– Supplementary Material –
Seeing Through Touch: Tactile-Driven Visual Localization of Material Regions

The contents in this supplementary material are as follows:

Contents

1. Clarifying Touch Instances in Touch-and-Go

The Touch-and-Go (TG) [11] dataset consists of approximately 246k visuo-tactile image pairs and 13.9k detected touches. The official split for visuo-tactile contrastive learning is available on the official GitHub page and includes 91,982 training and 29,879 testing samples. This split excludes samples labeled as "Inconclusive" which correspond to ambiguous material classification or non-contact moments. However, the original splits of the TG dataset include samples from the same videos in both the train and test sets. This creates a risk of information leakage, as test samples may contain visual scenes and tactile data highly similar to those encountered during training. To prevent this, we construct a new split, similar to [7], ensuring no video overlap between the train and test sets. Additionally, we exclude the "Others" category. This refined version yields 91,023 training and 29,786 testing visuo-tactile frame pairs across 18 categories.

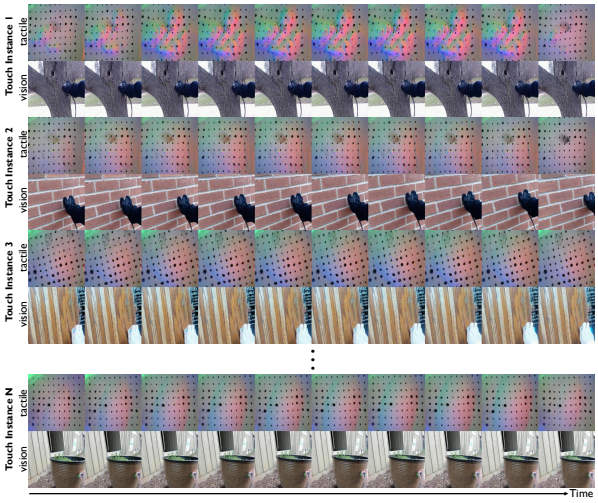

Figure 1. **Examples of Touch Instances.** Each example shows 10 frames evenly sampled from a touch instance.

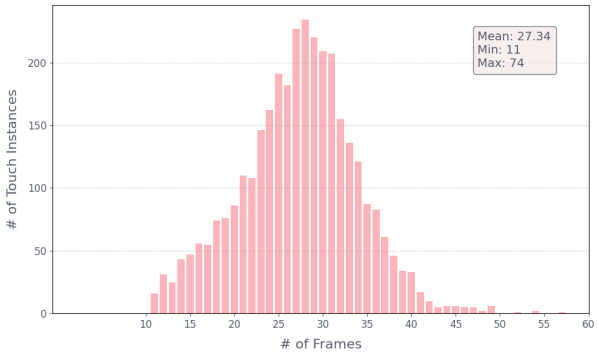

Figure 2. **Distribution of Frames per Touch Instance.**

We observe that the TG dataset contains consecutive frame sequences, each representing a single interaction in which the sensor is pressed onto and released from an object, as shown in Figure 1. We refer to each such sequence as a *Touch Instance*, as in Section 3.2 of the main paper. More technically, a Touch Instance is a temporally contiguous sequence of frames from the same video that share the same category label. The dataset provides shuffled im-

age–label pairs, where each image path encodes a video identifier and a frame index. To recover meaningful segments, we regroup samples by video ID and sort them by frame index. Consecutive frames with uninterrupted, increasing indices and identical labels are merged into a single Touch Instance. Each Touch Instance is thus defined by its frame range (*e.g.*, 332–347), its duration, and its category. The statistical distribution of Touch Instance lengths is shown in Figure 2, and the training split contains a total of 3,329 Touch Instances. As explained in Sections 3.2 and 3.3 of the main paper, we obtain our training pairs using these Touch Instances. Furthermore, we filter out overly redundant or visually overlapping neighboring Touch Instances in the test split, resulting in 579 testing instances.

2. Details on Our Web-Material Dataset

2.1. Dataset Construction

2.1.1. Image Collection

For each tactile category in the TG dataset, we prompt an LLM [2] to generate richer queries beyond simple class names to obtain descriptive context queries, which are then used to retrieve relevant and diverse web images.

Concept Query Generation. We generate concept queries using a {category} + {object} + {place} format to maximize data diversity. As this combination tends to yield wide-scene shots, we also add prompts in the format "A close-up shot of {category} + {object}" to capture close-up perspectives and improve viewpoint diversity. We use the prompt from Figure 3 with an LLM to generate our list of concept queries. For example, for the category "Brick", the LLM outputs phrases such as "brick house in a suburban neighborhood", "brick chimney in a cozy living room", and "brick bridge over a river" (as explained in Section 3.4 of the main paper).

Image Collection. Category-specific concept query management, image collection, and duplicate removal are all performed using a custom-built Gradio [1] page/tool, as shown in the top and middle panels of Figure 4.

2.1.2. Image Filtering

Since collected images may inherently contain irrelevant samples, a filtering step is essential. To reduce the workload of human annotation, we employ an automated CLIP-based filtering as a preprocessing step. Initially, we used a single positive prompt, "a photo of {category}", accepting samples where the similarity between the image and prompt embeddings exceeds a certain threshold. However, this approach proved insufficient for distinguishing subtle textural differences, often misclassifying visually similar materials such as "Brick" and "Concrete". To address this, we introduce negative prompts. We utilize an LLM to identify

ROLE
You are a highly specialized AI assistant for generating image crawling keywords. You have deep knowledge of object-scene relationships and how search engines interpret descriptive queries.

OBJECTIVE
To generate a comprehensive and diverse list of keywords for the material category {Category}, ensuring the resulting images will be varied, high-quality, and suitable for research use. Should exclude raw {Category}.

INSTRUCTIONS
1. Identify Objects: Generate a balanced set of 15–20 distinct objects characteristically made from the material {Category}. Each object itself must be unique.
2. Brainstorm Locations/Contexts: For each object, list 3 plausible and visually distinct locations or contexts. • Contexts may be reused across different objects if they naturally fit. Do not force all contexts to be unique. • Example: a Tile sink in a minimalist bathroom / in a trendy cafe / in an outdoor garden kitchen.
3. Exclusion Rule: You will be provided with an existing keyword list. Do not repeat or generate any keyword that overlaps semantically or textually with this given list. • It is acceptable to reuse the same context phrases (e.g., "in a rustic kitchen") as long as the object is new.
4. Combine and Format: Combine each object with its corresponding locations to create a final list of 60 new keywords.
Keyword Style: The keywords must be natural, descriptive, and specific. Avoid overly generic terms.

RULES
• The final output must be a single line list of keywords.
• Keywords must be separated by commas (,) in the format: keyword1, keyword2, keyword3, …
• Each keyword must follow the structure: "{object} in a {location/context}" or a similar natural phrase.
• Do not add numbers, bullet points, or line breaks.
• Do not generate keywords that carry political, cultural, or ideological bias.
• Do not include objects that are so massive or distant that the tactile material quality of {Category} cannot be visually perceived.

