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

Seongyu Kim[1] Seungwoo Lee[1] Hyeonggon Ryu[2] Joon Son Chung[1] Arda Senocak[3]

[1]Korea Advanced Institute of Science and Technology [2]Hankuk University of Foreign Studies
[3]Ulsan National Institute of Science and Technology

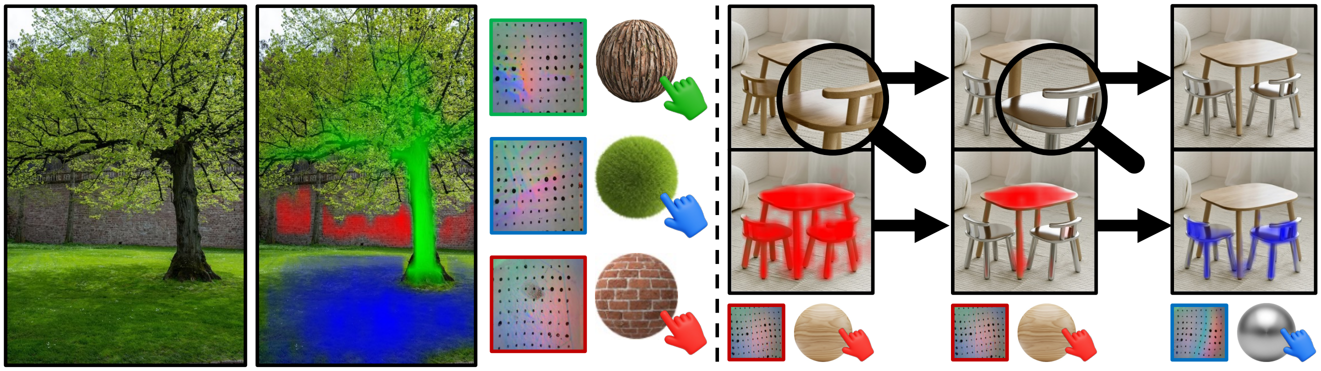

Figure 1. **Tactile Localization.** We introduce a task where the goal is to localize the visual regions that correspond to given tactile inputs. Left: The scene remains the same while different touch signals are given, and the model localizes the corresponding regions. Right: Some regions are replaced with different materials; the model should stop highlighting them when they no longer match the touch signal and highlight them again once the touch signal matches the new material.

Abstract

We address the problem of tactile localization, where the goal is to identify image regions that share the same material properties as a tactile input. Existing visuo-tactile methods rely on global alignment and thus fail to capture the fine-grained local correspondences required for this task. The challenge is amplified by existing datasets, which predominantly contain close-up, low-diversity images. We propose a model that learns local visuo-tactile alignment via dense cross-modal feature interactions, producing tactile saliency maps for touch-conditioned material segmentation. To overcome dataset constraints, we introduce: (i) in-the-wild multi-material scene images that expand visual diversity, and (ii) a material-diversity pairing strategy that aligns each tactile sample with visually varied yet tactilely consistent images, improving contextual localization and robustness to weak signals. We also construct two new tactile-grounded material segmentation datasets for quantitative evaluation. Experiments on both new and existing benchmarks show that our approach substantially outperforms prior visuo-tactile methods in tactile localization. Project page: `https://mm.kaist.ac.kr/projects/SeeingThroughTouch/`.

1. Introduction

Humans possess a natural ability to infer the tactile properties of the world around them [16]. With a single touch, we can grasp how a material feels, its softness, roughness, or texture, and immediately associate that sensation with visual cues in the environment [32]. When we touch a velvet cloth, we can easily identify other regions in a scene that would evoke the same tactile feeling, even without physically touching them. This cross-modal skill suggests that tactile perception and visual understanding are intertwined, allowing us to reason about how things feel simply by looking at them.

Inspired by humans' ability to imagine how surfaces feel from their visual appearance, we investigate whether machines can perform a similar reasoning process. To this end, we define the *tactile localization* task, where a model is given a tactile input and must identify regions in an image that share similar material properties. This task can be viewed as a material segmentation conditioned on tactile cues rather than purely visual ones [4, 30, 33], guiding the model to learn visual features that correspond to tactile properties. However, enabling such visuo-tactile reasoning poses several unique challenges.

We focus on learning local visuo-tactile representations that emerge from dense cross-modal feature interactions, a capability often missing in existing approaches. Prior work on visuo-tactile learning has predominantly focused on global alignment between modalities, using similarity between pooled representations or CLS tokens to capture coarse semantic correspondence across entire samples. While such methods can determine whether an image and a tactile input correspond to the same material, they fail to identify where in the visual scene a given tactile property exists. This limitation prevents existing models from supporting downstream tasks that require spatial reasoning, such as tactile localization. To address this, we propose a model that computes dense similarity maps between local tactile and visual features, producing tactile saliency maps that highlight image regions expected to evoke the same tactile sensation as the given touch.

In addition to architectural limitations, existing dataset constraints also make this task challenging to learn. Most existing datasets consist of close-up, texture-centric images in which nearly the entire image corresponds to a single tactile category, as the contact point or object is shown at a very close range. Additionally, the visual frames corresponding to the touch signals remain nearly identical (see Figure 4). This limited information and diversity result in only a few effective image–tactile pairs, making them insufficient for learning strong cross-modal alignment. To address these challenges, we adopt two key strategies. First, we extend the visual data with in-the-wild, open-world, scene-level images containing multiple material types. Second, leveraging the observation that similar materials evoke similar tactile sensations, we introduce a material diversity-based pairing strategy that associates one tactile sample with multiple visually diverse yet tactilely consistent images. This not only enriches the visuo-tactile correspondence space, allowing the model to learn more robust contextual tactile localization, but also provides an emergent ability to handle weak tactile signals more effectively.

