DATE: DUAL-SCALE ADAPTIVE TEMPORAL ENCODING FOR TACTILE PROPERTY PERCEPTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-based tactile sensing enables robots to infer physical properties from contact interactions. Yet the cues needed for different properties unfold at different temporal scales. Hardness is reflected in deformation accumulated during pressing, whereas roughness is revealed by local texture and shear changes during sliding. Models that rely on a single global summary or a fixed short window can miss complementary evidence needed for these properties. To address this, we propose Dual-scale Adaptive Temporal Encoding (**DATE**), which jointly models local contact changes and full-sequence dynamics. A frozen tactile backbone provides patch and frame-level features. A local branch applies spatiotemporal attention over patch tokens within short windows to preserve texture and shear transitions, while a global branch encodes the deformation history across the observed sequence. Property-specific gated fusion combines these temporal representations with statistical cues, allowing hardness and roughness to draw on different temporal evidence across pressing, sliding, and twisting. We further integrate the learned tactile representation with a frozen vision-language model for material description and property reasoning. Evaluations on two benchmarks with graded property labels demonstrate the effectiveness of **DATE**. On VitaSet, it improves hardness and roughness classification accuracy over AnyTouch by 7.38 and 5.14 percentage points, respectively. The downstream extension achieves an overall TVL description score of 5.66, exceeding TVL-LLaMA (ViT-Base) by 0.63 points.

Index Terms— Vision–tactile fusion, Temporal feature modeling, Material property recognition, Hardness and roughness classification, Multimodal perception

1 INTRODUCTION

Conventional robot manipulation relies on external vision to recognize objects, estimate geometry, and plan grasps (Gao et al., 2024). Vision supports localization and grasp planning, but it cannot reveal intrinsic physical properties of objects: objects of similar appearance may differ in hardness and surface roughness (Yuan et al., 2017b), which call for different force-aware contact strategies. Tactile sensing provides this complementary information. Among tactile modalities such as force or pressure, magnetic and optical tracking (Kappassov et al., 2015), vision-based sensors like GelSight (Yuan et al., 2017a) and DIGIT (Lambeta et al., 2020) capture the deformation of a compliant surface at high spatial resolution, and their image sequences encode contact geometry, local texture, and deformation over time. We target hardness and roughness because they rely on different temporal cues: hardness emerges from deformation accumulated during pressing (Yuan et al., 2017c), whereas roughness appears in short-term texture and shear variations during sliding or twisting (Li & Adelson, 2013). These complementary cues motivate learning physical properties (roughness or hardness) across the full temporal course of tactile interaction.

Prior tactile sequence models, however, capture only a single fixed temporal scale. Recurrent hardness estimation compresses each frame into a global feature before temporal modeling (Yuan et al., 2017c), and Octopi average-pools salient-frame embeddings into one video descriptor (Yu et al., 2024). Both track long-term deformation but smooth away brief local changes. Conversely, STAM (Cao et al., 2020) attends to informative regions and temporal samples for texture recognition, capturing local cues yet not the full loading-and-recovery process that hardness requires. Global aggregation thus blurs short sliding events, while local models miss the full deformation history, and

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

once these cues are discarded, later visual or language fusion cannot recover them. No existing model preserves short patch-level events and full-interaction evolution in parallel, or adapts their relative contributions to each property. This raises our central question: how to retain both temporal scales in one representation from which hardness and roughness classification jointly draw the evidence each needs.

To address this, we propose **DATE**, a local-global temporal fusion model for contact-based physical-property perception. From each frame, a frozen tactile backbone extracts patch tokens and a class (CLS) token. A ***Local Temporal Encoding*** branch applies spatiotemporal attention over patch tokens within short windows, preserving localized texture and shear changes that mainly support roughness. In parallel, a ***Global Temporal Encoding*** branch processes the full CLS sequence with a bidirectional recurrent model, tracking contact establishment, loading, and release to support hardness. Grad-CAM examples appear in Fig. 1 and Appendix C. ***Temporal-Dynamics-Aware Gated Fusion*** then learns per-property weightings over these local, global, and fine-grained statistical cues, letting each property draw on the evidence it needs. The resulting tactile representation is fused with visual features and grounded in a frozen language model for material description, three-object ranking, and physical-property reasoning.

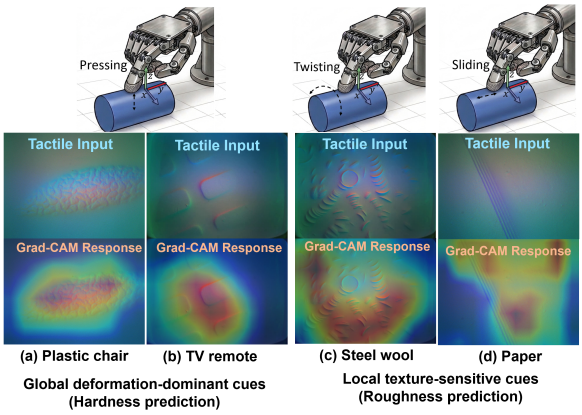

Figure 1: **Representative Grad-CAM responses for tactile property prediction.** After local-global spatiotemporal fusion, attention concentrates on the deformable regions most informative of each object's hardness and roughness: broad responses for softer objects (plastic chair) versus finer, localized focus for rough ones (steel wool). Warmer colors indicate higher saliency for the learned attention feature.

We evaluate **DATE** on VitaSet (Zong et al., 2026) and PhysiCLeAR (Yu et al., 2024) for hardness and roughness classification. We evaluate the **DATE (with vision)** for three-object ranking, and physical-property reasoning, and on the TVL benchmark (Fu et al., 2024) for tactile-language description. Across these settings, local short-term temporal features contribute most to roughness while global temporal features provide the main evidence for hardness, and the model remains effective across pressing, sliding, and twisting. Across score classification, language-grounded reasoning, and description tasks, **DATE** and **DATE (with vision)** consistently outperforms state-of-the-art vision-tactile baselines, supporting local-global temporal modeling for material-aware robot perception.

Our contributions are threefold:

- We propose a unified local-global temporal representation for vision-based tactile property perception. A local branch applies spatiotemporal attention over patch tokens within short windows to capture fine-grained texture and shear transitions, while a global branch encodes full-sequence deformation history. The two provide complementary cues for roughness and hardness classification.
- We introduce a temporal-dynamics-aware gated fusion that learns per-property weightings over the local, global, and statistical features, letting each property draw on the incremental and accumulated temporal evidence it needs. The model handles variable-length sequences and predicts graded property levels beyond binary classification, remaining effective across pressing, sliding, and twisting.
- We integrate the temporal tactile representation into a vision-tactile-language framework that grounds the predicted property levels in natural language, extending the model to material description and property reasoning, triplet object ranking by hardness or roughness, and yielding consistent gains on both classification and language-based property description tasks.

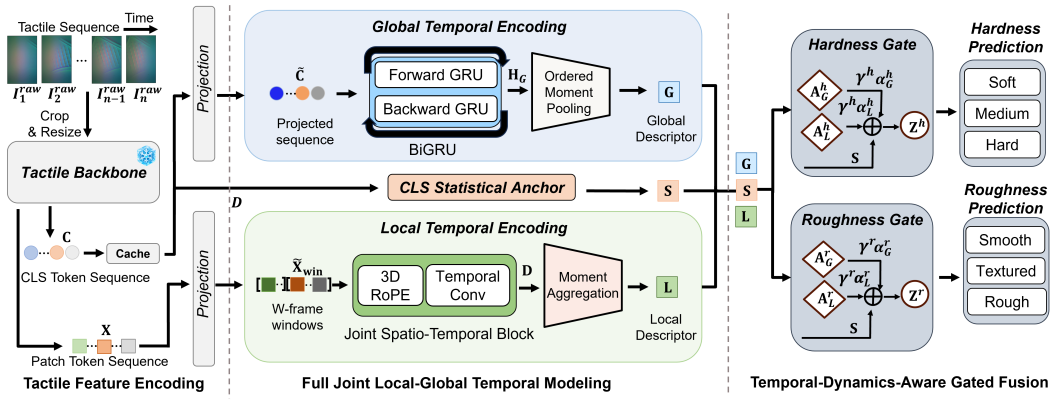

Figure 2: Architecture of the **DATE** tactile encoder. A frozen backbone extracts CLS and patch tokens from the observed tactile sequence $\mathcal{I}_n$. Cached CLS features support Global Temporal Encoding and the statistical anchor, while Local Temporal Encoding processes patch tokens within W-frame windows. Separate hardness and roughness gates fuse the local descriptor $\mathbf{L}$, global descriptor $\mathbf{G}$, and statistical anchor $\mathbf{S}$ into $\mathbf{Z}^h$ and $\mathbf{Z}^r$ for property classification.

