nuance of
361 This rendering provides a tactile sense of
362 This image yields a touch-felt experience of
363 This composition reveals a tactile characteristic of
364 This picture bestows a tactile attribute of
365 This image imparts a sense of tactile
366 This visual stimulates tactile sensations of
367 This artwork hints at a tactile experience of
368 This photo embodies a tactile quality of
369 This depiction resonates with tactile feelings of
370 This snapshot conveys tactile impressions of
371 This illustration suggests a tactile nature of
372 This rendering evokes tactile attributes of
373 This graphic communicates a tactile essence of
374 This visual piece reveals tactile characteristics of
375 This image portrays tactile elements of
376 This picture brings to mind tactile aspects of
377 This visual representation offers tactile nuances of
378 This composition provides tactile insights into
379 This visual art form captures tactile features of
380 This image projects tactile properties of
381 This visual work hints at tactile textures of
382 This image introduces tactile dimensions of
383 This visual scene manifests tactile facets of
384 This image presents tactile qualities of
385 This image elucidates tactile attributes of
```

386 C.3 Distribution of Vocabulary Words

387 The list and counts of human labels and pseudo-labels in the TVL dataset are reproduced here in
388 dictionary format (note that all typos are carried over from the dataset). A visual representation is
389 provided in Figure 6.

```
390 'smooth': 14577, 'textured': 12443, 'hard': 10758, 'cool': 10433, 'reflective': 8643, 'soft': 8415,  
391 'glossy': 6416, 'cushioned': 6011, 'rigid': 5799, 'firm': 5659, 'sleek': 5628, 'uneven': 5379, 'flat':
```

Distribution of Tactile Descriptor Words in the TVL Dataset

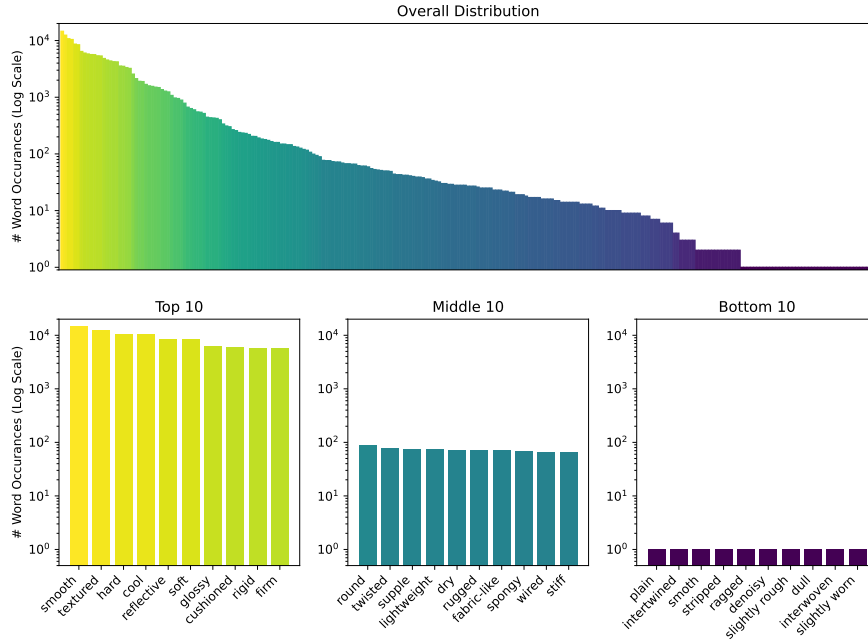


Figure 6: Distribution of Words in the TVL Dataset: The TVL dataset contains 254 unique tactile descriptors, ranging from common tactile descriptions (smooth, hard, firm) to unusual and optical descriptors. These less-common adjectives include a small fraction of misspellings and non-tactile descriptors which were generated by the VLM. The long-right-tailed distribution common in image classification [34] presents a challenge for learning predictors on tactile-semantic data as well.