INPUT
• Category: {Category}.
• Existing Keywords: {…}

OUTPUT
• A single line list of new keywords where the objects are unique and the contexts may repeat naturally.

Figure 3. **Prompt for Concept Query Generation.**

easily confused materials using the prompt shown in Figure 5, while also targeting low-quality and non-real images within our negative prompts. This enhances the selectivity of the filtering, enabling us to effectively reject false positives while preserving true positives. An example of a positive prompt and negative prompts are shown in Figure 6. For each image, we compute its similarity score with a positive prompt as well as with multiple negative prompts. We then retain only the samples for which the positive prompt achieves the highest similarity score among all prompts.

Human Annotation. After CLIP-based filtering, two human annotators manually verify whether the images contain objects corresponding to the target category. The bottom panel of Figure 4 shows the Gradio interface used for this annotation task. We emphasize that this is a minimal manual process.

2.2. Dataset Distribution

We construct 32,107 training samples across 18 categories. For the test set, we annotate masks for a single target category over 675 samples. From this set, we select 100 samples and provide additional segmentation masks for a second category in each scene, forming the Interactive Localization test set with 200 masks. The distributions of the Train, Test, and Interactive Localization sets are shown in Table 1.

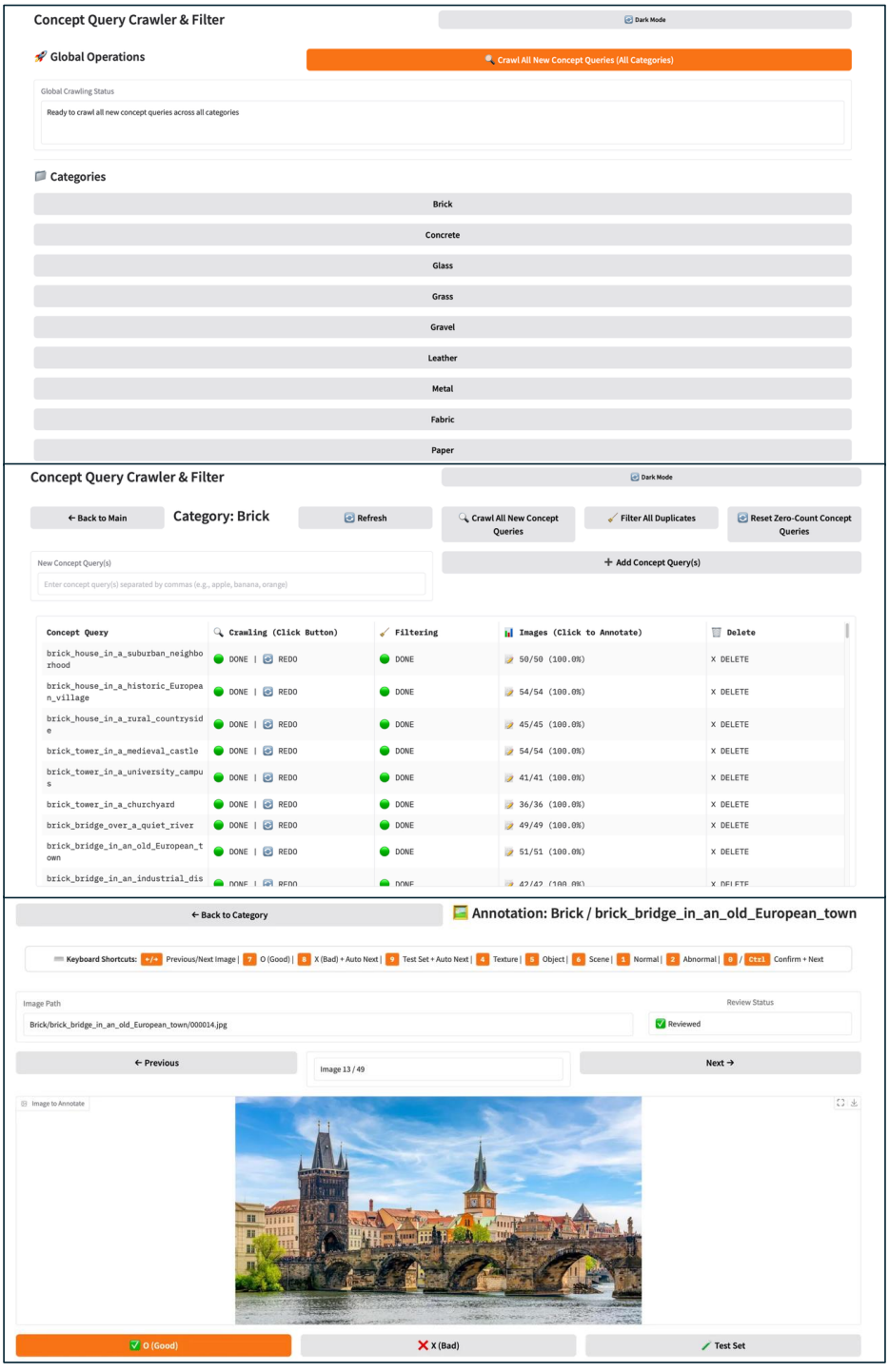

Figure 4. **Custom-built Gradio Page for Dataset Construction.**

3. Material Diversity-based Pairing

We establish three pairing strategies by leveraging two datasets: Touch-and-Go, which consists of aligned visuo-tactile data, and Web-Material, a vision-only dataset capturing diverse visual contexts for corresponding material categories. Figure 7 illustrates the data pairing strategies based on Touch Instance and Material Diversity, as proposed in Sections 3.2 and 3.3 of the main paper.

	Train			Test	IIoU Test
Category	Web Collected	MINC Curation	Sum	# of samples	# of samples
Concrete	984	-	984	41	9
Plastic	1,000	793	1,793	33	12
Glass	965	1,512	2,477	52	8
Wood	992	1,906	2,898	38	15
Metal	816	1,651	2,467	38	11
Brick	1,127	2,117	3,244	57	12
Tile	528	1,904	2,432	37	12
Leather	621	1,673	2,294	33	9
Fabric	597	1,407	2,004	36	10
Rubber	977	-	977	32	10
Paper	495	1,736	2,231	35	13
Tree	807	-	807	32	11
Grass	965	-	965	35	14
Soil	474	-	474	32	10
Rock	978	-	978	33	13
Gravel	966	-	966	39	6
Sand	1,203	-	1,203	36	17
Plants	991	1,922	2,913	36	8
Total	**15,486**	**16,621**	**32,107**	**675**	**200**

Table 1. **Category Distribution of the Train, Test, and Interactive Localization Test Sets of the Web-Material Dataset.**

Target material category: {Category}
I already use the positive prompt "a photo of {Category}".

Give me 5–10 negative prompts: things that can look similar to {Category} in photos but are made of different materials.