2 Related Works

Vision-Based Tactile Sensing for Physical-Property Perception. Touch reveals physical properties that vision alone cannot capture. Vision-based tactile sensors such as GelSight (Yuan et al., 2017a) and DIGIT (Lambeta et al., 2020) record surface deformation via an internal camera, enabling three broad lines of work: geometry estimation (Dong et al., 2017), texture/material recognition (Li & Adelson, 2013), and property estimation such as hardness or clothing material (Yuan et al., 2017c; 2018). These studies confirm that tactile images encode rich cues about geometry, motion, and material. However, most rely on a single contact frame, capturing only one instant of an interaction. Comparing neighboring frames instead reveals deformation and shear dynamics that are critical for reliable property perception.

Temporal Modeling of Tactile Interactions. Prior work differs mainly in how temporal information is aggregated. Frame-pooling methods, e.g., Octopi (Yu et al., 2024), average-pool embeddings from selected salient frames. This is simple but order-agnostic and smooths out short-lived contact changes. Recurrent models (RNN/LSTM/GRU (Cho et al., 2014)), such as the CNN-RNN hardness estimator of Yuan *et al.*, (Yuan et al., 2017c), preserve sequential order and handle variable-length input, but typically compress each frame into a single global feature before temporal modeling, losing fine-grained local changes. Temporal convolution offers an alternative: Grover et al. (Grover et al., 2022) apply local filters for slip detection, but the cost grows with interaction length. Attention-based models retain richer spatiotemporal detail: STAM (Cao et al., 2020) attends to salient regions and time steps for texture recognition, and TimeSformer (Bertasius et al., 2021) models spatiotemporal patch interactions. Recent tactile works follow suit, such as AnyTouch (Feng et al., 2025) and AnyTouch 2 (Feng et al., 2026) learn from tactile images and video, capturing fine deformation and force dynamics, but attention cost grows with the number of frames and patches. MidasTouch (Suresh et al., 2023) integrates long-term tactile information, though for localization rather than property perception. Overall, existing methods capture either local temporal patterns or full clip-level context well, but not both, leaving open how to model variable-length interactions while retaining fine-grained local change and global temporal evolution.

Multimodal Physical-Property Reasoning. Vision-language models can infer material and fragility priors from images (Gao et al., 2024), but visual appearance alone is often ambiguous for hardness and roughness. Vision-tactile methods instead learn joint visual and tactile representations: (Yuan et al., 2017b) associates visual and tactile material properties, ViTac (Luo et al., 2018) shares features for cloth texture recognition, and UniTouch (Yang et al., 2024) aligns multiple tactile sensors with pretrained multimodal spaces. These methods improve cross-modal recognition and transfer but emphasize feature alignment over temporal structure. Vision-tactile-language models extend this further: TVL (Fu et al., 2024) aligns tactile, visual, and textual signals for classification and description, and VitaTouch (Zong et al., 2026) fuses visual and tactile features for property-aware

inspection. In all cases, downstream reasoning depends directly on the tactile encoder's output. If the encoder compresses an interaction into a single feature, information about frame order and local change is lost and cannot be recovered during fusion, underscoring the need to preserve tactile dynamics before multimodal reasoning.

These limitations motivate a local-global temporal representation that preserves both short-lived contact events and full-sequence context for physical-property prediction.

3 METHODS

3.1 PROBLEM FORMULATION AND FRAMEWORK OVERVIEW

We predict hardness and roughness from a vision-based tactile sensor. After n frames have been observed, the available tactile sequence is $\mathcal{I}_n = (\mathbf{I}_1, \ldots, \mathbf{I}_n)$. Here, n is the number of observed frames, and $\mathbf{I}_i \in \mathbb{R}^{3\times H_{\text{img}}\times W_{\text{img}}}$ is the preprocessed frame at index $i \in \{1, \ldots, n\}$. Here, H_{img} and W_{img} are the resized image dimensions. Preprocessing is described in Section 4.1. The sequence begins at contact onset. For a recorded clip, n is its full length. During incremental acquisition, $\mathcal{I}_n$ is the observed frame sequence. Throughout, "full sequence" refers to all frames available for the current prediction.

We use $p \in \{h, r\}$ for hardness property and roughness property. Each target score $y^p \in \{0, 1, 2\}$ represents one of three ordered levels: soft/medium/hard for hardness and smooth/textured/rough for roughness. Both temporal branches process tactile images. External vision is introduced only in the downstream framework. Fig. 2 details the tactile encoder for local and global temporal feature learning. Appendix B describes the downstream extension. Below, sequence-level features refer to $\mathcal{I}_n$ without repeating the index n. Named operators denote module types and have independent parameters unless sharing is explicitly stated.

3.2 TACTILE FEATURE ENCODING

A frozen AnyTouch backbone (Feng et al., 2025) extracts a CLS feature $\mathbf{c}_i \in \mathbb{R}^{d_0}$ and spatial patch tokens from each frame, where $d_0 = 768$. We pool the original 16×16 patch grid to 4×4, obtaining $\mathbf{X}_i \in \mathbb{R}^{N\times d_0}$ with $N = 16$ patches. Stacking the observed features gives $\mathbf{C} \in \mathbb{R}^{n\times d_0}$ and $\mathbf{X} \in \mathbb{R}^{n\times N\times d_0}$. The same learned projector maps each CLS or patch token to dimension $d = 192$, yielding $\widetilde{\mathbf{C}} \in \mathbb{R}^{n\times d}$ and $\widetilde{\mathbf{X}} \in \mathbb{R}^{n\times N\times d}$.

During incremental acquisition, each new frame is encoded once and its CLS feature is appended to the cached tokens. The cached tokens supply the global branch and statistical anchor, while local processing uses patch tokens from the available local temporal windows. The cache is reset at the start of each new interaction. The CLS statistical anchor $\mathbf{S} \in \mathbb{R}^{d_s}$, with $d_s = 64$, is,

$$\mathbf{S} = \text{Proj}([\mu(\mathbf{C}); \sigma(\mathbf{C})]), \tag{1}$$

here, brackets $[\cdot\,;\,\cdot]$ denote concatenation, μ and σ are the feature-wise mean and standard deviation over the n frames. Their concatenation is projected from $2d_0$ to d_s dimensions.

3.3 LOCAL-GLOBAL TEMPORAL ENCODING

The two branches operate in parallel at complementary temporal scales. Local encoding preserves short contact transition features, while global encoding summarizes the long-term interaction history.

Local Temporal Encoding: Let $\widetilde{\mathbf{X}}_{\text{win}} \in \mathbb{R}^{W\times N\times d}$ denote a local temporal window containing W observed frames. The window size W is distinct from the observed sequence length n. We omit window indices below for clarity. By default, the local module operates on inputs of this window size.

Flattening the temporal and patch axes yields $W \times N$ tokens, each with feature dimension d. Each token has coordinate $\boldsymbol{\rho} = (\tau, y, x) \in \mathbb{R}^3$, where τ is its relative time within the window and (y, x) locates its patch on the 4×4 grid. Independent linear mappings produce queries $\mathbf{Q}$, keys $\mathbf{K}$, and values $\mathbf{V}$, each in $\mathbb{R}^{(W\times N)\times d_a}$ for one attention head of dimension d_a. We apply 3D rotary position embedding (RoPE) (Su et al., 2024) to each query and key vector $\mathbf{z}_{\mathbf{Q},\mathbf{K}} \in \mathbb{R}^{d_a}$:

$$\text{RoPE}(\mathbf{z}_{\mathbf{Q},\mathbf{K}}, \boldsymbol{\rho}) = [\mathcal{R}(\tau) \oplus \mathcal{R}(y) \oplus \mathcal{R}(x)]\,\mathbf{z}, \tag{2}$$

here, $\mathcal{R}$ denotes the standard rotary operator on the corresponding temporal or spatial dimension. The rotated queries and keys contain temporal and spatial channel-group tokens, respectively. The direct sum $\oplus$ forms a block-diagonal rotation of size $d_a \times d_a$. We denote the temporal feature tokens by $\mathbf{Q}^{\text{rot}}_{\text{temp}}, \mathbf{K}^{\text{rot}}_{\text{temp}} \in \mathbb{R}^{(W\times N)\times d_{\text{temp}}}$, and the spatial groups by $\mathbf{Q}^{\text{rot}}_{\text{spat}}, \mathbf{K}^{\text{rot}}_{\text{spat}} \in \mathbb{R}^{(W\times N)\times d_{\text{spat}}}$. Their channel dimensions satisfy $d_{\text{temp}} + d_{\text{spat}} = d_a$. The temporal groups use $\mathcal{R}(\tau)$, while the spatial groups use $\mathcal{R}(y) \oplus \mathcal{R}(x)$.