392 5343, 'fibrous': 4825, 'plush': 4534, ',': 4363, 'matte': 4230, 'polished': 4203, 'flexible': 3553,
 393 'grainy': 3513, 'solid': 3337, 'warm': 3227, 'woven': 2559, 'fabric': 2124, 'yielding': 1908, 'rough':
 394 1889, 'slippery': 1683, 'slick': 1587, 'rubbery': 1553, 'coarse': 1504, 'lined': 1480, 'durable':
 395 1362, 'pliable': 1281, 'curved': 1240, 'bumpy': 1076, 'metallic': 970, 'patterned': 949, 'cloth-like':
 396 889, 'resilient': 785, 'abrasive': 668, 'plastic': 631, 'ridged': 599, 'gritty': 551, 'deformable': 544,
 397 'compressible': 517, 'synthetic': 444, 'fuzzy': 434, 'varnished': 430, 'dimpled': 423, 'wooden':
 398 399, 'thin': 337, 'irregular': 311, 'splotchy': 301, 'even': 267, 'uniform': 257, 'perforated': 239,
 399 'granular': 234, 'indistinct': 230, 'plastic-like': 220, 'grooved': 204, 'paper-like': 203, 'blurred':
 400 191, 'sewn': 183, 'elastic': 179, 'contoured': 173, 'shiny': 165, 'blurry': 159, 'level': 159, 'taut':
 401 149, 'grid-like': 149, 'creased': 145, 'porous': 145, 'grippy': 135, 'cushiony': 132, 'speckled': 126,
 402 'leather-like': 120, 'grained': 116, 'knitted': 107, 'padded': 99, 'worn': 94, 'round': 89, 'twisted':
 403 77, 'supple': 76, 'lightweight': 76, 'dry': 73, 'rugged': 72, 'fabric-like': 72, 'spongy': 69, 'wired':
 404 67, 'stiff': 67, 'unclear': 66, 'indented': 66, 'dense': 62, 'dark': 61, 'iridescent': 61, 'undefined':
 405 59, 'knobby': 55, 'grid-patterned': 53, 'layered': 52, 'resonant': 51, 'fluffy': 50, 'translucent': 50,
 406 'soft-focus': 49, 'absorbent': 44, 'slightly textured': 43, 'leathery': 43, 'obscured': 42, 'cylindrical':
 407 42, 'wrinkly': 41, 'unfocused': 40, 'ribbed': 39, 'rippled': 39, 'thick': 38, 'sturdy': 36, 'striated':
 408 36, 'hairy': 34, 'hazy': 33, 'embroidered': 32, 'raised': 30, 'cottony': 30, 'colorful': 29, 'slightly
 409 compressible': 29, 'straight': 28, 'silky': 28, 'braided': 28, 'straight-edged': 28, 'overexposed': 27,
 410 'angular': 27, 'ethereal': 27, 'glowing': 26, 'lettered': 25, 'tough': 25, 'edged': 25, 'rounded': 25,
 411 'transparent': 23, 'smeared': 23, 'carpeted': 23, 'stretchy': 22, 'slightly squishy': 22, 'fleshy': 21,
 412 'ceramic': 21, 'engraved': 19, 'opaque': 19, 'clothlike': 19, 'bright': 18, 'folded': 17, 'striped':
 413 17, 'embossed': 17, 'brushed': 17, 'mesh': 16, 'stable': 16, 'bendable': 16, 'slightly bendable': 16,
 414 'frayed': 15, 'printed': 15, 'vague': 14, 'cardboard': 14, 'clickable': 14, 'organic': 14, 'delicate':
 415 14, 'undulating': 14, 'clear': 13, 'stringy': 13, 'clicky': 13, 'smooth edges': 13, 'sticky': 12, 'out-of-
 416 focus': 12, 'lace': 11, 'brittle': 11, 'regular': 10, 'open': 10, 'continuous': 10, 'muted': 10, 'slightly
 417 abrasive': 10, 'malleable': 9, 'incised': 9, 'motion-blurred': 9, 'slightly warm': 9, 'intricate': 9,
 418 'obscure': 9, 'laced': 8, 'slightly curved': 8, 'compliant': 8, 'metal': 7, 'sewed': 7, 'pressed': 7,

419 'flimsy': 6, 'sandy': 6, 'insulated': 6, 'convex': 6, 'sharp': 4, 'crinkled': 4, 'springy': 3, 'complex': 3,
420 'grainy fabric': 3, 'line': 3, 'slightly gritty': 3, 'consistent': 2, 'loose': 2, 'paper': 2, 'fraying': 2,
421 'lustrous': 2, 'spotty': 2, 'light': 2, 'bristly': 2, 'woolen': 2, 'wrinkled': 2, 'griany': 2, 'precise': 2,
422 'non-glossy': 2, 'wavy': 2, 'lacey': 1, 'meshed': 1, 'imprinted': 1, 'flat smooth': 1, 'sewn fabric':
423 1, 'shadow': 1, 'bendy': 1, 'rigit': 1, 'jagged': 1, 'flash': 1, 'frabric': 1, 'patterened': 1, 'floor': 1,
424 'flawless': 1, 'long': 1, 'spolotchy': 1, 'granulated': 1, 'cloth': 1, 'thready': 1, 'patterend': 1, 'smooth
425 fabric': 1, 'deformalbe': 1, 'smmoth': 1, 'wirey': 1, 'fabric granular': 1, 'graint': 1, 'lined sewn':