Lastly, the main goal of this paper is to achieve tactile-grounded material segmentation; accordingly, we aim to evaluate our model's ability to perform this task. However, no existing dataset provides both image-tactile pairs and corresponding segmentation maps suitable for this purpose. Therefore, we constructed two new datasets: the first extends the Touch-and-Go (TG) [36] dataset by adding pixel-level material segmentations for the corresponding touch samples, and the second is a newly curated dataset of scene-level in-the-wild images collected from the web, annotated with material regions aligned to tactile categories. Evaluation on both new and existing datasets, including OpenSurfaces [3], shows that our model outperforms prior visuo-tactile methods and baselines in tactile localization.

Our main contributions are as follows:

- We propose a local visuo–tactile alignment model that produces dense tactile saliency maps to identify image regions sharing the same tactile sensation as a given touch input for material segmentation.
- We curate in-the-wild, scene-level multi-material images and propose a material-diversity pairing strategy that enriches local visuo-tactile correspondence and improves robustness to weak tactile signals.
- We construct two new tactile-grounded material segmentation datasets for evaluation and demonstrate that our model outperforms existing methods.

2. Related Work

Visuo-Tactile Representation Learning. Early work on modeling cross-modal associations between touch and vision jointly learned a shared representation by training CNNs across visual, tactile, and depth images for fabric classification [39]. More recently, this direction has shifted toward self-supervised learning [8, 10, 12, 19, 36, 37, 40]. While [19, 36] employ general multimodal contrastive learning between vision and touch, [8] extends this idea by incorporating both inter- and intra-modal relationships. Further studies [37] expand beyond touch–vision learning by connecting the tactile modality to language and sound as well, by aligning touch embeddings with image embeddings [15] already aligned with language and audio. Likewise, [12] connects touch to vision and language but adopts a pairwise approach rather than binding via images. Although these methods address visuo-tactile learning, they primarily emphasize global alignment, producing embeddings that capture coarse semantic correspondences across entire samples but lack localization ability in their local features. In contrast, our work focuses on fine-grained local alignment, where tactile sensations correspond to specific visual regions that feel similar, using contrastive learning at the local level.

Visuo-Tactile Localization. Tactile localization has been explored in various forms across the field, with differing objectives but a shared formulation in which, given a tactile signal, the goal is to identify the corresponding region in a visual scene. [19] defines this task as contact localization on garment data using a robotic arm, following pre-training with spatially aligned tactile–vision data. TaRF [10] integrates tactile sensing into neural radiance fields to learn a shared 3D representation of vision and touch, enabling spatially aligned tactile prediction within 13 reconstructed scenes and localization of contact points in 3D space. In contrast, our work focuses on tactile localization in in-the-wild 2D RGB images, where no explicit geometry or scene reconstruction is available and the domain is not restricted to garments. Instead of pinpointing a single contact location, our model segments all image regions that exhibit

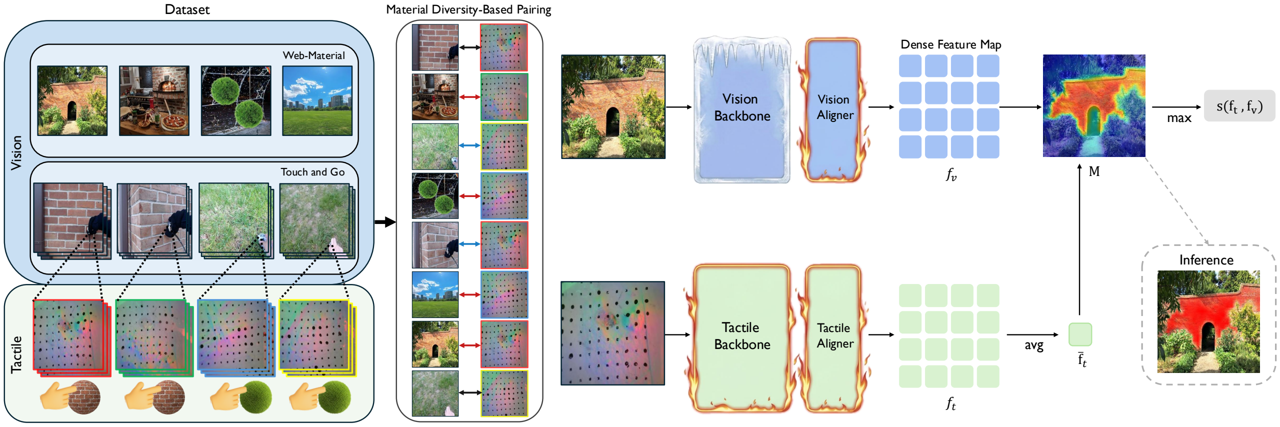

Figure 2. ***Pipeline of Seeing Through Touch.*** Tactile and visual encoders extract features from touch signals and paired images. These features are used to compute visuo-tactile similarities for contrastive learning. We extend visuo-tactile pairing by linking each tactile signal to diverse in-domain (Touch-and-Go) and out-domain (Web-Material) images of the same material category, leveraging the insight that similar materials evoke similar tactile sensations.

similar material properties to the given touch. From this perspective, our task is related to [30], which selects image regions sharing the same material with a user-selected query pixel. However, unlike [30], our method uses tactile sensory input as the query rather than a pixel within the same image, making the problem more challenging, as it requires discovering cross-modal correspondences rather than performing in-domain segmentation.

3. Methodology

Our goal is tactile localization – identifying image regions that share the same tactile sensation as a given touch input for material segmentation. We propose *Seeing Through Touch (STT)*, a framework that encodes paired tactile and visual inputs into a shared space and learns fine-grained local cross-modal alignment through contrastive learning. To enhance alignment, a material diversity-based pairing strategy leverages intra-category material variation, while additional in-the-wild web images further improve generalization. An overview is shown in Figure 2.