Concatenating the temporal and spatial groups along the feature dimension gives $\mathbf{Q}^{\text{rot}}, \mathbf{K}^{\text{rot}} \in \mathbb{R}^{(W\times N)\times d_a}$. The single attention-head output $\mathbf{H}_{\text{atten}} \in \mathbb{R}^{(W\times N)\times d_a}$ is,

$$\mathbf{H}_{\text{atten}} = \text{softmax}\left(\frac{\mathbf{Q}^{\text{rot}}(\mathbf{K}^{\text{rot}})^\top}{\sqrt{d_a}}\right)\mathbf{V}. \tag{3}$$

All tokens within the observed window can attend to one another. For J attention heads, let $\mathbf{H}^{(j)}_{\text{atten}} \in \mathbb{R}^{(W\times N)\times d_a}$ denote the output of head $j \in \{1, \ldots, J\}$. We concatenate these outputs along the feature dimension. The local output $\mathbf{H}_L \in \mathbb{R}^{W\times N\times d}$ is,

$$\mathbf{H}_L = \text{Refine}\Big(\widetilde{\mathbf{X}}_{\text{win}}, [\,\mathbf{H}^{(1)}_{\text{atten}}; \ldots; \mathbf{H}^{(J)}_{\text{atten}}\,]\Big), \tag{4}$$

here, Refine denotes the local processing after attention. It includes output projection, reshaping, temporal depthwise convolution, and a gated feed-forward layer. The subscript L denotes local features.

To summarize each local temporal window, we combine its overall feature statistics with the temporal change between the beginning and end of the window. Specifically, we average the patch features within the first and last frames of $\mathbf{H}_L$ to obtain the frame-level features $\mathbf{H}^{\text{start}}_L, \mathbf{H}^{\text{end}}_L \in \mathbb{R}^d$. The window descriptor $\mathbf{D} \in \mathbb{R}^d$ is then computed as,

$$\mathbf{D} = \text{Proj}\big([\,\mu(\mathbf{H}_L); \sigma(\mathbf{H}_L); \mathbf{H}^{\text{end}}_L - \mathbf{H}^{\text{start}}_L\,]\big), \tag{5}$$

here, μ and σ summarize the overall feature distribution within the window, while $\mathbf{H}^{\text{end}}_L - \mathbf{H}^{\text{start}}_L$ captures its relative temporal change. Their concatenation has dimension $3d$, which is projected back to d dimensions.

Each local temporal window is therefore represented by one descriptor $\mathbf{D}$. We aggregate the descriptors from all available local windows to obtain a single local descriptor for the observed interaction. Let $\mathbf{D}_1, \ldots, \mathbf{D}_K \in \mathbb{R}^d$ denote the descriptors of the K available local windows. Moment Aggregation produces the local descriptor $\mathbf{L} \in \mathbb{R}^d$:

$$\mathbf{L} = \text{Proj}([\,\mu(\mathbf{D}_1, \ldots, \mathbf{D}_K); \sigma(\mathbf{D}_1, \ldots, \mathbf{D}_K)\,]). \tag{6}$$

The statistics now operate across window descriptors, and the projection maps $2d$ to d dimensions. Thus, $\mathbf{L}$ aggregates short-term information without a fixed number of windows.

Global Temporal Encoding: A local window may not contain the full deformation history. Long-term deformation evolution provides useful context for property classification prediction. This indicates the stage of maximal deformation reached during the whole interaction. Maximal deformation motivates full-sequence encoding.

We use a bidirectional GRU (BiGRU) (Cho et al., 2014; Schuster & Paliwal, 1997) with 96 hidden units per direction. It processes $\widetilde{\mathbf{C}}$ and produces $\mathbf{H}_G \in \mathbb{R}^{n\times d}$, where G denotes global states. Let $\mathbf{H}^{\max}_G, \mathbf{H}^{\min}_G \in \mathbb{R}^d$ denote the feature-wise maxima and minima of $\mathbf{H}_G$ over the n observed frames. Sequence projection yield the global descriptor $\mathbf{G} \in \mathbb{R}^d$:

$$\mathbf{G} = \text{Proj}\big([\,\mu(\mathbf{H}_G); \sigma(\mathbf{H}_G); \mathbf{H}^{\max}_G - \mathbf{H}^{\min}_G\,]\big). \tag{7}$$

These statistics summarize the mean, standard deviation, and feature-wise range over observed frames, with projection from $3d$ to d. Each prediction reevaluates the BiGRU on the cached history.

3.4 Temporal-Dynamics-Aware Gated Fusion

We use separate fusion gates for hardness property and roughness property. For property $p \in \{h, r\}$, independent projections and layer normalizations give $\mathbf{A}^p_L = \text{LN}(\text{Proj}(\mathbf{L}))$ and $\mathbf{A}^p_G =$

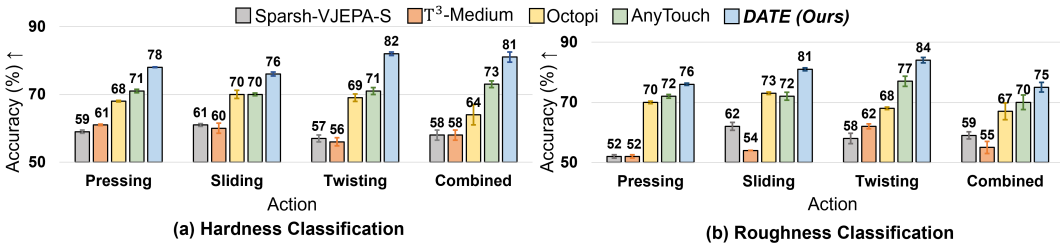

Figure 3: Classification accuracy (%, vertical axis) across exploration settings (horizontal axis) for (a) hardness and (b) roughness. Pressing, sliding, and twisting use extended PhysiCLeAR. Combined uses VitaSet. Bars and error bars show three-seed means and standard deviations.

$\mathrm{LN}(\mathrm{Proj}(\mathbf{G}))$, both in $\mathbb{R}^{d_s}$. These property-specific vectors align the temporal descriptors with the statistical feature space. The normalized anchor is $\overline{\mathbf{S}} = \mathrm{LN}(\mathbf{S}) \in \mathbb{R}^{d_s}$. For property p, the gate takes three inputs: the normalized statistical feature $\overline{\mathbf{S}}$, the local feature $\mathbf{A}_L^p$, and the global feature $\mathbf{A}_G^p$. It concatenates these features and then predicts triple scalar logits $\mathbf{u}^p = (u_L^p, u_G^p, u_\gamma^p) \in \mathbb{R}^3$:

$$\mathbf{u}^p = \mathrm{MLP}_p\big([\overline{\mathbf{S}}; \mathbf{A}_L^p; \mathbf{A}_G^p]\big)\,. \tag{8}$$

The gated weight scalars are $(\alpha_L^p, \alpha_G^p) = \mathrm{softmax}([u_L^p, u_G^p])$. The sigmoid $\gamma^p = \mathrm{sigmoid}(u_\gamma^p)$ merges the local and global descriptors by weighted sum, while $\gamma^p \in (0, 1)$ controls the temporal residual. The fused feature $\mathbf{Z}^p \in \mathbb{R}^{d_s}$ is,

$$\mathbf{Z}^p = \mathrm{LN}\big(\overline{\mathbf{S}} + \gamma^p(\alpha_L^p \mathbf{A}_L^p + \alpha_G^p \mathbf{A}_G^p)\big)\,. \tag{9}$$

The predicted property score $\widehat{y}^p \in \{0, 1, 2\}$ is the class index with the largest logit o_c^p, where $c \in \{0, 1, 2\}$ indexing classes.