426 1, 'smoth': 1, 'wiry': 1, 'torn': 1, 'vauge': 1, 'facrib': 1, 'gariny': 1, 'plain': 1, 'intertwined': 1,
427 'smoth': 1, 'stripped': 1, 'ragged': 1, 'denoisy': 1, 'slightly rough': 1, 'dull': 1, 'interwoven': 1,
428 'slightly worn': 1

429 C.4 Prompting for Psuedo-Label Generation

430 We use the following prompt with GPT-4V in order to label the images with tactile descriptions:

```
431 Surface Type: [Specify the surface type, e.g., "metal," "fabric"]  
432 2 Images: The first image is from a camera observing the tactile sensor  
433 (shiny, near the top of the image) and the surface. The second  
434 image is a cropped version of the first image that focuses on the  
435 contact patch.  
436 3 Example: For a smooth and cold surface, the description might be "  
437 slick, chilly, hard, unyielding, glossy."  
438 4 Task: Based on these images, describe the possible tactile feelings of  
439 the contact patch using sensory adjectives. Limit your response  
440 up to five adjectives, separated by commas.
```

441 C.5 Prompting GPT-4 for Evaluation

442 We use the following prompt for TVL Benchmark:

```
443 1 [User Question]: {prompt}  
444 2 [Assistant Response]: {assistant_response}  
445 3 [Correct Response]: {correct_response}  
446 4  
447 5 We would like to request your feedback on the performance of an AI  
448 assistant in response to the user question displayed above.  
449 6 The user asks the question on observing an image. The assistant's  
450 response is followed by the correct response.  
451 7  
452 8 Please evaluate the assistant's response based on how closely it  
453 matches the correct response which describes tactile feelings.  
454 Please compare only the semantics of the answers. DO NOT consider  
455 grammatical errors in scoring the assistant. The assistant  
456 receives an overall score on a scale of 1 to 10, where a higher  
457 score indicates better overall performance.  
458 9  
459 0 Please first output a single line containing only one value indicating  
460 the score for the assistant.  
461 11  
462 2 In the subsequent line, please provide a comprehensive explanation of  
463 your evaluation, avoiding any potential bias.
```

464 C.6 Improved Prompting Format

465 To investigate the effect of the prompting format, we conduct reference-guided grading for evaluation.
466 In addition, to mitigate the position bias mentioned in [33], we randomly shuffle the order of the
467 agent's response and human label on the test set. The prompt is adjusted to the following:

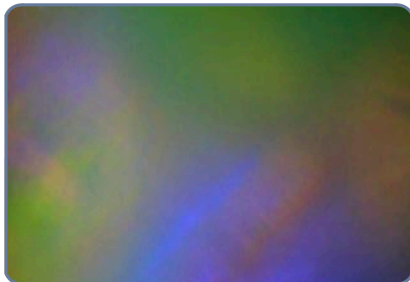
```
468 1 [User Question]: {prompt}  
469 2 {assistant_response or human_label}  
470 3 {human_label or assistant_response}  
471 4 We would like to request your feedback on the performance of an AI  
472 assistant in response to the user question displayed above.
```

473 5 The user asks the question on observing an image. The assistant’s
474 response is followed by the correct response.
475 6 Please evaluate the assistant’s response based on how closely it
476 matches the correct response which describes tactile feelings.
477 Please compare only the semantics of the answers. DO NOT consider
478 grammatical errors in scoring the assistant. The assistant
479 receives an overall score on a scale of 1 to 10, where a higher
480 score indicates better overall performance. Please first output a
481 single line containing only one value indicating the score for the
482 assistant. In the subsequent line, please provide a comprehensive
483 explanation of your evaluation, avoiding any potential bias.