3.1. Preliminaries

Contrastive Learning encourages positive pairs to be close and negative pairs to be far apart. In visuo-tactile learning, let E_t and E_v be the tactile and visual encoders, respectively. Given a tactile feature $f_{t_i} = E_t(t_i)$ and its positive visual counterpart $f_{v_i} = E_v(v_i)$, with negatives f_{v_j} $(i \neq j)$ from a dataset $\mathcal{D} = \{(v_i, t_i)\}_{i=1}^{N}$, the loss is:

$$\mathcal{L}_i = -\log \frac{\exp(s(f_{t_i}, f_{v_i})/\tau)}{\sum_j \exp(s(f_{t_i}, f_{v_j})/\tau)}, \tag{1}$$

where s is a cross-modal similarity and τ is a temperature [35]. As in prior works [23, 24, 27, 28, 36, 37], this loss is applied symmetrically.

Vision and Tactile Encoders with Aligner. Given an image v_i and its paired tactile sample t_i, the encoders extract modality-specific features. Both encoders consist of a backbone network followed by a shallow aligner network. This process maps the visual and tactile inputs into a shared representation space, producing a visual feature map $f_v \in \mathbb{R}^{C\times H\times W}$ and a tactile feature map $f_t \in \mathbb{R}^{C\times H\times W}$. Here, H and W denote the spatial dimensions, and C represents the channel dimension of the shared feature space.

Similarity Function. To achieve fine-grained visuo-tactile alignment, it is essential to use a function that computes the similarity between tactile and visual features with consideration of the task. As we aim to find the regions in the image that correspond to the given tactile input, we first aggregate the tactile feature into a 1-D vector:

$$\bar{f}_t = \text{avg}_{h,w}\left(f_t[h,w]\right), \tag{2}$$

where $f_t[h,w] \in \mathbb{R}^C$ refers to the 1-D vector at location $[h,w]$ of $f_t \in \mathbb{R}^{C\times H\times W}$. We then construct a similarity map $M \in \mathbb{R}^{H\times W}$ from the aggregated tactile feature and the visual feature map:

$$M[h,w] = \bar{f}_t \cdot f_v[h,w], \tag{3}$$

where $f_v[h,w] \in \mathbb{R}^C$ denotes the 1-D vector at location $[h,w]$, and $\cdot$ represents the inner product. Thus, we obtain a similarity map between the tactile input and its corresponding image. The final similarity score is the max-pooled value of the similarity map:

$$s(f_t, f_v) = \max(M). \tag{4}$$

3.2. Training Pairs

By analyzing the popular visuo-tactile benchmark TG [36], we make several observations and use them to design our training pair strategy as follows.

Touch Instance. A touch instance refers to the action in which the collector presses and releases the sensor on an object surface, during which synchronized tactile and image frames are recorded [36]. Formally, it consists of a sequence of frames $(v_1, t_1), (v_2, t_2), \ldots, (v_T, t_T)$, where T is the number of frames in the instance. Although tactile signals vary throughout a touch instance, the corresponding visual frames remain nearly identical despite slight camera pose changes, making the visual modality temporally invariant.

Positive Pair Construction. Images and tactile signals from the same touch instance inherently correspond to the same material. Thus, any tactile frame t_i and image frame v_i from that instance can form a positive pair, as done in prior contrastive learning–based methods. However, given our observation that the visual modality is temporally invariant, positive pairs can also be constructed by randomly sampling a tactile frame t_j and an image frame v_i from the same instance, even when $i \neq j$. The similarity between tactile and visual features is then computed using Eq. 4 as $s(f_{t_j}, f_{v_i})$, which represents our training pair strategy. Moreover, this temporal invariance of the visual modality enables leveraging the fact that similar materials evoke similar tactile sensations. As a result, tactile frames can be mapped to visually similar images without requiring precise temporal correspondence, as we discuss in the next section.

3.3. Material Diversity-Based Pairing

In the visuo-tactile context, a tactile sensation is treated as a positive pair with its corresponding image, while negative pairs are sampled from other images. However, as mentioned earlier, the images within the same touch instance are highly similar, so regardless of the pairing combinations within each instance, only a few effective image–tactile pairs are obtained. This limited diversity makes the learning objective insufficient for strong cross-modal alignment. To address this, we extend contrastive learning by pairing each tactile signal with diverse images from both in-domain and out-domain samples of the same material category, thereby improving visuo-tactile alignment.

In-domain Pairing. Let our dataset consist of N touch instances, each represented as (c_n, y_n), where c_n is the n-th instance and y_n is its material category label: $\mathcal{D} = \{(c_1, y_1), (c_2, y_2), \ldots, (c_N, y_N)\}$. Each instance c_n contains a sequence of synchronized tactile and image frames: $c_n = \{(v_1^n, t_1^n), (v_2^n, t_2^n), \ldots, (v_{T_n}^n, t_{T_n}^n)\}$. Even within the same material category y_n, tactile signals can vary due to inherent structural differences between instances. To capture this diversity, we consider all instances that share the same material category and randomly sample tactile and image frames across these instances to construct positive pairs. The similarity between a tactile feature $f_{t_j}^n$ from instance c_n and an image feature $f_{v_i}^m$ from instance c_m of the same material category is computed using Eq. 4 as $s(f_{t_j}^n, f_{v_i}^m)$.