Requirements:
- Each item must:
  - start with "a photo of"
  - be realistic and usable for web image search
  - not contain {Category} or its synonyms
- Make the list a mix of:
  - very simple prompts, e.g. a photo of a rock
  - more detailed prompts, e.g. a photo of a transparent acrylic panel
- Output only the prompts, one per line, no numbers, no explanations.

Figure 5. **Prompt for Negative CLIP prompt generation**

```
{
    "prompts": [
        "a photo of real, high-quality glass",
        "a photo of low-quality, blurry, or fake glass",
        "a 3D render of glass",
        "a drawing of glass",
        "a photo of something else entirely",
        "a photo of clear plastic or acrylic",
        "a photo of a block of clear ice",
        "a photo of transparent epoxy resin",
        "a photo of a highly polished quartz crystal"
        ],

    "answers": [
        0
        ]
}
```

Figure 6. **Positive and Negative Prompts of "Glass" Category for CLIP filtering.**

First, **Touch Instance Pairing** pairs visual and tactile samples from the same touch instance within the Touch-and-Go dataset. Second, **In-domain Pairing** matches visual and tactile samples extracted from different touch instances belonging to the same category within the Touch-and-Go dataset. Finally, **Out-domain Pairing** combines visual images from the Web-Material dataset with tactile information from Touch-and-Go, paired based on matching categories.

4. Implementation Details

We train the model using the AdamW optimizer with $\beta = (0.9, 0.95)$, a weight decay of 0.05, and a base learning rate of 1×10^{-5}, with an effective batch size of 64. Both encoders use a pretrained DINOv3 Small model. For the first 3 epochs, we freeze both backbones and optimize only the aligners for stable adaptation. We then unfreeze the tactile backbone while keeping the image backbone frozen for the remainder of training.

Compared to the TG dataset, the Web-Material dataset exhibits substantially greater diversity in the visual appearance of target materials, particularly in object scale and spa-

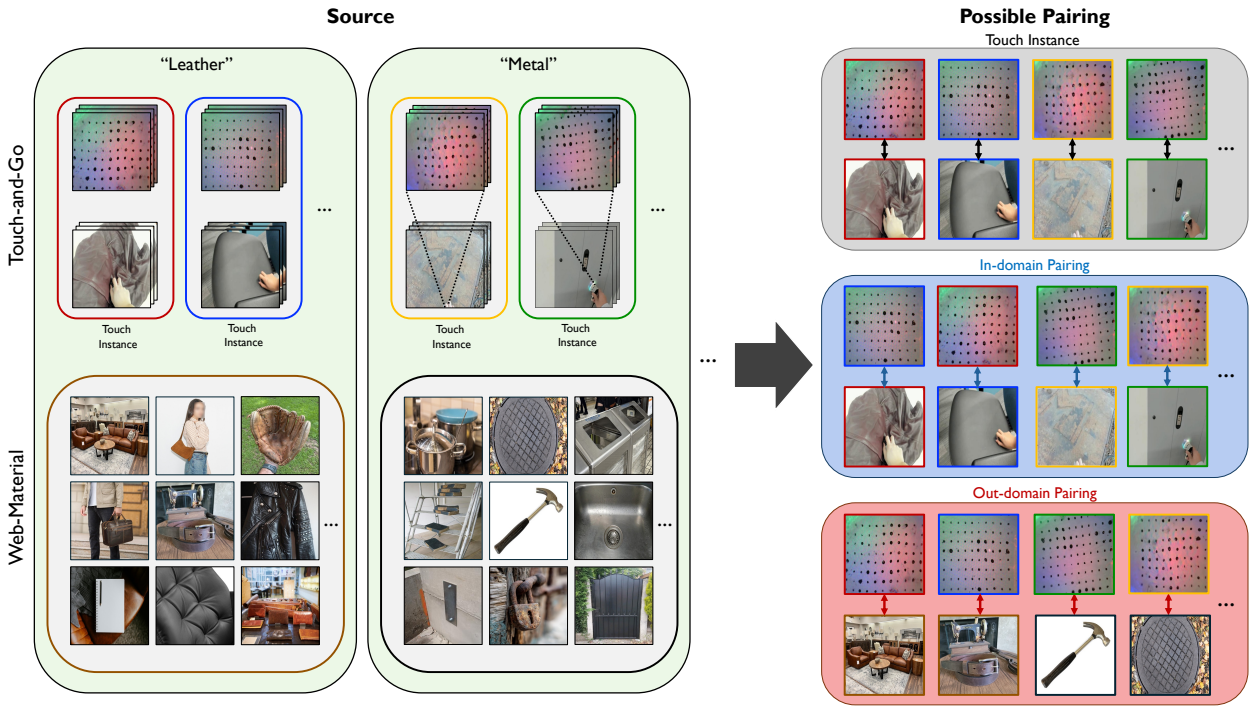

Figure 7. **Material Diversity Pairing.** Each colored borderline corresponds to a sample ID, allowing readers to see which samples are paired.

tial location. Given this complexity, learning visuo-tactile correspondence from scratch is not effective. Therefore, we adopt a curriculum learning strategy to achieve better convergence. We first establish a basic alignment by training on the TG dataset alone, then further optimize the model by introducing the more challenging out-of-domain pairs. These two stages use 100 and 50 epochs, respectively.

5. Details on Evaluation

5.1. Test Sets

Touch-and-Go Dataset. As described in Section 1, the original TG dataset undergoes several refinement steps to remove any risk of information leakage. Afterward, Touch Instance chunks are extracted, and additional filtering is performed to reduce redundancy, particularly in the test set. The final tactile localization test set contains 579 Touch Instances, and as stated in Section 4.1 of the main paper, segmentation masks are annotated for every test sample for the given category. We name this as TG-Test.

OpenSurfaces Test Set. We use a refined subset of OpenSurfaces [3] as our test set because the original dataset is unsuitable for evaluation: it provides only partial segmentation, annotating a few salient regions rather than all regions matching the ground-truth material. To construct our refined test set, we group images and masks by material category and filter out samples with negligible mask coverage. We then manually inspect all remaining pairs to ensure that the masks comprehensively cover all material instances. Finally, we verify the dataset against the curated training samples from MINC [4] to prevent data leakage, removing any overlaps. The resulting OpenSurfaces test set for tactile localization contains 211 samples across 13 categories.

Web-Material Test Set. As described in Section 2.2 of this suppl. material, this test set contains 657 samples, each with an annotation mask for its corresponding tactile category.