3.5 Loss Function

For hardness and roughness classification, we use a label-smoothed cross-entropy loss $\mathcal{L}_{\mathrm{ce}}(\mathbf{o}, \mathbf{y})$, where $\mathbf{o} \in \mathbb{R}^3$ are the logits and $\mathbf{y} \in \{0, 1, 2\}$ is the target class. The overall objective combines this main classification loss with auxiliary supervision:

$$\mathcal{L}_{\mathrm{tac}} = \frac{1}{2} \sum_{p \in \{h,r\}} \mathcal{L}_{\mathrm{ce}}(\mathbf{o}^p, \mathbf{y}^p) + \lambda_{\mathrm{aux}} \mathcal{L}_{\mathrm{aux}}. \tag{10}$$

The first term averages the label-smoothed classification losses of the fused hardness and roughness predictions. The second term directly supervises the local and global branches. Its weight is controlled by λ_{aux}.

Training-only heads applied to $\mathbf{L}$ and $\mathbf{G}$ produce auxiliary logits $\mathbf{o}_L^p, \mathbf{o}_G^p \in \mathbb{R}^3$. The auxiliary term averages the four losses from the two branches and two properties:

$$\mathcal{L}_{\mathrm{aux}} = \frac{1}{4} \sum_{p \in \{h,r\}} [\mathcal{L}_{\mathrm{ce}}(\mathbf{o}_L^p, y^p) + \mathcal{L}_{\mathrm{ce}}(\mathbf{o}_G^p, y^p)]\,. \tag{11}$$

This auxiliary supervision encourages the local and global branches to remain individually discriminative for both properties before fusion, and the corresponding heads are discarded at inference time.

4 Experiments

4.1 Experimental Setup

Datasets and preprocessing. We evaluate hardness and roughness on VitaSet (Zong et al., 2026) and PhysiCLeAR (Yu et al., 2024). The VitaSet covers 186 objects and 21,510 tactile images, including the associated TacQuad (Feng et al., 2025) data. Although VitaTouch evaluates hardness as a binary task, we use the released Soft/Medium/Hard annotations across the full VitaSet, with the

Table 1: Three-level tactile property classification on VitaSet.

Method	Accuracy (%) ↑		Macro F1 (%) ↑	
	Hardness	Roughness	Hardness	Roughness
TVL (ViT-Tiny) (Fu et al., 2024)	57.49±3.70	49.89±0.39	42.11±5.30	39.23±2.12
TVL (ViT-Small)	57.72±1.16	48.32±1.78	40.92±1.60	39.84±3.39
TVL (ViT-Base)	59.28±6.89	53.47±2.36	45.82±7.44	46.10±4.35
T^3-Small (Zhao et al., 2025)	57.74±2.68	52.57±1.03	52.91±5.12	48.88±3.31
T^3-Medium	58.17±1.40	55.48±2.05	54.73±1.10	53.06±3.21
T^3-Large	58.61±2.05	56.15±1.69	55.08±0.51	53.13±2.09
Sparsh-VJEPA-S (Higuera et al., 2025)	57.70±1.34	59.06±1.16	42.79±0.31	53.41±0.54
Octopi (Yu et al., 2024)	63.53±3.03	67.11±2.79	55.10±2.74	61.87±3.38
AnyTouch (Feng et al., 2025)	73.38±1.03	70.25±2.36	69.49±1.33	65.41±1.87
DATE (Ours)	**80.76**±1.04	**75.39**±1.55	**73.24**±1.89	**70.33**±0.53

Results are reported as mean ± standard deviation over three seeds. **Bold** and underlining indicate the highest and second-highest.

Medium class provided by its reannotated TacQuad subset. Original PhysiCLeAR contains 408 tactile videos of 74 everyday objects and provides three-level property annotations. We also use extended PhysiCLeAR from Octopi-1.5 (Yu et al., 2025) for action-wise (Figure 3) and label-granularity (Figure 5) experiments. Extended PhysiCLeAR published 2,689 videos of 100 objects. We rescale the extended dataset's $[0, 10]$ property scores into three $[0, 2]$ or four ordered intervals $[0, 3]$ and train a classifier for each setting. These tasks examine finer distinctions than binary classification. All tactile inputs are contact-cropped, resized to 224×224, and normalized. Methods within each comparison use the same split. The action-wise (Figure 3) and label-granularity (Figure 5) experiments use different splits.

Baselines. We compare the tactile encoders of TVL, T^3, Sparsh,Octopi, and AnyTouch. The model sizes are listed in Table 1 and Table 5. The downstream description comparison additionally uses our vision-tactile-language framework.

Training and evaluation. We train **DATE** on four NVIDIA RTX 4090 GPUs using AdamW with a learning rate of 3×10^{-4}, a batch size of 8, and 30 epochs. The AnyTouch backbone remains frozen. The temporal and prediction modules are jointly optimized with Eq. (10). The default temporal local window is $W = 5$. Classification results report accuracy and Macro-F1 over three seeds, with standard deviations shown unless omitted for compactness. Macro-F1 weights each class equally. The VitaSet main comparison and ablations share the same split and seeds. Accuracy and F1 gains below are expressed in percentage (%). TVL uses its original evaluation protocol.

4.2 Evaluation Results

Tactile property classification. Table 1 compares three-level classification on VitaSet. **DATE** achieves the highest mean in all four metrics. Its hardness and roughness accuracies reach 80.76% and 75.39%, exceeding AnyTouch by 7.38% and 5.14%, respectively. Macro-F1 also increases from 69.49% to 73.24% for hardness and from 65.41% to 70.33% for roughness. AnyTouch is also our backbone, so this comparison shows that the added temporal representation drives the improvement in property prediction.

Comparison of actions. Fig. 3 compares pressing, sliding, and twisting on extended PhysiCLeAR. Combined from vitaSet. **DATE** outperforms the listed baselines for both properties in all three action groups. Its rounded hardness accuracy is 78% under pressing and 76% under sliding. Roughness accuracy achieves 81%. Twisting gives the highest values, at 82% and 84% for hardness and roughness, respectively. This pattern is consistent with pressing exposing deformation cues for hardness and sliding exposing local surface changes via roughness. Twisting combines both forms of contact variation.

Moving from three to four label levels lowers accuracy and Macro-F1, but **DATE** retains the highest mean for both properties and metrics among the compared methods (Appendix A).

Frame-wise qualitative comparison. Fig. 4 shows predictions from two baselines and our model on four tactile sequences. The single-frame ViT-Base has no access to earlier contacts, while the

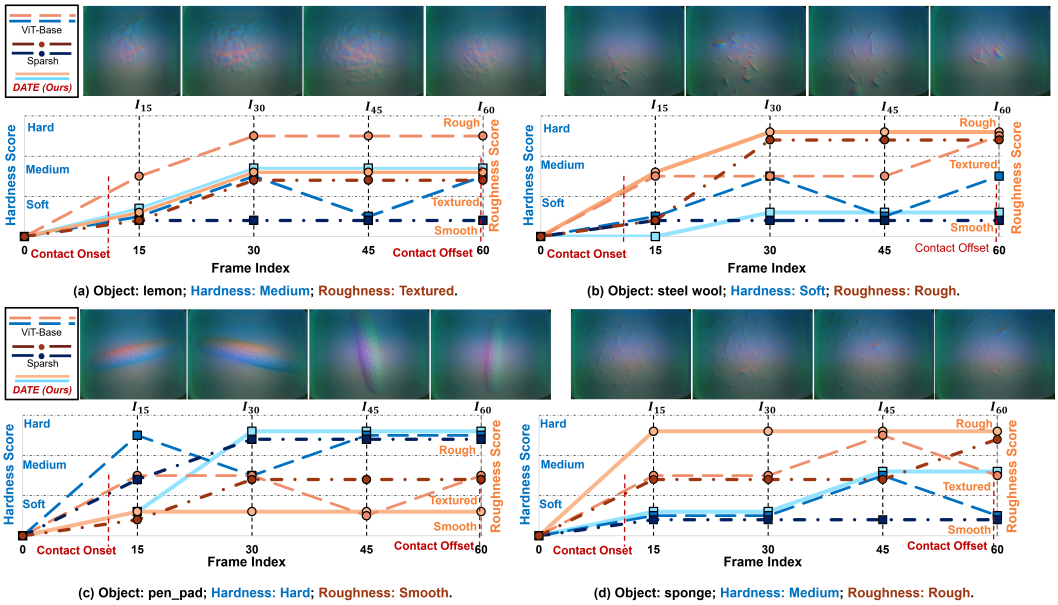

Figure 4: Qualitative property predictions versus frame index (horizontal axis), with hardness and roughness on the left and right vertical axes. Examples from PhysiCLeAR are (a) lemon, (b) steel wool, (c) pen pad, and (d) sponge. Blue and orange indicate hardness and roughness. Line styles distinguish models. Tactile images show frames 15, 30, 45, and 60. Red dashed lines mark contact onset and offset.