Out-domain Pairing. Let the additional image dataset be $D_{\text{out}} = \{(v_1, y_1), (v_2, y_2), \ldots, (v_l, y_l)\}$ where each image v_i is labeled with its material category y_i. Since these images do not have temporal sequences and are not naturally paired with tactile data, we construct positive pairs by sampling tactile frames t_j^n from existing instance c_n in the main dataset whose material category y_n matches y_i. The similarity function between a tactile feature and an out-domain image feature is computed using Eq. 4 as $s(f_{t_j}^n, f_{v_i})$. This formulation allows all l images in D_{out} to be leveraged for cross-modal alignment, using tactile frames drawn from the existing dataset.

3.4. Collecting Additional Images

In the Touch-and-Go dataset, the images are extremely close-up, with the material filling almost the entire scene except for the collector's hand and the sensor, as illustrated in Figure 4. Consequently, the dataset is ineffective as both a training source and an evaluation benchmark for tactile localization. To address this, we collect additional scene-level images containing multiple material types from the web and a prior material understanding dataset [4]. As described in Section 3.3, these open-world images are paired with tactile samples from the TG based on material categories, since similar materials evoke similar tactile sensations.

Image Collection. We collect images from search engines using descriptive phrases that capture diverse real-world contexts for each material. For every tactile category in the TG dataset, we prompt an LLM [1, 7] to generate richer queries beyond simple class names, which are then used to retrieve relevant and diverse web images. The LLM instruction for this process is provided in the supplementary material. For example, for the category "Brick", the LLM generates phrases such as "brick house in a suburban neighborhood", "brick chimney in a cozy living room", and "brick bridge over a river". By incorporating such phrases covering varied objects, structures, and environments, the collected images represent each material across a broad range of visual and contextual variations. Additionally, we collect images from MINC [4], a dataset of materials in the wild. Since we use tactile data from the TG dataset, 11 of its 18 categories overlap with MINC, yielding about 17K samples.

Image Filtering. After collecting images from the web, we filter out misclassified samples belonging to visually similar but incorrect categories. For each category, the LLM suggests potentially confusing alternatives, and we compute a CLIP [24] similarity score between images and category names, retaining only images whose original category has the highest similarity score. For example, for “Brick”, we use prompts such as “a photo of brick”, “a photo of concrete”, “a photo of stone”, and “a photo of rock”, keeping only images most similar to the first prompt. After this automated filtering, human annotators make minimal refinements by removing remaining unrelated images.

3.5. Implementation and Training Details

3.5.1. Architecture Details

Image encoder $E_v(\cdot)$ and Tactile encoder $E_t(\cdot)$. We adopt DINOv3 [31], pre-trained on large scale image datasets [9, 26, 34], in a self-supervised manner, for both encoders. An aligner module consisting of a Channel-wise LayerNorm [2] block and a 1x1 convolutional layer is appended to the end of each backbone. During training, we freeze the image backbone and optimize the image aligner together with the entire tactile encoder according to our training objective.

3.5.2. Implementation Details

Our model takes a single image and a tactile signal as input, each of size 224 × 224. Both image and tactile data are preprocessed with standard image augmentations. The training is conducted on 4 A5000 GPUs with an effective batch size of 64. See the supplementary material for details.

4. Experiments

4.1. Datasets

Training datasets. Our training data come from two sources: 1) **Touch-and-Go (TG)** [36]: It contains approximately 246k pairs of images and tactile signals. Although several visuo–tactile datasets exist [5, 13, 14, 19, 21], most are limited to tabletop, object-centric, or simulated settings. Since we target tactile localization in real scenes, we use TG, which was collected by human operators across diverse sub-scene environments beyond controlled setups. 2) **Our image dataset (Section 3.4)**: It consists of 32,107 web-collected images with diverse scenes and multiple material categories. Based on the training setup described in Section 3.2 and Section 3.3, these datasets are used accordingly.

Testing datasets. We evaluate localization performance using the following datasets. As no existing dataset fits this purpose, we created new benchmarks.

- **TG-Test:** We manually annotated material segmentation masks based on ground-truth tactile categories. Since this dataset already contains tactile signals, it naturally forms visuo-tactile pairs. The pairs are taken from the test split of [36], totaling 579 samples across 18 categories.
- **Web-Material:** We manually annotated material segmentation masks for web-crawled images that are completely disjoint from the training split. Tactile signals are mapped from the Touch-and-Go dataset based on category matching, yielding 675 samples covering 18 categories.
- **OpenSurfaces** [3]**:** As it already provides segmentation masks for material recognition, we use them directly and map tactile signals from TG based on overlapping categories, resulting in 211 samples across 13 categories. As these datasets lack corresponding tactile signals, we pair each image with a prototype feature computed by averaging the start, middle, and end tactile frames.

We use an online annotation tool [18] built on top of the Segment Anything Model (SAM) [20]. By selecting keypoints on images with simple mouse clicks, annotators obtain high-quality segmentation masks. They segment each region according to the given tactile categories. Example annotations are shown in Figure 3.

4.2. Baselines

We compare our methods against the baselines and prior visuo-tactile works, grouping them according to the perspectives used for analysis and comparison.

Visual Bias. As mentioned earlier, the visual images may consist of close-up shots or exhibit visual bias, as the region of interest is often single, centered, or easily identifiable without tactile information. We introduce the following baselines to assess and reveal such bias: (1) *Full Square* and (2) *Full Circle* binary masks: These involve no visual or tactile understanding and simply apply a fixed square (224×224) or circular (diameter 224) mask; (3) *DINOv3 Attention Map*: a vision-only baseline that captures visual objectness without tactile cues.