5.2. Computing Prototype

Because the Web-Material and OpenSurfaces test sets do not contain corresponding tactile signals and selecting a single tactile frame is non-trivial, we use a prototype tactile feature obtained by averaging the tactile features from the start, middle, and end frames. This provides a simple and fair strategy for tactile frame selection. We define the set of all categories as $C = \{c_1, c_2, \ldots, c_K\}$, where $K = |C|$. For a category $c \in C$, we define the set of its corresponding Touch Instances as $I_c = \{I_{c1}, I_{c2}, \ldots, I_{cN}\}$, where $N = |I_c|$. From a Touch Instance $I_{ci} \in I_c$, we obtain its start, middle, and end tactile frames: $t_{I_{ci},\text{start}}$, $t_{I_{ci},\text{middle}}$, and $t_{I_{ci},\text{end}}$. The prototypes for the start, middle, and end frames of category c are calculated as follows, respectively:

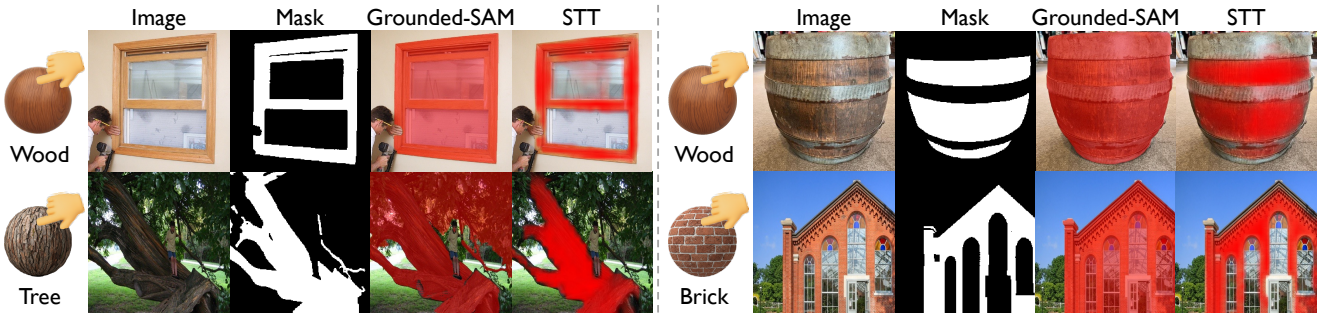

Figure 8. **Visual Comparison of Grounded-SAM and STT.**

$$P_{c,\text{start}} = \frac{1}{N}\sum_{i=1}^{N} \bar{f}_{t_{I_{ci,\text{start}}}}, \tag{1}$$

$$P_{c,\text{middle}} = \frac{1}{N}\sum_{i=1}^{N} \bar{f}_{t_{I_{ci,\text{middle}}}}, \tag{2}$$

$$P_{c,\text{end}} = \frac{1}{N}\sum_{i=1}^{N} \bar{f}_{t_{I_{ci,\text{end}}}}, \tag{3}$$

which are used in Table 2 in the main paper. The average of the start, middle, and end prototypes serves as the final prototype of category c:

$$P_c = \frac{1}{3}(P_{c,\text{start}} + P_{c,\text{middle}} + P_{c,\text{end}}), \tag{4}$$

which is utilized in Table 1, Table 3 in the main paper.

6. Comparison with Cascaded System

Model	Method	mIoU
UniTouch + Grounded-SAM	Cascaded	69.40
STT	End-to-End	76.82
Ground-Truth + Grounded-SAM	Cascaded	77.22

Table 2. **Comparison of Cascaded and End-to-End Methods.**

Table 2 compares our method with a cascaded baseline that first classifies the tactile input using UniTouch [12] and then feeds the predicted category into Grounded-SAM for visual grounding. Despite using a strong vision–language segmenter, this cascaded approach underperforms our model, demonstrating that tactile localization cannot be reduced to tactile classification followed by vision-only segmentation and must be solved as a standalone task. We further provide a qualitative analysis of the cascaded system by replacing the tactile classification output with the ground-truth material label as input to Grounded-SAM, revealing an additional limitation of this approach. As shown in Figure 8, Grounded-SAM performs primarily ***object-wise*** rather than ***material-wise*** segmentation; for objects composed of multiple materials, it often fails to separate distinct material regions even when given the correct material text prompt.

7. Material Classification on the Original Split

Model	Material (%)
VT CMC [11]	54.7
MViTac [6]	57.6
UniTouch [12]	61.3
VIT-LENS-2 [8]	63.0
TLV-Link [5]	67.2
OmniBind [9]	67.45
STT	**67.77**

Table 3. **Material Classification Linear Probing Accuracy.**

In Table 3, we report material classification results obtained by training and evaluating on the original Touch-and-Go train-test split for fair comparison. We compare our method with some of the recent tactile representation learning methods. Our method achieves the highest accuracy among them, indicating that it learns discriminative material representations while also enabling tactile localization.

8. Ablation on Tactile Backbone

Method	TG-Test		Web-Material		OpenSurfaces	
	mAP	mIoU	mAP	mIoU	mAP	mIoU
T3 (Local)	66.83	68.23	47.79	39.22	27.74	24.95
STT (Local)	85.12	76.79	67.72	52.34	37.25	29.47
T3 (Out-domain)	85.66	74.55	69.83	54.03	38.91	30.19
STT (Out-domain)	87.56	76.82	77.43	60.94	48.06	36.73

Table 4. **Ablation on Tactile Backbone.**

Since the tactile signals are captured by vision-based tactile sensors, we can initialize the tactile backbone with either tactile or vision-pretrained models. As shown in Table 4, our method outperforms the T3 [13]-initialized method in both the local and out-domain settings. This indicates that the texture and local representations acquired from DINO's large-scale pre-training are also effective for interpreting tactile frames.

9. Additional Qualitative Results

Due to space limitations in the main manuscript, we present a selected subset of qualitative results. In this supplementary material, we provide a comprehensive visualization to further demonstrate the generalization capability and robustness of our model.

Localization Results on Test Sets. Figure 9, Figure 10, and Figure 11 show additional localization results for TG-Test, Web-Material, and OpenSurfaces, respectively. These results indicate the model's consistent performance across various environments and material textures.

Interactive Localization Results. Figure 12 illustrates the interactive capabilities of our model. We visualize localization outputs for two distinct tactile signals on the same image, represented in red and green. The results demonstrate that our model can precisely distinguish and localize different material properties within a single scene based on specific tactile queries.

Real-World Scenarios. We present results on two practical use cases: (1) one-touch material-based item collection in a warehouse using 360° views, and (2) material-based robotic recycling. These results are shown in Figure 13. In both cases, our model accurately localizes the target items based on tactile input, despite distorted 360° imagery or cluttered conveyor-belt scenes. We believe these examples provide an initial indication of the real-world applicability of tactile localization in robotic systems.

Illumination Change Results. We also test our method under illumination changes. We provide qualitative results in two scenarios: (1) we adjust the color contrast of the images to test whether the model's predictions remain consistent under strong artificial variations, and (2) we examine natural illumination differences between daytime and nighttime conditions. The results, shown in Figure 14 and Figure 15, demonstrate that our method remains robust under these changes.

Material Replacement Results. Figure 16 addresses the challenge of shape bias through material replacement scenarios. We compare original images with manipulated versions where specific regions are replaced with different materials while preserving the original object shape (*e.g.*, a glass cup replaced by a plastic cup). Despite the identical geometric structure, our model accurately localizes the target regions according to their distinct tactile signals. This effectively validates that our method learns to localize based on material semantics (tactile properties) rather than relying on object shape.