Table 2: Cross-dataset transfer and mixed-data training on VitaSet and PhysiCLeAR.

Method	**Accuracy (%) ↑**					
	VitaSet → PhysiCLeAR		**PhysiCLeAR → VitaSet**		**Mixed → Mixed**	
	Hardness	**Roughness**	**Hardness**	**Roughness**	**Hardness**	**Roughness**
TVL (ViT-Small)	30.84±0.34	44.25±1.64	38.74±3.37	24.32±2.21	60.29±0.69	64.33±0.68
T^3-Large	28.09±0.27	**47.18**±3.72	64.86±0.00	45.95±0.00	71.42±0.62	73.46±0.89
Sparsh-VJEPA-S	27.48±0.63	38.88±12.58	44.14±6.37	24.32±0.00	69.58±0.56	70.50±0.41
AnyTouch	27.87±0.12	23.42±3.35	65.77±1.27	48.65±0.00	75.04±0.26	78.38±0.20
DATE (Ours)	**32.94**±1.27	45.83±2.48	**70.27**±0.00	**52.25**±1.27	**88.54**±0.26	**91.08**±0.42

Arrows indicate training → test datasets. Mixed denotes the union of the two datasets' respective training or held-out test splits. Results are reported as mean ± standard deviation over three seeds. **Bold** and underlining indicate the highest and second-highest mean, respectively.

fixed-scale Sparsh model gives temporally consistent but often incorrect labels (e.g., lemon, pen pad, sponge). In contrast, **DATE** combines local changes with the cached history in $\mathcal{I}_n$. These examples show how the two property outputs develop as contact progresses.

Cross-dataset evaluation. Table 2 evaluates transfer between VitaSet and PhysiCLeAR. Models are trained and selected on the source domain, then tested on the target without fine-tuning. For PhysiCLeAR → VitaSet, **DATE** reaches 70.27% hardness and 52.25% roughness accuracy. In the reverse direction, it has the highest hardness accuracy, Mixed pools the training splits of both datasets and evaluates their pooled held-out test samples. **DATE** reaches 88.54% hardness and 91.08% roughness accuracy. The result shows that **DATE** can benefit from exposure to both data sources, while source-only transfer remains more challenging.

Downstream applicability. We integrate **DATE** with Qwen3-VL-8B-Instruct for tactile-language description on TVL (Fu et al., 2024). **DATE (with vision)** achieves an overall score of 5.66, exceeding TVL-LLaMA (ViT-Base) by 0.63 points. Appendix B gives full details. Appendix D illustrates the robotic workflow.

Table 3: Local-global temporal ablation of **DATE** on VitaSet. **H** denotes hardness, and **R** indicated roughness.

Model	Accuracy (%) ↑				Macro F1 (%) ↑			
	Hardness	Roughness	Δ_h	Δ_r	Hardness	Roughness	Δ_h	Δ_r
w/o Temporal	75.62	72.93	–	–	70.78	66.97	–	–
w/o Local	78.52	72.48	+2.90	-0.45	72.36	66.74	+1.58	-0.23
w/o Global	75.84	74.94	+0.22	+2.01	71.07	69.91	+0.29	+2.94
DATE (Full)	**80.76**	**75.39**	**+5.14**	**+2.46**	**73.24**	**70.33**	**+2.46**	**+3.36**

Results are averaged over three seeds. Standard deviations are omitted for compactness. Within each metric group, Δ_h and Δ_r denote changes relative to w/o Temporal, in percentage points. w/o Temporal removes both temporal branches. w/o Local and w/o Global remove the corresponding branch.

Table 4: Local window-length ablation of **DATE** on VitaSet.

Window size W	Accuracy (%) ↑		Macro F1 (%) ↑	
	Hardness	Roughness	Hardness	Roughness
$W = 1$	80.31±1.69	71.81±2.32	69.34±4.37	68.65±0.94
$W = 3$	80.54±1.78	72.48±1.78	72.81±3.63	69.56±0.97
$W = 5$ (default)	**80.76**±1.04	**75.39**±1.55	73.24±1.89	**70.33**±0.53
$W = 8$	80.54±1.34	71.81±1.78	**73.75**±4.05	68.33±0.91

W denotes the number of frames in a local window. Results are reported as mean ± standard deviation over three seeds. **Bold** and underlining indicate the highest and second-highest mean.

Table 5: Comparison of model size and inference runtime.

Model	Params (M) ↓	Infer (ms) ↓	Model	Params (M) ↓	Infer (ms) ↓
Octopi	303.38	627.62	TVL-Tiny	**5.68**	**16.83**
T^3-Small	21.67	41.32	Sparsh-S	21.90	169.71
T^3-Medium	85.80	122.82	AnyTouch	305.23	585.49
T^3-Large	189.94	252.49	**DATE (Ours)**	306.61	186.35

Parameter counts are in millions. We evaluate **DATE**'s runtime using small pre-trained encoder and post-encoder modules at test time.

4.3 Ablation Study

Local and global temporal encoding. Table 3 compares temporal branches. The internal w/o Temporal variant retains the statistical anchor and prediction heads. It differs from standalone AnyTouch. Adding Global improves hardness accuracy by 2.90% but slightly lowers roughness accuracy. Adding Local improves roughness accuracy and Macro-F1 by 2.01% and 2.94%. Full achieves the highest mean in all four metrics. Removing Global from Full lowers hardness accuracy by 4.92%, removing Local lowers roughness accuracy by 2.91%, supporting complementary temporal contributions in this setting.

Local window length. Table 4 varies window size (W) from 1 to 8. The global branch remains active at $W = 1$, so this setting does not remove all temporal information. The default $W = 5$ yields the highest mean in three of four metrics and improves roughness accuracy by 3.58% over $W = 1$. Extending the window to $W = 8$ raises hardness Macro-F1 slightly, from 73.24% to 73.75%, but lowers roughness performance. A moderate window gives a better balance.

Model size and runtime. Table 5 reports the computational footprint. **DATE** has 306.61M parameters, only 1.38M more than, AnyTouch, an increase of approximately 0.45%.

5 Conclusion

We present **DATE**, which perceives hardness and roughness from tactile interaction sequences by fusing local and full-sequence temporal encoding. On VitaSet, it improves accuracy over the SOTA baseline by 7.38 and 5.14 points, and our vision-tactile-language framework attains the top TVL score among compared models. It currently covers these two properties and depends on a costly backbone. Future work will add geometry, mass, and impact dynamics while reducing backbone and temporal-module cost.

AI USE STATEMENT

Generative AI tools were used to improve the grammar and readability of the manuscript. The authors take responsibility for the final content.

REFERENCES

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 813–824. PMLR, 2021. URL https://proceedings.mlr.press/v139/bertasius21a.html.

Guanqun Cao, Yi Zhou, Danushka Bollegala, and Shan Luo. Spatio-temporal attention model for tactile texture recognition. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 9896–9902, 2020. doi: 10.1109/IROS45743.2020.9341333.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734, 2014. doi: 10.3115/v1/D14-1179.

Siyuan Dong, Wenzhen Yuan, and Edward H. Adelson. Improved GelSight tactile sensor for measuring geometry and slip. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 137–144, 2017. doi: 10.1109/IROS.2017.8202149.

Ruoxuan Feng, Jiangyu Hu, Wenke Xia, Tianci Gao, Ao Shen, Yuhao Sun, Bin Fang, and Di Hu. AnyTouch: Learning unified static-dynamic representation across multiple visuo-tactile sensors. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=XToAemis1h.

Ruoxuan Feng, Yuxuan Zhou, Siyu Mei, Dongzhan Zhou, Pengwei Wang, Shaowei Cui, Bin Fang, Guocai Yao, and Di Hu. AnyTouch 2: General optical tactile representation learning for dynamic tactile perception. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ndilONnABZ.