Global vs. Local Alignment. Existing visuo-tactile works primarily employ global alignment between the two modalities, whereas localization requires dense, fine-grained local alignment. The following baselines are used to validate this hypothesis and our learning objective: (1) *TVL* [12]: A recent state-of-the-art visuo-tactile method that aligns CLS tokens. We use its without-language setting for fair comparison. A variant with a frozen CLIP-Large pretrained image encoder and a ViT-Tiny tactile encoder is trained from scratch using the same data as ours. (2) *STT-CLS*: A variant of our model trained with a CLS-token alignment objective. (3) *STT*-Local: A variant of our model trained with a proposed local alignment objective only with positive pair construction of Section 3.2. (4) *STT*-Indomain: A version of *STT*-Local that incorporates In-domain material diversity-based pairing. (5) *Seeing Through Touch (STT)*: Our final model, which extends *STT*-Local by employing Out-domain material diversity-based pairing.

Model	TG-Test		Web-Material		OpenSurfaces [3]	
	mAP	mIoU	mAP	mIoU	mAP	mIoU
Binary-mask						
Full Square	-	67.25	-	32.13	-	18.13
Full Circle	-	61.75	-	34.20	-	18.19
Visual Heatmap						
DINOv3 Att. Map [31]	83.74	74.27	62.73	47.12	18.91	19.04
Global Alignment						
TVL w/o Language [12]	70.61	68.12	32.16	32.16	17.93	18.61
STT-CLS	73.63	73.49	39.35	34.74	17.98	19.07
Local Alignment						
STT-Local	85.12	76.79	67.72	52.34	37.25	29.47
STT-Indomain	86.95	77.58	71.33	55.73	42.54	34.10
STT	87.56	76.82	77.43	60.94	48.06	36.73
Upper Bound Baselines						
GroundedSAM [25]	-	77.22	-	67.03	-	50.23
Materialistic [30]	96.29	87.91	91.22	76.14	88.77	69.83

Table 1. **Tactile Localization Results on TG-Test, Web-Material, and OpenSurfaces.**

Upper Bound Baselines. We also include two upper-bound references. These baselines are not intended for direct comparison, but serve as reference points to indicate how far localization performance could reach under more favorable conditions: (1) *GroundedSAM* [25]: A large-scale vision–language segmentation model that uses text prompts for segmentation; in our experiments, the tactile category name is used as a prompt. (2) *Materialistic* [30]: A material segmentation model that uses a user-clicked visual prompt; in our setup, a pixel from the ground-truth segmented area is provided as the prompt.

4.3. Main Results

4.3.1. Comparison with Prior Works and Baselines

We evaluate our method against prior works and strong baselines on three test sets, TG-Test, Web-Material, and OpenSurfaces, defined in Section 4.1. We use mAP and mIoU as evaluation metrics following the standard multi-modal grounding protocols [6, 11, 17, 22, 27]. The results are presented in Table 1. Our model consistently outperforms prior works and relevant baselines. **Key findings** are as follows:

1. Local visuo-tactile alignment is essential for tactile localization. Our results show that global alignment is not a suitable objective for learning tactile localization. All variants of our local alignment objective outperform both TVL and *STT-CLS* across all test sets by a large margin, as these methods rely on global alignment. This validates the motivation of our work and shows the necessity of the proposed method over existing visuo-tactile approaches.

2. Material diversity-based pairing effectively improves visuo-tactile alignment. We observe that applying material diversity-based pairing consistently yields clear gains over *STT-Local*, both in the in-domain and out-domain settings. Incorporating additional in-the-wild images further improves performance, with *+5.21* mIoU on Web-Material and *+2.63* mIoU on the OpenSurfaces benchmark, suggesting enhanced semantic alignment between the two modalities through exposure to greater visual diversity. These results support our hypothesis that the limited diversity of existing datasets restricts cross-modal alignment, and that leveraging the insight that similar materials evoke similar tactile sensations offers a straightforward yet effective solution. Moreover, diversity-based pairing enables better utilization of limited and expensive tactile data, as even in-domain pairing leads to noticeable improvements.

3. Tactile localization requires true visuo-tactile alignment. The gap between the DINOv3 Attention Map and any variant of our local visuo-tactile alignment shows that this task requires true cross-modal understanding between the two modalities, as the difference between our method and the visual heatmap baseline remains substantial, except on the TG-Test dataset, which will be discussed later.

4. Upper-bound baselines indicate that achieving highly accurate localization still remains a challenge. As discussed in Section 4.2, these baselines are not meant for direct comparison but serve as reference points to show how far localization performance could reach under more favorable conditions. GroundedSAM reflects the level of performance achievable when substantial explicit tactile understanding is assumed, while Materialistic estimates an upper bound under ideal cross-modal correspondence, where visual and tactile spaces are perfectly aligned. This observation highlights a new challenge for the community.

5. TG is a limited benchmark for tactile localization. TG is a popular visuo-tactile dataset with synchronized tactile–image pairs. As part of our benchmark construction, we annotated TG to create TG-Test. However, our results reveal some dataset bias. Full Square and Full Circle baselines, which perform no meaningful reasoning, already achieve *67.25* and *61.17* mIoU on TG-Test, respectively. Global alignment methods, unsuitable for localization, reach *68.12* and *73.49* mIoU, and the DINOv3 attention map obtains *74.27* mIoU *without tactile input*. These results indicate that TG-Test inflates localization performance even in the absence of visuo-tactile reasoning. This aligns with our earlier observation that TG consists of close-up, texture-centric images where nearly the entire frame corresponds to a single tactile category. In contrast, the same baselines perform substantially worse on Web-Material and OpenSurfaces, and the performance gap between our method and others is more prominent, suggesting that these datasets provide more reliable benchmarks for evaluating visuo-tactile localization.