Failure Cases. We analyze two representative failure cases in Figure 17 to discuss the current limitations of our method. *Transparent Objects.* Segmenting transparent objects is a challenging problem due to their optical property of transmitting background objects and textures [10]. Our model shares this limitation. As shown in our previous qualitative results, the model successfully segments glass objects or regions (relatively smaller in size) that exhibited internal refraction or specular highlights, making them easier to detect. However, as seen in the top two rows of Figure 17, the model struggles with larger highly transparent regions where the background texture is clearly visible, leading to inaccurate segmentation results.

Painted Surface. We observe another failure case when the target material is visually overlaid with high-contrast painted patterns, such as logos or symbols on concrete or asphalt surfaces. Although the underlying texture is largely preserved and implies a comparable tactile sensation, the model often suppresses these painted regions. This occurs because the visual appearance of the paint introduces strong color and edge cues that dominate the material's true surface characteristics, leading the model to treat painted areas as a separate visual category. As a result, the model localizes only the unpainted concrete regions, despite the tactile equivalence across the entire surface.

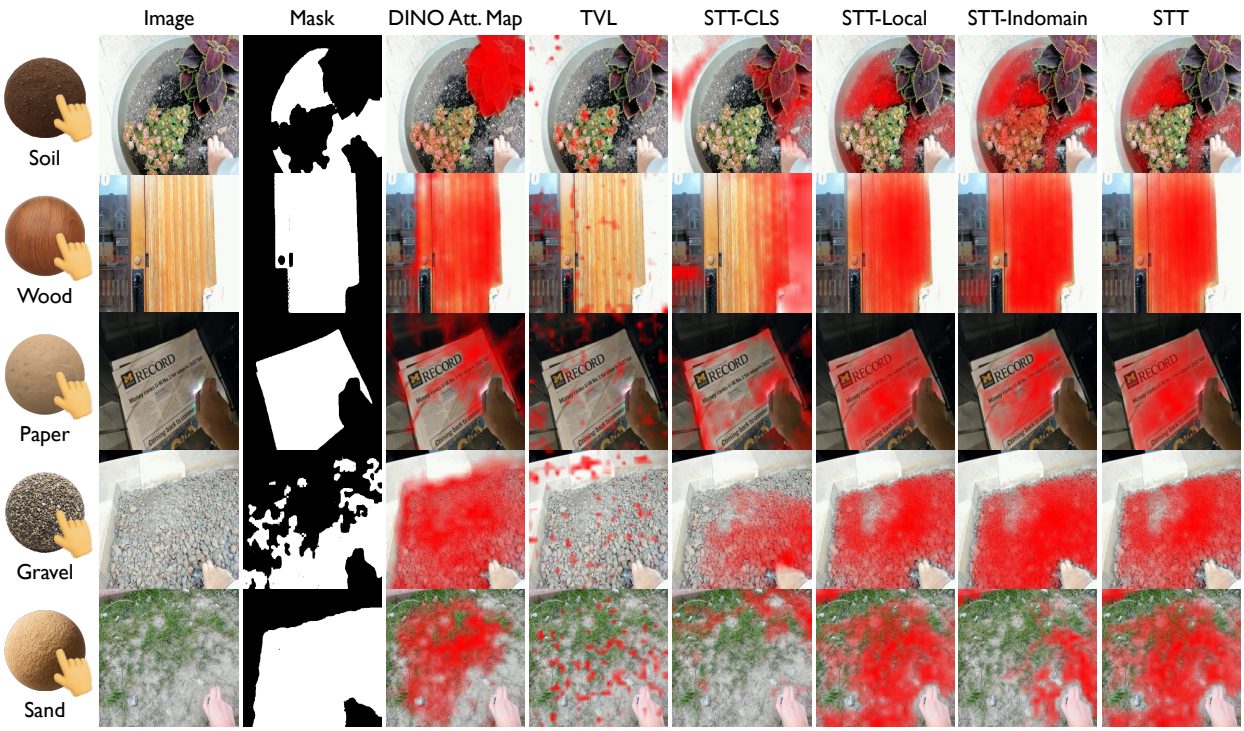

Figure 9. **Qualitative Tactile Localization Results on TG-Test.**

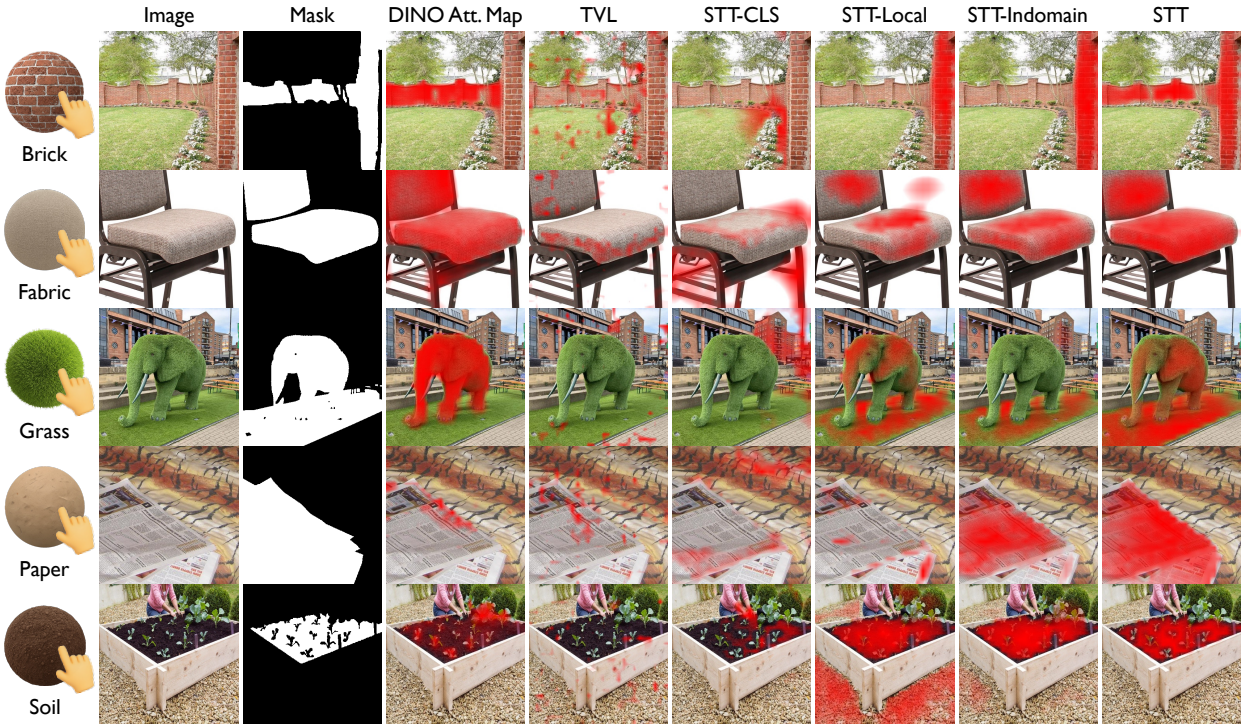

Figure 10. **Qualitative Tactile Localization Results on Web-Material.**

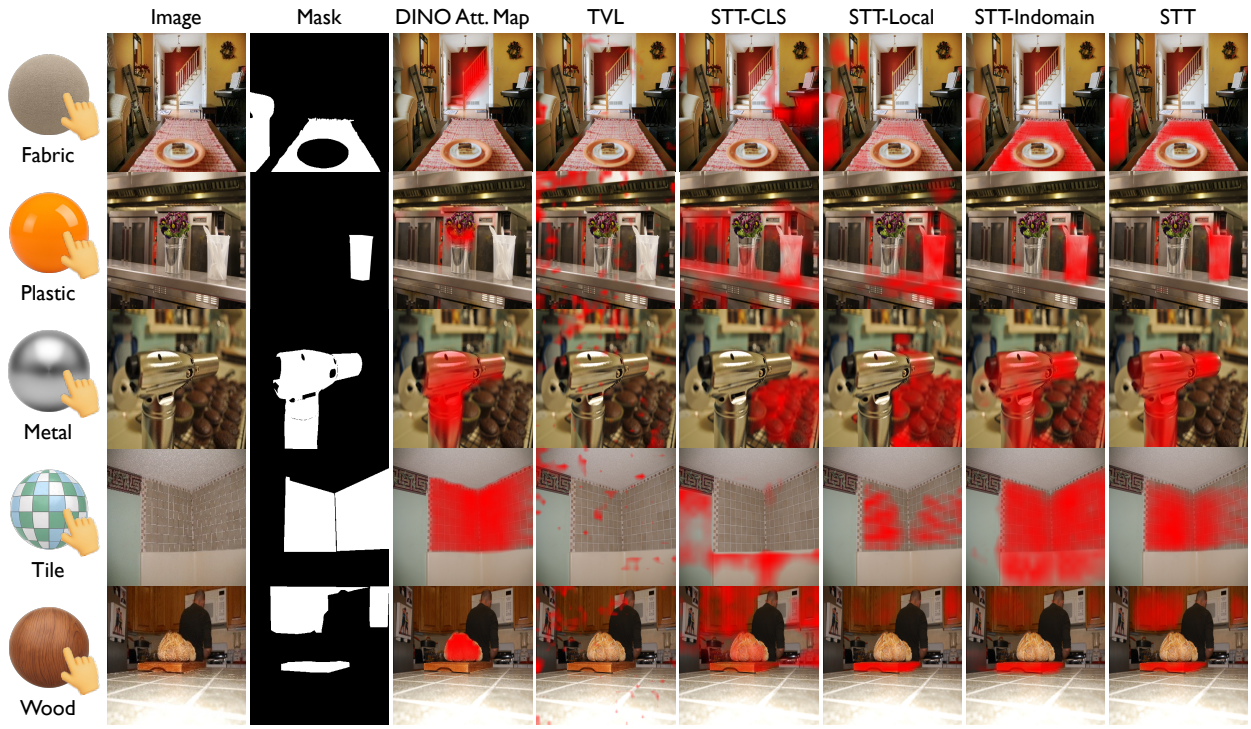

Figure 11. **Qualitative Tactile Localization Results on OpenSurfaces.**

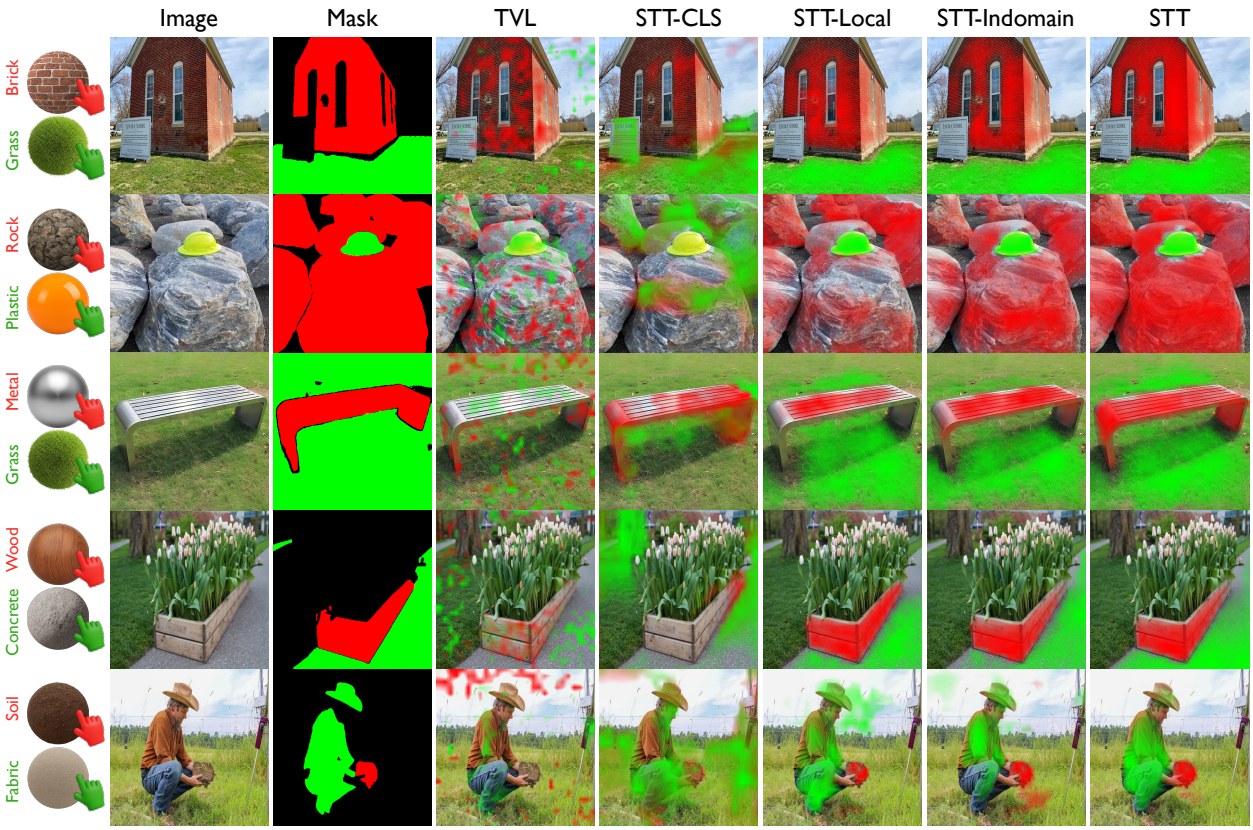

Figure 12. **Qualitative Results on Interactive Localization.**

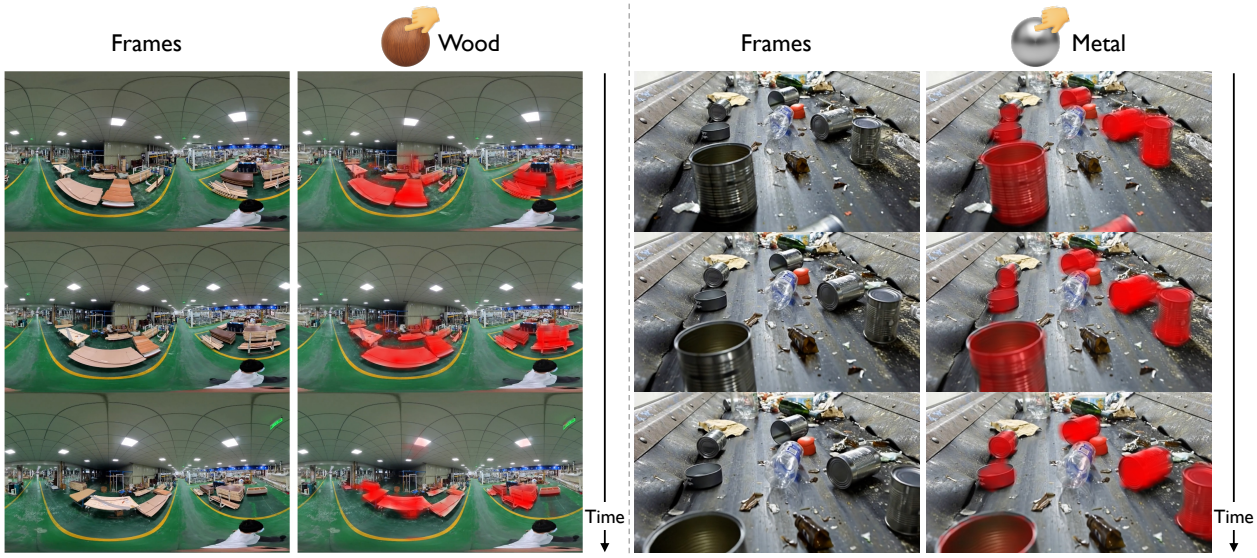

Figure 13. **Real-world Scenarios: 360° views and robot recycling.**

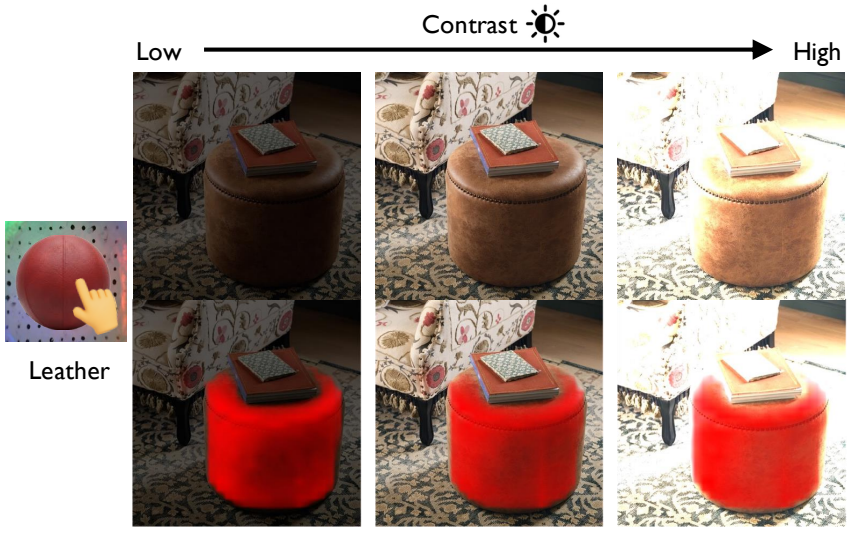

Figure 14. **Qualitative Results on Artificial Illumination Change.**

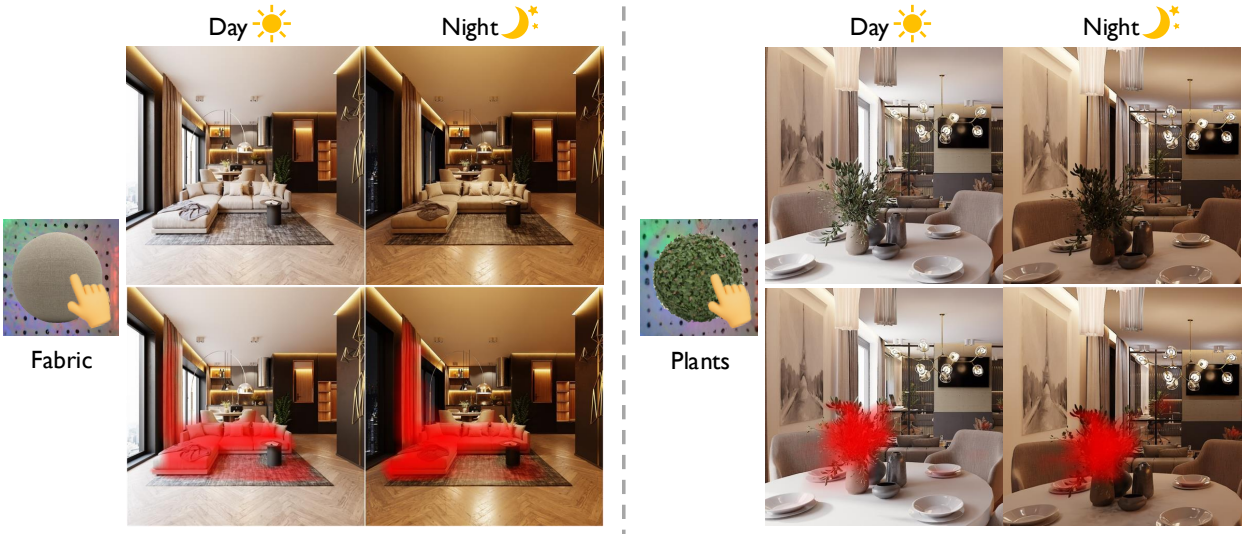

Figure 15. **Qualitative Results on Natural Illumination Change.**

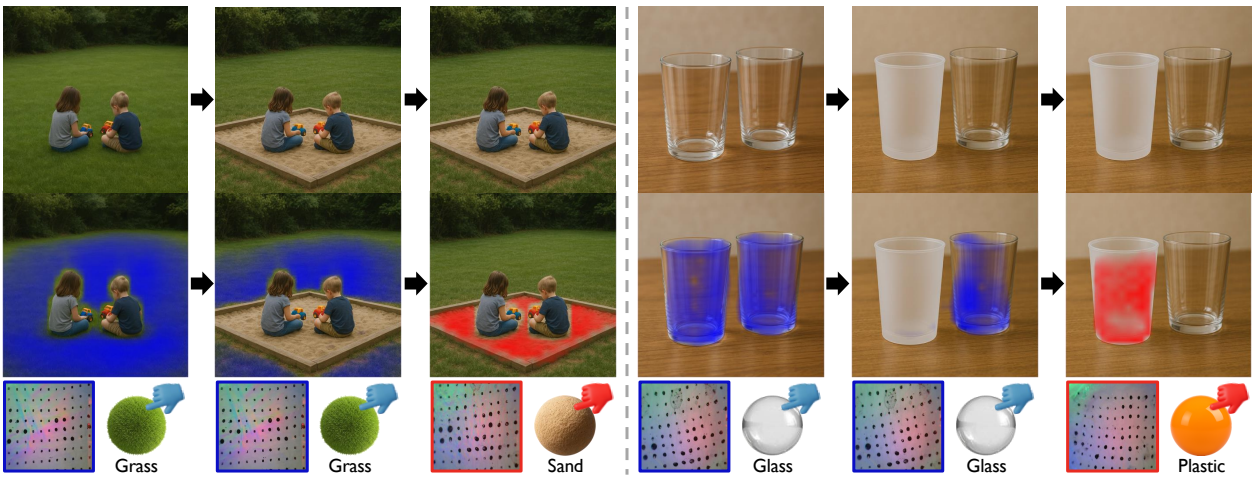

Figure 16. **Qualitative Results on Material Replacement.**

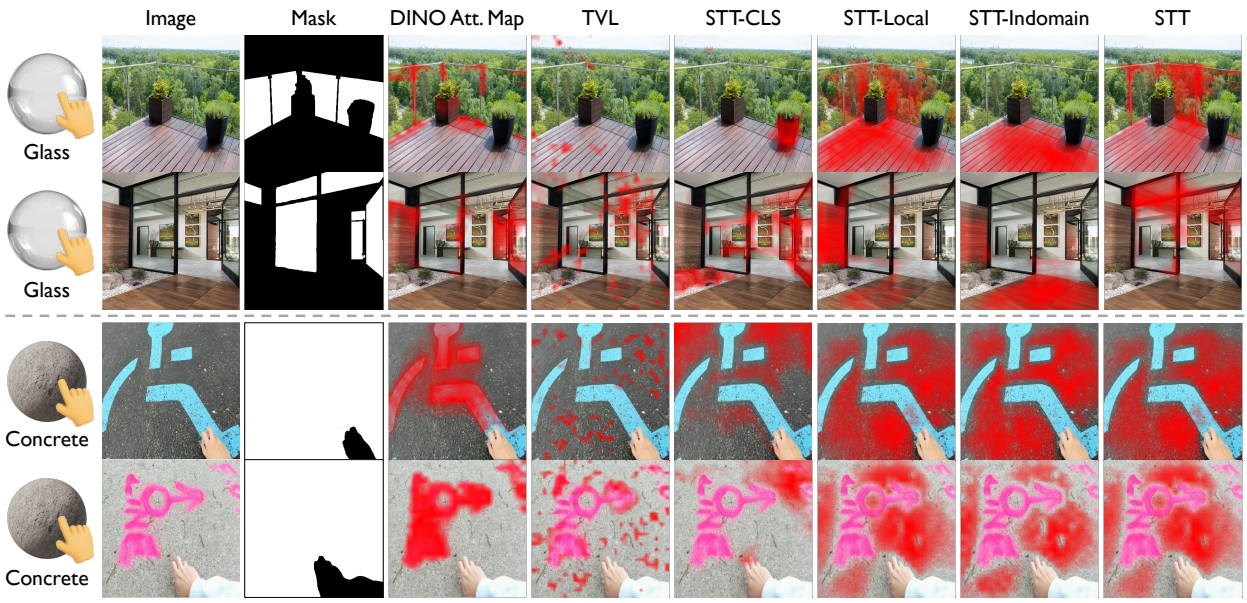

Figure 17. **Failure Cases.**