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 *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 14080–14101. PMLR, 2024. URL https://proceedings.mlr.press/v235/fu24b.html.

Jensen Gao, Bidipta Sarkar, Fei Xia, Ted Xiao, Jiajun Wu, Brian Ichter, Anirudha Majumdar, and Dorsa Sadigh. Physically grounded vision-language models for robotic manipulation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 12462–12469, 2024. doi: 10.1109/ICRA57147.2024.10610090.

Abhinav Grover, Philippe Nadeau, Christopher Grebe, and Jonathan Kelly. Learning to detect slip with barometric tactile sensors and a temporal convolutional neural network. In *2022 International Conference on Robotics and Automation (ICRA)*, pp. 570–576, 2022. doi: 10.1109/ICRA46639.2022.9811592.

Carolina Higuera, Akash Sharma, Chaithanya Krishna Bodduluri, Taosha Fan, Patrick Lancaster, Mrinal Kalakrishnan, Michael Kaess, Byron Boots, Mike Lambeta, Tingfan Wu, and Mustafa Mukadam. Sparsh: Self-supervised touch representations for vision-based tactile sensing. In *Proceedings of the 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 885–915. PMLR, 2025. URL https://proceedings.mlr.press/v270/higuera25a.html.

Zhanat Kappassov, Juan-Antonio Corrales, and Véronique Perdereau. Tactile sensing in dexterous robot hands—review. *Robotics and Autonomous Systems*, 74:195–220, 2015. doi: 10.1016/j.robot.2015.07.015.

Mike Lambeta, Po-Wei Chou, Stephen Tian, Brian Yang, Benjamin Maloon, Victoria Rose Most, Dave Stroud, Raymond Santos, Ahmad Byagowi, Gregg Kammerer, Dinesh Jayaraman, and Roberto Calandra. DIGIT: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation. *IEEE Robotics and Automation Letters*, 5(3):3838–3845, 2020. doi: 10.1109/LRA.2020.2977257.

Rui Li and Edward H. Adelson. Sensing and recognizing surface textures using a GelSight sensor. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1241–1247, 2013. doi: 10.1109/CVPR.2013.164.

Shan Luo, Wenzhen Yuan, Edward H. Adelson, Anthony G. Cohn, and Raul Fuentes. ViTac: Feature sharing between vision and tactile sensing for cloth texture recognition. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2722–2727, 2018. doi: 10.1109/ICRA.2018.8460494.

Qwen Team. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Mike Schuster and Kuldip K. Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997. doi: 10.1109/78.650093.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. doi: 10.1016/j.neucom.2023.127063.

Sudharshan Suresh, Zilin Si, Stuart Anderson, Michael Kaess, and Mustafa Mukadam. MidasTouch: Monte-carlo inference over distributions across sliding touch. In *Proceedings of the 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pp. 319–331. PMLR, 2023. URL https://proceedings.mlr.press/v205/suresh23a.html.

Fengyu Yang, Chao Feng, Ziyang Chen, Hyoungseob Park, Daniel Wang, Yiming Dou, Ziyao Zeng, Xien Chen, Rit Gangopadhyay, Andrew Owens, and Alex Wong. Binding touch to everything: Learning unified multimodal tactile representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26330–26343, 2024. doi: 10.1109/CVPR52733.2024.02488.

Samson Yu, Kelvin Lin, Anxing Xiao, Jiafei Duan, and Harold Soh. Octopi: Object property reasoning with large tactile-language models. In *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, July 2024. doi: 10.15607/RSS.2024.XX.066.

Samson Yu, Kelvin Lin, and Harold Soh. Demonstrating the Octopi-1.5 visual-tactile-language model. *arXiv preprint arXiv:2507.09985*, 2025. doi: 10.48550/arXiv.2507.09985.

Wenzhen Yuan, Siyuan Dong, and Edward H. Adelson. GelSight: High-resolution robot tactile sensors for estimating geometry and force. *Sensors*, 17(12):2762, 2017a. doi: 10.3390/s17122762.

Wenzhen Yuan, Shaoxiong Wang, Siyuan Dong, and Edward H. Adelson. Connecting look and feel: Associating the visual and tactile properties of physical materials. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4494–4502, 2017b. doi: 10.1109/CVPR.2017.478.

Wenzhen Yuan, Chenzhuo Zhu, Andrew Owens, Mandayam A. Srinivasan, and Edward H. Adelson. Shape-independent hardness estimation using deep learning and a GelSight tactile sensor. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 951–958, 2017c. doi: 10.1109/ICRA.2017.7989116.

Wenzhen Yuan, Yuchen Mo, Shaoxiong Wang, and Edward H. Adelson. Active clothing material perception using tactile sensing and deep learning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4842–4849, 2018. doi: 10.1109/ICRA.2018.8461164.

Jialiang Zhao, Yuxiang Ma, Lirui Wang, and Edward Adelson. Transferable tactile transformers for representation learning across diverse sensors and tasks. In *Proceedings of the 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 3766–3779. PMLR, 2025. URL https://proceedings.mlr.press/v270/zhao25c.html.

Junyi Zong, Qingxuan Jia, Meixian Shi, Tong Li, Jiayuan Li, Zihang Lv, Gang Chen, and Fang Deng. VitaTouch: Property-aware vision-tactile-language model for robotic quality inspection in manufacturing. *arXiv preprint arXiv:2604.03322*, 2026. doi: 10.48550/arXiv.2604.03322. URL https://arxiv.org/abs/2604.03322.

APPENDIX

A EFFECT OF LABEL GRANULARITY

A.1 EXPERIMENTAL MOTIVATION AND SETUP

The main classification experiments evaluate hardness and roughness using three ordered levels. The label-granularity experiment examines whether the advantage of **DATE** is also observed when the prediction task requires four levels. Increasing the number of categories requires the classifier to distinguish additional ranges of the annotated property scores. This provides a complementary evaluation of the representation beyond the three-level setting. Hardness and roughness are evaluated separately because their temporal evidence differs, and the comparison reports both accuracy and Macro-F1 for each property.

We use the extended PhysiCLeAR data released with Octopi-1.5 (Yu et al., 2025). Its property annotations are scores in $[0, 10]$. For the two classification settings, these annotations are discretized into three or four ordered categories, represented by class indices $\{0, 1, 2\}$ and $\{0, 1, 2, 3\}$, respectively. A classifier is trained for each label granularity. The inputs follow the contact cropping, 224×224 resizing, and normalization described in Section 4.1.

Figure 5 compares TVL (ViT-Base), T^3-Large, Sparsh-VJEPA-S, AnyTouch, and **DATE**. Methods within each comparison use the same data split, as specified in the main experimental setup. Accuracy measures the fraction of correctly classified examples, while Macro-F1 gives equal weight to the classes. Reporting both metrics characterizes performance under each label granularity without relying on accuracy alone. Classification results follow the three-seed reporting convention in Section 4.1.

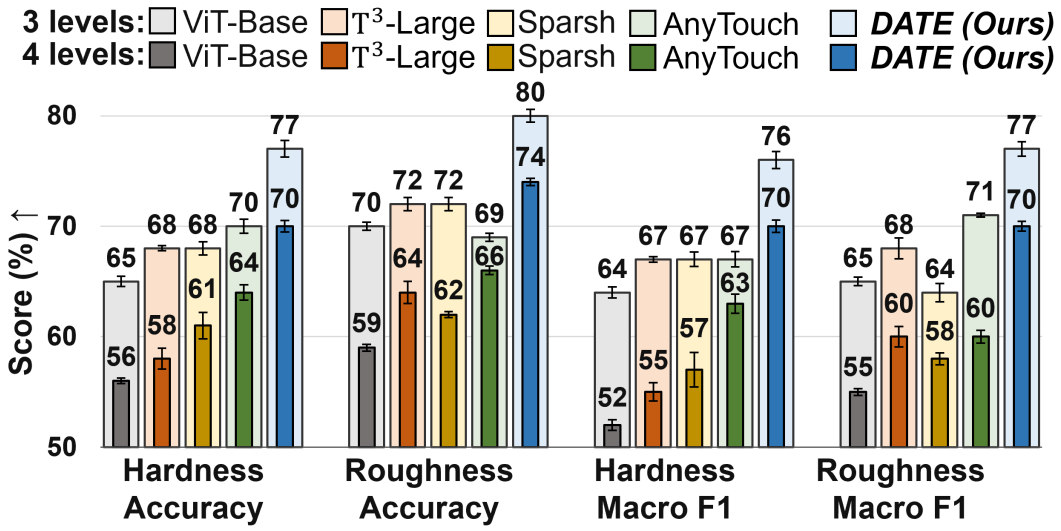

Figure 5: Accuracy or Macro-F1 (%, vertical axis) across property-metric pairs (horizontal axis) for three- and four-level classification on extended PhysiCLeAR. Wide, light bars show three levels. Narrow, dark bars show four. ViT-Base and Sparsh denote TVL (ViT-Base) and Sparsh-VJEPA-S.