Qualitative Results. Figure 3 compares our method and its variant with the visuo-tactile baseline TVL and DINOv3 attention map. Consistent with the results in Table 1, our models accurately localize the tactile signal, whereas TVL struggles due to its global-alignment objective and DINOv3 attention highlights only visually salient objects rather than

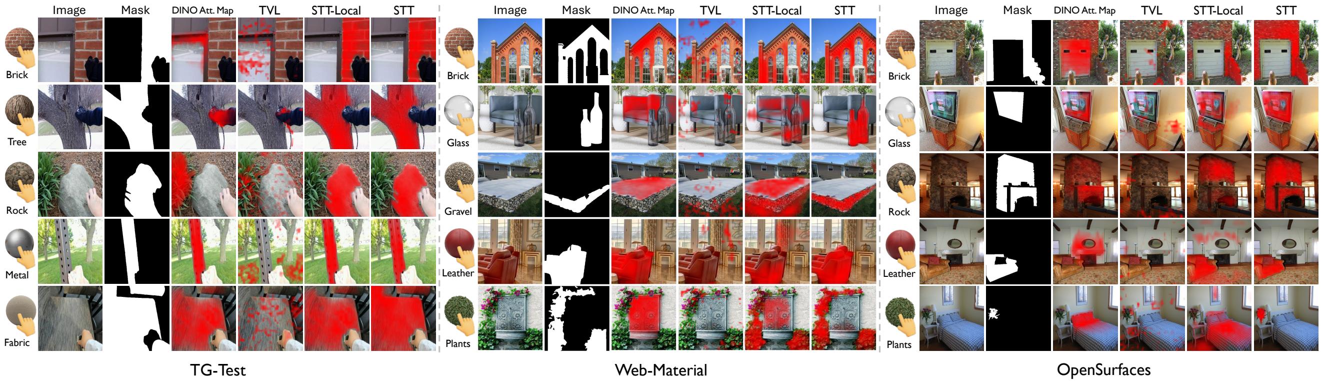

Figure 3. **Qualitative Tactile Localization Results.** Our model localizes more accurately than prior works and baselines across all benchmarks.

the true tactile correspondence. The material-diversity pairing strategy further improves localization quality. Overall, our approach successfully localizes a wide range of materials and objects, including small regions such as the 'Plants' example in the last row of OpenSurfaces.

4.3.2. Robustness to Weaker Tactile Signals

During tactile data collection, the sensor [38] is gradually pressed onto and released from the surface, so tactile signals at the beginning and end of a touch instance are typically weaker than those in the middle, where contact is firm. To analyze the encoder's ability to capture such weak signals, we evaluate our method using three types of tactile frames, Start, Middle, and End as shown in Table 2. Start and End correspond to the initial and final moments of a tactile sequence, while Middle refers to the frames in between. Each visual image is paired with one of these tactile signals, and localization is performed accordingly. This experiment is conducted only on our model variants, as earlier sections already show that our method achieves the best performance for this task. As shown in Table 2, our method without material diversity-based pairing exhibits a clear performance drop on Start and End frames compared to Middle frames, highlighting the challenge posed by weaker tactile inputs. Applying material diversity-based pairing significantly mitigates this gap, with further improvement when incorporating out-domain in-the-wild images. The performance of weaker signals becomes closer to that of Middle-frame inputs, indicating that material diversity-based pairing effectively compensates for faint tactile cues. This is especially important given the limited and costly nature of tactile data, allowing the model to use all available signals more efficiently without discarding them. We also present a qualitative example from the Touch-and-Go dataset in Figure 4, showing how tactile signals change over time within a single touch instance, where the start and end are relatively weaker. This example illustrates both the temporal variation of tactile signals and the robustness of our model with material diversity-based pairing to weaker tactile inputs.

Model	**M.D.P.**	Start		Middle		End	
		mAP	mIoU	mAP	mIoU	mAP	mIoU
			TG-Test				
STT-Local	✗	81.31	72.67	85.12	76.79	81.96	72.68
STT-Indomain	In-domain	86.34	76.15	86.95	77.58	85.51	74.60
STT	Out-domain	86.20	74.56	87.56	76.82	84.54	73.57
			Web-Material				
STT-Local	✗	64.45	49.45	69.60	54.69	61.52	48.72
STT-Indomain	In-domain	69.45	53.83	71.14	56.33	67.31	53.24
STT	Out-domain	76.19	59.99	78.98	62.08	75.08	58.56
			OpenSurfaces [3]				
STT-Local	✗	33.63	26.77	40.14	32.54	35.86	28.60
STT-Indomain	In-domain	40.73	32.24	44.39	35.11	39.08	31.45
STT	Out-domain	45.33	35.00	55.06	42.12	44.57	34.20

Table 2. **Robustness to Weaker Tactile Signals.** We evaluate our local alignment variants using three types of tactile frames: Start, Middle, and End. M.D.P. denotes material diversity-based pairing.

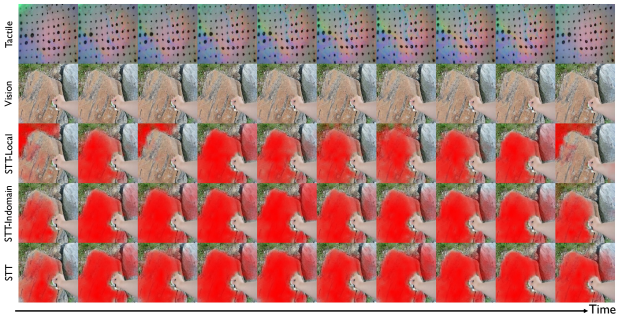

Figure 4. **Touch-and-Go dataset examples.** The tactile signal is typically weaker at the beginning and end of a touch instance while strongest in the middle. Models with material diversity-based pairing achieve robust localization regardless of variations in signal strength.

4.3.3. Interactive Localization

To further analyze fine-grained visuo-tactile alignment, we evaluate the models from an interactive localization perspective following [29]. A reliable visuo-tactile localization method should accurately associate tactile inputs with their corresponding materials, meaning the localized region in the image should change when paired with a different tactile signal from the scene.