References

[1] Abubakar Abid, Ali Abdalla, Ali Abid, Dawood Khan, Abdulrahman Alfozan, and James Zou. Gradio: Hassle-free sharing and testing of ml models in the wild. *arXiv preprint arXiv:1906.02569*, 2019. 2

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2

[3] Sean Bell, Paul Upchurch, Noah Snavely, and Kavita Bala. Opensurfaces: A richly annotated catalog of surface appearance. In *SIGGRAPH*, 2013. 5

[4] Sean Bell, Paul Upchurch, Noah Snavely, and Kavita Bala. Material recognition in the wild with the materials in context database. In *CVPR*, 2015. 5

[5] Ning Cheng, Jinan Xu, Changhao Guan, Jing Gao, Weihao Wang, You Li, Fandong Meng, Jie Zhou, Bin Fang, and Wenjuan Han. Touch100k: A large-scale touch-language-vision dataset for touch-centric multimodal representation. *Information Fusion*, 2025. 6

[6] Vedant Dave, Fotios Lygerakis, and Elmar Rueckert. Multimodal visual-tactile representation learning through self-supervised contrastive pre-training. In *ICRA*, 2024. 6

[7] Cagri Gungor, Derek Eppinger, and Adriana Kovashka. Towards generalization of tactile image generation: Reference-free evaluation in a leakage-free setting. *arXiv preprint arXiv:2503.06860*, 2025. 1