484 7 Example:
485 8 [User Question]: This image gives tactile feelings of?
486 9 [Assistant Response]: fabric, grainy.
487 10 [Correct Response]: coarse, fabric, deformable.
488 11 9.5
489 12 The assistant’s response is very close to the correct response. Fabric
490 and grainy are similar to coarse and fabric. The assistant’s
491 response is missing the word deformable, which is a minor error.
492 13 Example:
493 14 [User Question]: This image gives tactile feelings of?
494 15 [Assistant Response]: flat, hard
495 16 [Correct Response]: soft, smooth, deformable
496 17 1
497 18 The assistant’s response is not close to the correct response. Hard
498 and flat are opposite to soft and smooth.

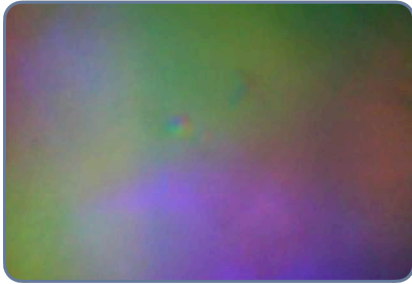
499 We tested TVL-LLaMA (ViT-B) with the reformed prompt. The score achieved by the model on
500 the prompt above is similar to the prompt mentioned in Appendix C.5 used for Table 1 (5.15 v.s.
501 5.03) with a slightly smaller p-value (1.08e-8 v.s. 3.46e-6). We encourage future works to further
502 investigate the effect of prompting on multimodal models.

503 D Generation Examples

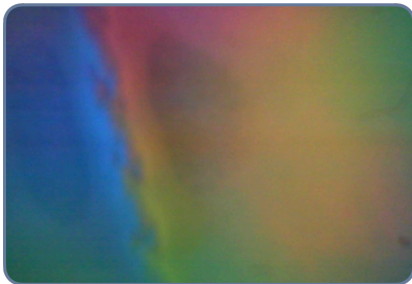
504 We provide a few positive and negative samples of image-tactile pairs from our dataset and the
505 language descriptions generated for them by our various baseline models.



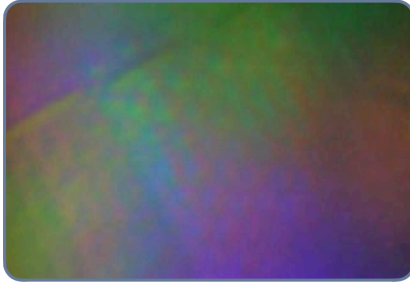
Model	Description	Score (1-10)
Human Labels	lined, fabric, sewn	GT
TVL-LLaMA	fabric, lined	7.5
GPT-4V	textured, woven, soft, flexible, durable	6
SSVTP-LLaMA	smooth, reflective, hard, cool, sleek.	1
LLaVA 1.5 7B:	Soft, grippy, textured, warm, comfortable	4
LLaVA 1.5 13B	Rough, uneven, sticky, warm, slightly damp	2
ViP-LLaVA 7B	Textured, uneven, pitted, grainy, rough###	3
ViP-LLaVA 13B	Faded, worn, creased, soft, cozy.	3
LLaMA-Adapter	1. Slick, hard, unyielding, cold, glossy.	2
BLIP-2 OPT-6.7	There is no tactile surface on this image.	1
InstructBLIP 7B	the close up picture of a blue denim jacket with button on the right side	2
InstructBLIP 13B	light blue jean jacket nothing 3	1



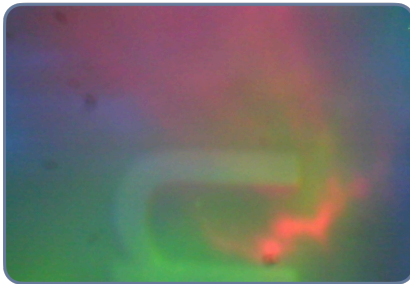
Model	Description	Score (1-10)
Human Labels	hard, flat, splotchy, grainy	GT
TVL-LLaMA	flat, hard.	6.5
GPT-4V	Textured, solid, cool, rough, uneven	4
SSVTP-LLaMA	smooth, reflective, hard, cool, sleek.	2
LLaVA 1.5 7B:	Slick, rough, cold, hard, glassy.	5
LLaVA 1.5 13B	For a white, speckled surface, the description 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



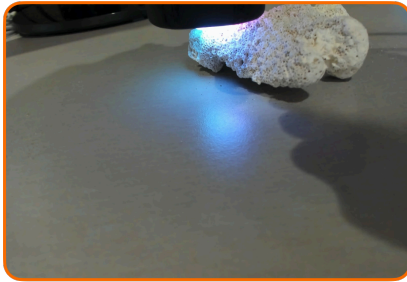