Baselines				
Model	TVL [12]	DINOv3 Att. Map [31]	GroundedSAM [25]	Materialistic [30]
IIoU	1.0	18.0	49.0	83.0
Ours				
Model	*STT*-CLS	*STT*-Local	*STT*-Indomain	*STT*
IIoU	4.0	30.0	32.0	37.0

Table 3. **Quantitative Results on Interactive Localization.** *STT* outperforms other methods and shows reliable interactive localization ability.

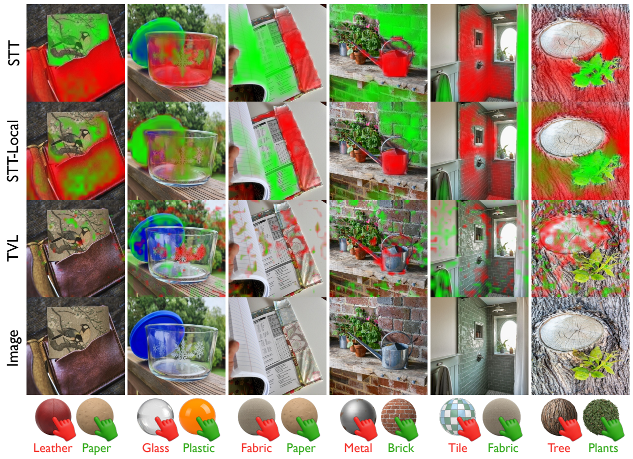

Figure 5. **Qualitative Results on Interactive Localization.** Our model accurately localizes the objects corresponding to the given tactile inputs and shifts the localized region appropriately when the tactile signal is changed.

Implementation. For interactive localization, we annotate each image in the Web-Material set with segmentation masks corresponding to the two tactile regions in the scene, forming the Web-Material-Interactive dataset. The model then predicts a separate localization map for each tactile signal. A sample is considered successful if the IoU for both tactile regions exceeds 0.5, ensuring that the model can reliably localize each tactile signal in the scene.

Quantitative Results. Table 3 reports interactive IoU (IIoU) on the Web-Material-Interactive test set. These results clearly indicate that our local-alignment objective is essential for interactive tactile localization. Notably, global-alignment methods almost fail in the interactive setting, and their already limited performance in single tactile localization drops drastically: TVL and *STT*-CLS fall from 32.16 and 34.74 mIoU to only 1.0 and 4.0 IIoU when distinguishing multiple tactile signals in the same scene. Moreover, the results show that material diversity-based pairing, whether with in-domain or better with out-domain images, is another key factor for accurate visuo-tactile association. Overall, our approach reliably captures interactive visuo-tactile relationships while maintaining strong standard localization performance. As discussed earlier, the other baselines in the table serve as upper-bound references or illustrate the characteristics of the task when using only the visual modality.

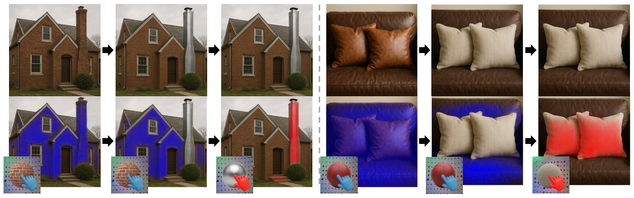

Figure 6. **Qualitative Results on Material Replacement.** The model interactively and consistently updates localization in response to changes in material and corresponding tactile inputs.

Qualitative Results. Figure 5 demonstrates the interactive localization ability of our method and its variant. Accurate tactile localization should identify the material regions corresponding to a given touch signal. Compared to TVL, our model reliably highlights different regions in the same scene depending on the tactile input, while the competing method fails to do so. The material-diversity pairing variant further improves localization accuracy. In the 5th column, our model precisely localizes the towel and tiles based on their respective tactile cues. Overall, our model not only handles interactive localization but does so with high precision.

Unlike the previous interactive localization setup, where the scene remains identical while the tactile signal changes, here we consider a second scenario: the scene is mostly unchanged, but some regions are replaced with different materials. In this case, regions that were previously highlighted should no longer be activated when their material no longer matches the touch signal. However, if the touch signal is updated to match the new material, the model should highlight the replaced region again. We show this scenario by editing images with off-the-shelf image editing tool [7] and pairing them with the corresponding tactile signals. Visualizations in Figure 1 and Figure 6 show that our method localizes the touched material in an interactive and consistent manner.

5. Conclusion and Discussion

In this paper, we introduce a framework for tactile localization that learns fine-grained alignment between tactile signals and visual scenes. By leveraging dense local cross-modal feature interactions, in-the-wild multi-material images, and a material diversity-based pairing strategy, our approach overcomes the limitations of existing visuo-tactile methods that employ global-alignment objectives, as well as the constraints of current visuo-tactile datasets. Through extensive evaluation on both new and established benchmarks, we demonstrate significant improvements in touch-conditioned material segmentation and robust localization, even under weak tactile inputs. Our results highlight the importance of local visuo-tactile alignment and dataset diversity for grounding tactile perception in images. We hope this work provides a foundation for future efforts in visuo-tactile reasoning, interactive perception, and multisensory scene understanding.

6. Acknowledgment

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2025-02215122, Development and Demonstration of Lightweight AI Model for Smart Homes (60%); RS-2020-II201336, Artificial Intelligence Graduate School Program (UNIST) (10%)), and by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2026-25496684) (30%).