A.2 RESULTS AND INTERPRETATION

Label granularity. Fig. 5 tests **DATE** whether the advantage persists with finer property labels. Moving from three to four levels lowers accuracy and Macro-F1 for all methods. Nevertheless, **DATE** retains the highest mean for each property and metric. Its rounded hardness and roughness accuracies change from $77\%/80\%$ to $70\%/74\%$, and its four-level Macro-F1 is 70% for both properties.

The four groups of bars show hardness accuracy, roughness accuracy, hardness Macro-F1, and roughness Macro-F1. Wide, light bars indicate three-level classification, and narrow, dark bars

indicate four-level classification. The numerical labels are rounded percentages. Error bars show variation across seeds and do not establish statistical significance. The vertical axis begins at 50%, so differences should be read using the labelled values rather than the apparent ratios of bar heights.

For **DATE**, the three-level hardness and roughness Macro-F1 values shown in the figure are 76% and 77%, respectively. Together with the accuracy values, these correspond to approximate three-to-four-level decreases of seven and six percentage points in hardness and roughness accuracy, and six and seven percentage points in their respective Macro-F1 scores. These differences are computed from the rounded figure labels. In the four-level setting, **DATE** has the highest displayed mean for all four metrics among the methods included in this comparison. The result supports the usefulness of the representation under both reported label spaces.

A.3 Relation to the Main Evaluations

This experiment complements the main three-level classification results by examining a finer categorization of the same two physical properties. It remains a classification evaluation: the reported accuracy and Macro-F1 do not measure continuous-score calibration or the magnitude of an error between adjacent property levels. The observed decrease in the four-level scores is consistent with a more demanding discrimination task, but the aggregate metrics alone do not identify which category boundaries account for that decrease.

The label-granularity experiment also uses a different split from the action-wise analysis in Figure 3, as stated in Section 4.1. Consequently, absolute scores from the two figures should be interpreted within their respective evaluation settings. The action-wise comparison examines performance across exploration settings, while this appendix examines the reported performance under three- and four-level labels.

B Supplementary TVL Results

B.1 Vision-Tactile-Language Integration

We integrate **DATE (with vision)** with Qwen3-VL-8B-Instruct (Qwen Team, 2025), keeping its language decoder frozen (Fig. 6). The native vision transformer and vision-language merger produce $\mathbf{F}_{\text{vis}} \in \mathbb{R}^{N_v \times d_m}$. A learned tactile connector maps $(\mathbf{Z}^h, \mathbf{Z}^r)$ to $\mathbf{F}_{\text{tac}} \in \mathbb{R}^{N_t \times d_m}$, where d_m is the language-model embedding dimension and N_v, N_t are the visual and tactile token counts. Temporal Token Fusion concatenates these tokens along the token axis. The predicted property levels and semantic labels enter the task prompt, whose encoding and projection produce $\mathbf{E}_{\text{txt}} \in \mathbb{R}^{N_x \times d_m}$, with N_x text tokens. The final input is $\mathbf{E}_{\text{concat}} \in \mathbb{R}^{(N_v+N_t+N_x)\times d_m}$:

$$\mathbf{E}_{\text{concat}} = [\,[\mathbf{F}_{\text{vis}}; \mathbf{F}_{\text{tac}}]; \mathbf{E}_{\text{txt}}\,]. \tag{12}$$

The tactile heads predict discrete property levels, and the language model generates corresponding semantic descriptions.

For the property-ranking task, the tokens and predicted property levels of the compared objects are placed in a shared prompt together with their object classes.

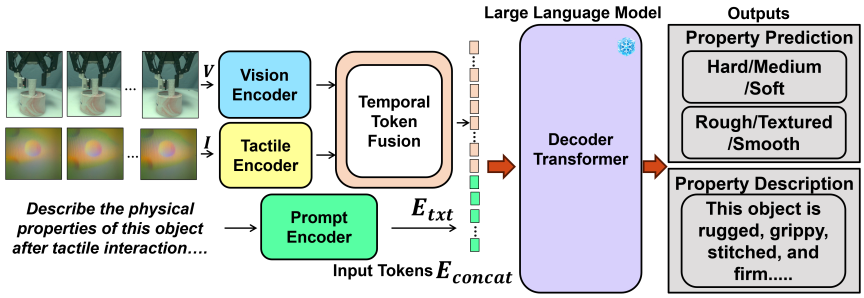

Figure 6: Downstream vision-tactile-language integration using Qwen3-VL-8B-Instruct. The inputs V and I denote external visual observations and the tactile sequence $\mathcal{I}_n$, respectively. Temporal Token Fusion concatenates the visual and tactile tokens. These are combined with the text tokens from Prompt Encoder to form $\mathbf{E}_{\text{concat}}$. The language model generates property labels and descriptions.

B.2 TVL Evaluation

Vision-tactile-language description. We evaluate the downstream framework (Fig. 6) on the Touch-Vision-Language (TVL) benchmark (Fu et al., 2024), using its original test set and generation/evaluation protocol. The test set contains 402 human-annotated image-touch pairs. SSVTP (Self-Supervised Visuo-Tactile Pretraining) contains robot-collected observations. HCT (Human Collected Tactile) contains human-collected observations. A text-only GPT-4 judge scores agreement with human-annotated tactile descriptions on a scale from 1 to 10. The TVL column is the overall test score. We evaluate the Octopi variants and **DATE (with vision)** following the same TVL protocol. We achieves the highest HCT and overall scores among these models: 5.62 and 5.66, respectively. The overall gain over TVL-LLaMA (ViT-Base) is 0.63 judge-score points.

Table 6: Tactile-language description on the TVL benchmark

Model	Modalities			Judge Score (1–10) ↑		
	V	T	L	SSVTP	HCT	TVL
LLaVA-1.5 7B	✓	–	✓	3.64	3.55	3.56
LLaVA-1.5 13B	✓	–	✓	3.55	3.63	3.62
ViP-LLaVA 7B	✓	–	✓	2.72	3.44	3.36
ViP-LLaVA 13B	✓	–	✓	4.10	3.76	3.80
LLaMA-Adapter	✓	–	✓	2.56	3.08	3.02
BLIP-2 Opt-6.7b	✓	–	✓	2.02	2.72	2.64
InstructBLIP 7B	✓	–	✓	1.40	1.30	1.31
InstructBLIP 13B	✓	–	✓	1.44	1.21	1.24
GPT-4V	✓	–	✓	5.02	4.42	4.49
SSVTP-LLaMA	✓	✓	–	2.58	3.67	3.54
Octopi 7B	–	✓	✓	3.85	3.93	3.92
Octopi 13B	–	✓	✓	3.70	4.02	3.98
Octopi-1.5	–	✓	✓	2.76	3.99	3.85
TVL-LLaMA (ViT-Tiny)	✓	✓	✓	<u>6.09</u>	4.79	4.94
TVL-LLaMA (ViT-Small)	✓	✓	✓	5.81	4.77	4.89
TVL-LLaMA (ViT-Base)	✓	✓	✓	**6.16**	<u>4.89</u>	<u>5.03</u>
DATE (with vision) (Ours)	✓	✓	✓	5.97	**5.62**	**5.66**

SSVTP and HCT denote the two test subsets. TVL denotes the overall test score.

C Paired Temporal Grad-CAM Visualizations

Figure 7 extends the examples in Figure 1 with paired explanations of the same tactile frames for two properties. The three pressing sequences show sausage, stone, and post box surfaces, with complete contexts of 48, 106, and 108 frames. Six chronological frames are displayed for each sequence. The target for each property is its predicted class from the complete sequence and remains fixed across the displayed frames.

Grad-CAM is computed from the **DATE** for each displayed frame while retaining the cached features of the other frames. Each map is independently normalized after projection onto the original image. Consequently, the color scale describes relative spatial attribution within a map.