[8] Weixian Lei, Yixiao Ge, Kun Yi, Jianfeng Zhang, Difei Gao, Dylan Sun, Yuying Ge, Ying Shan, and Mike Zheng Shou. Vit-lens: Towards omni-modal representations. In *CVPR*, 2024. 6

[9] Yuanhuiyi Lyu, Xu Zheng, Dahun Kim, and Lin Wang. Omnibind: Teach to build unequal-scale modality interaction for omni-bind of all. *arXiv preprint arXiv:2405.16108*, 2024. 6

[10] Enze Xie, Wenjia Wang, Wenhai Wang, Mingyu Ding, Chunhua Shen, and Ping Luo. Segmenting transparent objects in the wild. In *ECCV*, 2020. 7

[11] Fengyu Yang, Chenyang Ma, Jiacheng Zhang, Jing Zhu, Wenzhen Yuan, and Andrew Owens. Touch and go: Learning from human-collected vision and touch. In *NeurIPS - Datasets and Benchmarks Track*, 2022. 1, 6

[12] Fengyu Yang, Chao Feng, Ziyang Chen, Hyoungseob Park, Daniel Wang, Yiming Dou, Ziyao Zeng, Xien Chen, Rit Gangopadhyay, Andrew Owens, et al. Binding touch to everything: Learning unified multimodal tactile representations. In *CVPR*, 2024. 6

[13] Jialiang Zhao, Yuxiang Ma, Lirui Wang, and Edward H Adelson. Transferable tactile transformers for representation learning across diverse sensors and tasks. *arXiv preprint arXiv:2406.13640*, 2024. 6