References

[1] 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. 4

[2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 5

[3] Sean Bell, Paul Upchurch, Noah Snavely, and Kavita Bala. Opensurfaces: A richly annotated catalog of surface appearance. In *SIGGRAPH*, 2013. 2, 5, 6, 7

[4] Sean Bell, Paul Upchurch, Noah Snavely, and Kavita Bala. Material recognition in the wild with the materials in context database. In *CVPR*, 2015. 1, 4

[5] Roberto Calandra, Andrew Owens, Manu Upadhyaya, Wenzhen Yuan, Justin Lin, Edward H Adelson, and Sergey Levine. The feeling of success: Does touch sensing help predict grasp outcomes? *arXiv preprint arXiv:1710.05512*, 2017. 5

[6] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022. 6

[7] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 4, 8

[8] Vedant Dave, Fotios Lygerakis, and Elmar Rueckert. Multimodal visual-tactile representation learning through self-supervised contrastive pre-training. In *ICRA*, 2024. 2

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 5

[10] Yiming Dou, Fengyu Yang, Yi Liu, Antonio Loquercio, and Andrew Owens. Tactile-augmented radiance fields. In *CVPR*, 2024. 2

[11] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *IJCV*, 2015. 6

[12] Letian Fu, Gaurav Datta, Huang Huang, William Chung-Ho Panitch, Jaimyn Drake, Joseph Ortiz, Mustafa Mukadam, Mike Lambeta, Roberto Calandra, and Ken Goldberg. A touch, vision, and language dataset for multimodal alignment. In *ICML*, 2024. 2, 5, 6, 8

[13] Ruohan Gao, Zilin Si, Yen-Yu Chang, Samuel Clarke, Jeannette Bohg, Li Fei-Fei, Wenzhen Yuan, and Jiajun Wu. Objectfolder 2.0: A multisensory object dataset for sim2real transfer. In *CVPR*, 2022. 5

[14] Ruohan Gao, Yiming Dou, Hao Li, Tanmay Agarwal, Jeannette Bohg, Yunzhu Li, Li Fei-Fei, and Jiajun Wu. The objectfolder benchmark: Multisensory learning with neural and real objects. In *CVPR*, 2023. 5

[15] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *CVPR*, 2023. 2

[16] Michael SA Graziano and Charles G Gross. The representation of extrapersonal space: A possible role for bimodal, visual-tactile neurons. 1995. 1

[17] Mark Hamilton, Andrew Zisserman, John R Hershey, and William T Freeman. Separating the" chirp" from the" chat": Self-supervised visual grounding of sound and language. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13117–13127, 2024. 6

[18] Intel Corporation. Cvat: Computer vision annotation tool. https://www.cvat.ai/, 2025. 5

[19] Justin Kerr, Huang Huang, Albert Wilcox, Ryan Hoque, Jeffrey Ichnowski, Roberto Calandra, and Ken Goldberg. Self-supervised visuo-tactile pretraining to locate and follow garment features. *arXiv preprint arXiv:2209.13042*, 2022. 2, 5

[20] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, 2023. 5

[21] Yunzhu Li, Jun-Yan Zhu, Russ Tedrake, and Antonio Torralba. Connecting touch and vision via cross-modal prediction. In *CVPR*, 2019. 5

[22] Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *CVPR*, 2022. 6

[23] Shentong Mo and Pedro Morgado. Localizing visual sounds the easy way. In *European Conference on Computer Vision*, pages 218–234. Springer, 2022. 3

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 3, 5

[25] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024. 6, 8

[26] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015. 5

[27] Hyeonggon Ryu, Seongyu Kim, Joon Son Chung, and Arda Senocak. Seeing speech and sound: Distinguishing and locating audio sources in visual scenes. In *CVPR*, 2025. 3, 6

[28] Arda Senocak, Hyeonggon Ryu, Junsik Kim, Tae-Hyun Oh, Hanspeter Pfister, and Joon Son Chung. Sound source localization is all about cross-modal alignment. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7777–7787, 2023. 3

[29] Arda Senocak, Hyeonggon Ryu, Junsik Kim, Tae-Hyun Oh, Hanspeter Pfister, and Joon Son Chung. Toward interactive sound source localization: Better align sight and sound! *IEEE TPAMI*, 2025. 7

[30] Prafull Sharma, Julien Philip, Michaël Gharbi, Bill Freeman, Fredo Durand, and Valentin Deschaintre. Materialistic: Selecting similar materials in images. In *SIGGRAPH*, 2023. 1, 3, 6, 8

[31] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025. 5, 6, 8

[32] Linda Smith and Michael Gasser. The development of embodied cognition: Six lessons from babies. *Artificial life*, 2005. 1

[33] Paul Upchurch and Ransen Niu. A dense material segmentation dataset for indoor and outdoor scene parsing. In *ECCV*, 2022. 1

[34] Frederik Warburg, Soren Hauberg, Manuel Lopez-Antequera, Pau Gargallo, Yubin Kuang, and Javier Civera. Mapillary street-level sequences: A dataset for lifelong place recognition. In *CVPR*, 2020. 5

[35] Zhirong Wu, Yuanjun Xiong, Stella Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance-level discrimination. In *CVPR*, 2018. 3

[36] 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. 2, 3, 4, 5

[37] 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. 2, 3

[38] Wenzhen Yuan, Siyuan Dong, and Edward H Adelson. Gelsight: High-resolution robot tactile sensors for estimating geometry and force. *Sensors*, 2017. 7

[39] Wenzhen Yuan, Shaoxiong Wang, Siyuan Dong, and Edward Adelson. Connecting look and feel: Associating the visual and tactile properties of physical materials. In *CVPR*, 2017. 2

[40] Martina Zambelli, Yusuf Aytar, Francesco Visin, Yuxiang Zhou, and Raia Hadsell. Learning rich touch representations through cross-modal self-supervision. In *Conference on Robot Learning*, 2021. 2