Model	Description	Score (1-10)
Human Labels	fibrous, textured, uneven, pliable, patterned	GT
TVL-LLaMA	woven, flexible, soft, glossy.	6.5
GPT-4V	smooth, reflective, hard, cool, glossy	1
SSVTP-LLaMA	smooth, reflective, hard, cool, sleek.	1
LLaVA 1.5 7B:	Cool, gritty, textured, adhesive, reflective	6.5
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



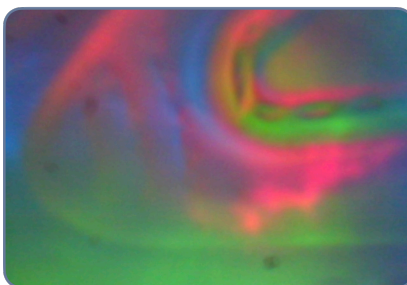
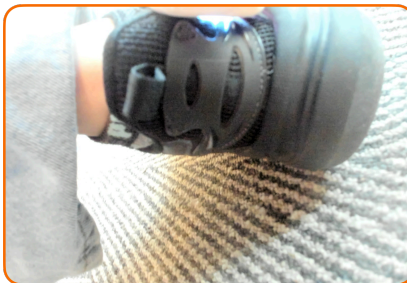
Model	Description	Score (1-10)
Human Labels	sewn, coarse, fabric, deformable	GT
TVL-LLaMA	fabric, grainy.	7.5
GPT-4V	Textured, flexible, woven, soft, uneven	7
SSVTP-LLaMA	smooth, reflective, hard, cool, sleek.	2
LLaVA 1.5 7B:	Torn, frayed, worn, stitched, black and white	6
LLaVA 1.5 13B	For the black and white striped material, the tactile sensor is making contact at its edge, creating a sense of tension or pressure, resulting in a "tight, firm, sturdy, sharp, and precise" feeling.	1
ViP-LLaVA 7B	Shiny, squishy, cloth-like, soft, textured.	4
ViP-LLaVA 13B	Textured, frayed, torn, rough, jagged.	7
LLaMA-Adapter	1. Slick, hard, unyielding, cold, glossy.	3
BLIP-2 OPT-6.7	I'm not exactly sure, but it looks like the patch has markings on it, so I guess it makes the sensor feel rough and hard	4
InstructBLIP 7B	there is a white and black striped object	2
InstructBLIP 13B	a close up of an old football uniform 1950s	1



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



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



Model	Description	Score (1-10)
Human Labels	rough, uneven, firm, patterned, ridged	GT
TVL-LLaMA	textured, flexible, woven, soft, elastic.	4
GPT-4V	Textured, firm, woven, rough, uneven	7
SSVTP-LLaMA	smooth, reflective, hard, cool, sleek.	1
LLaVA 1.5 7B:	Glossy, smooth, and cold.	1
LLaVA 1.5 13B	For the shiny patch in the image, the description would be "shiny, reflective, smooth, bright, glimmering."	2
ViP-LLaVA 7B	Rubbery, clammy, slippery, sticky, and smudged.	2
ViP-LLaVA 13B	Slick, firm, grainy, cushioned, textured	5.5
LLaMA-Adapter	1. For a rough and hot surface, the description might be "scratchy, searing, scorching, jagged, scalding." 2. For a smooth and cold surface, the description might be "slippery, icy, hard, unyielding, glassy." 3. For a rough and hot surface, the description might be "scratchy, scorching, jagged, scalding, searing."	2
BLIP-2 OPT-6.7	"cool"	4
InstructBLIP 7B	a person holding the back of a pair of shoes with a shoe lace tie.	2
InstructBLIP 13B	a person wearing a pair of shoes on a floor 300px	2