In these selected examples, roughness responses tend to occupy more limited regions, while hardness responses are broader. The contrast varies across frames, and some responses extend towards the image periphery.

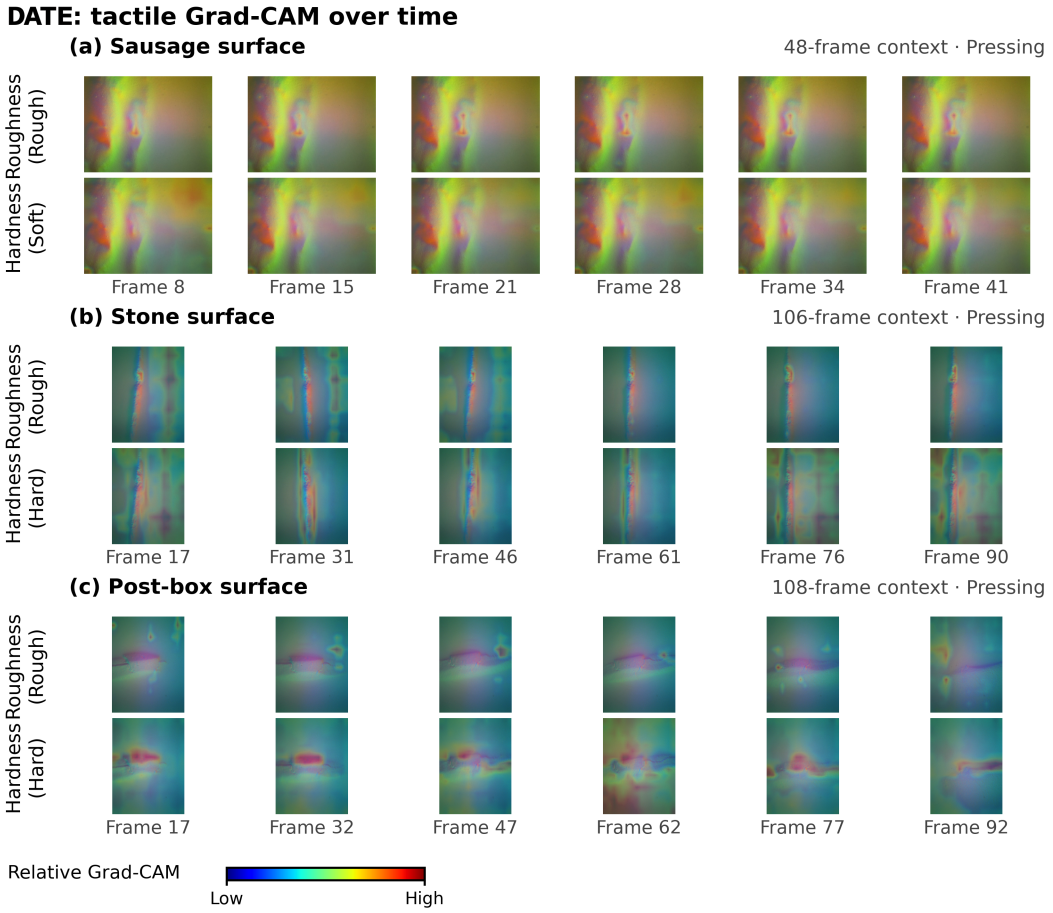

Figure 7: Paired Grad-CAM visualizations for roughness and hardness on three tactile sequences. Each group shows six chronological frames, with roughness responses above hardness responses for the same inputs. The complete sequence context is retained, and the predicted target class is fixed for each property. Maps are normalized independently. Warmer colors indicate higher relative attribution within each map.

D Robotic Interaction and Object Ranking

Figure 8 illustrates the robotic workflow associated with the multimodal framework in Appendix B. During pressing on a sponge and sliding on a glass cup, the interface shows hardness and roughness labels at successive stages of contact. The later snapshots show the final categories displayed for each example.

After interactions with the sponge, rubber duck, and walnut, the output presents their property rankings. Hardness places the tied sponge and rubber duck below walnut. Roughness places rubber duck below the tied sponge and walnut. Equal signs denote shared categories, without implying equality of calibrated physical properties. The timestamps identify snapshots within the interaction.

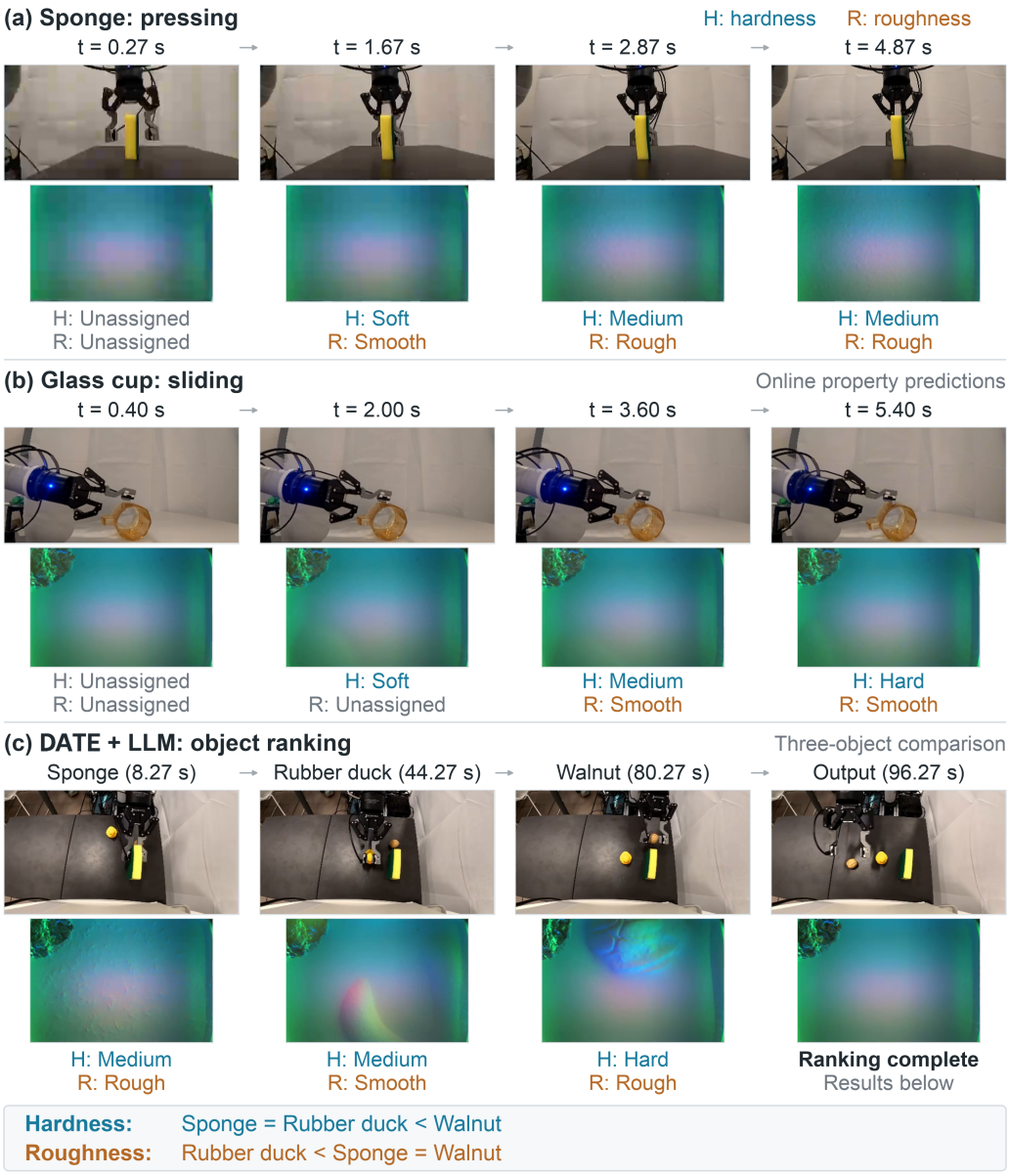

Figure 8: Robotic interaction and object ranking demonstration. Panels (a) and (b) show pressing a sponge and sliding on a glass cup, respectively. RGB and tactile observations are paired at the displayed times. The displayed hardness and roughness labels change during contact and reach the final categories shown. Panel (c) shows successive interactions with a sponge, rubber duck, and walnut, followed by the displayed property rankings.