Collaborative Representation Learning for Alignment of Tactile, Language, and Vision Modalities

Yiyun Zhou[1*], **Mingjing Xu**[2*], **Jingwei Shi**[3*], **Quanjiang Li**[4], **Jingyuan Chen**[1†]

[1]Zhejiang University
[2]Swansea University
[3]Shanghai University of Finance and Economics
[4]National University of Defense Technology
{yiyunzhou, jingyuanchen}@zju.edu.cn, mingjing.xu@swansea.ac.uk, shijingwei@stu.sufe.edu.cn, liquanjiang@nudt.edu.cn

Abstract

Tactile sensing offers rich and complementary information to vision and language, enabling robots to perceive fine-grained object properties. However, existing tactile sensors lack standardization, leading to redundant features that hinder cross-sensor generalization. Moreover, existing methods fail to fully integrate the intermediate communication among tactile, language, and vision modalities. To address this, we propose TLV-CoRe, a CLIP-based Tactile-Language-Vision Collaborative Representation learning method. TLV-CoRe introduces a ***Sensor-Aware Modulator*** **to unify tactile features across different sensors** and employs **tactile-irrelevant decoupled learning to disentangle irrelevant tactile features**. Additionally, a ***Unified Bridging Adapter*** **is introduced to enhance tri-modal interaction within the shared representation space**. To fairly evaluate the effectiveness of tactile models, we further propose the RSS evaluation framework, focusing on Robustness, Synergy, and Stability across different methods. Experimental results demonstrate that TLV-CoRe significantly improves sensor-agnostic representation learning and cross-modal alignment, offering a new direction for multimodal tactile representation.

Extended version — https://arxiv.org/abs/2511.11512

1 Introduction

Tactile is one of the essential senses of human perception. Through tactile interaction, we can sense both static and dynamic attributes of objects (*e.g.*, material texture, roughness, and hardness), many of which are too subtle to be reliably perceived by other perception systems like vision (Cheng et al. 2024; Shi et al. 2025; Dave, Lygerakis, and Rueckert 2024; Li et al. 2025a,b; Lv et al. 2025b). In recent years, researchers have been striving to help robots understand the complex and realistic physical world by designing high-resolution tactile sensors (Yuan, Dong, and Adelson 2017; Donlon et al. 2018; Lambeta et al. 2020; Inc. 2020; Zhang et al. 2024) comparable to human touch and collecting large-scale indoor and outdoor tactile image datasets (Yang et al. 2022; Kerr et al. 2022a; Fu et al. 2024; Yu et al. 2024; Feng et al. 2025).

However, **tactile sensors are not yet fully standardized**. Due to external factors (*e.g.*, camera type, lighting position, color, and illumination), tactile images can differ significantly even under identical touch object conditions (Fig. 1 `(i)` and `(ii)`). To address these variations, previous studies (Yang et al. 2024; Feng et al. 2025) have borrowed the concept of positional encoding from language models (Su et al. 2024; Zhao et al. 2023; Jiang et al. 2025a), introducing learnable tokens to model sensor-specific characteristics. However, **these methods overlook a crucial fact: even when touch objects differ noticeably, the styles of the tactile images can still be quite similar** (Fig. 1 `(iii)`), which poses a challenge for tactile models to disentangle tactile-irrelevant features.

Vision and language are also core channels for human-environment interaction. In real-world tasks, their integration has been extensively studied. A large body of work (Radford et al. 2021; Jia et al. 2021; Alayrac et al. 2022; Li et al. 2023) has successfully built semantic bridges between visual and linguistic modalities through contrastive learning (Oord, Li, and Vinyals 2018), achieving remarkable progress. This success has since extended to additional modalities, including audio, point clouds, event etc (Girdhar et al. 2023; Guo et al. 2023; Wang et al. 2024; Lyu et al. 2024). Despite the flourishing development of multimodal learning catalyzed by vision-language pretraining, **the tactile modality remains significantly underexplored**.

Recent research on tactile-language-vision learning has focused on representation learning based on CLIP (Radford et al. 2021; Lv et al. 2025c; Cherti et al. 2023). For instance, TLV-Link (Cheng et al. 2024), designed specifically for the GelSight sensor (Yuan, Dong, and Adelson 2017), trains a tactile encoder via curriculum learning to achieve effective tri-modal alignment. AnyTouch (Feng et al. 2025) proposes a unified representation learning framework for static-dynamic and multi-sensor tactile data, employing masked modeling, self-supervised multimodal alignment, and cross-

*These authors contributed equally.
†Corresponding author.

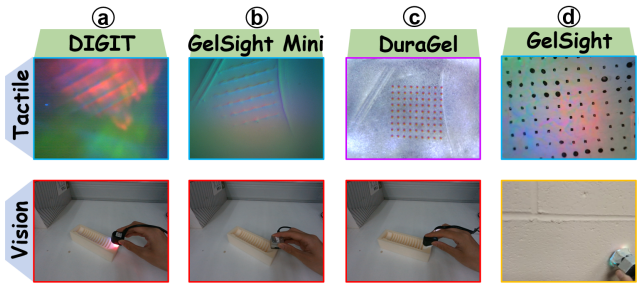

Figure 1: Three properties of heterogeneous sensors are identified: (i). Tactile sensors lack full standardization, leading to significant tactile images variation (Yang et al. 2024). (ii). Tactile images from the identical touch object can differ inconsistently (*e.g.*, ⓐ and ⓑ are similar, both differing greatly from ⓒ). (iii). Despite different touch objects, tactile images may share a consistent style (*e.g.*, ⓐ and ⓑ resemble ⓓ in a dark tone tinged with red).

sensor matching to improve generalization across different sensors. Vit-Lens-2 (Lei et al. 2024) introduces a generic multimodal encoding approach that first transforms various modality inputs into intermediate representations using lightweight modules and then feeds them into a frozen pretrained ViT (Dosovitskiy et al. 2020), enabling efficient representation learning. However, these methods face two challenges: (1) Most methods adapt LoRA (Hu et al. 2022; Lv et al. 2025a; Zhou, Yao, and Chen 2025) within single modality branches, **without explicitly modeling the synergy among the three modalities before fusion, limiting their deep fusion capability**. (2) **There is a lack of standardized evaluation settings** (*e.g.*, base models, batch sizes), making fair comparisons difficult.

To this end, we propose a CLIP-based **T**actile-**L**anguage-**V**ision **Co**llaborative **Re**presentation (**TLV-CoRe**) learning method, aiming to learn sensor-agnostic tactile representations. Unlike existing methods (Yang et al. 2024; Feng et al. 2025), TLV-CoRe introduces a learnable *Sensor-Aware Modulator* (SAM) that adaptively maps tactile features from different sensors into a unified parameter space, combined with tactile-irrelevant decoupled learning to effectively disentangle tactile-irrelevant features. Furthermore, to strengthen cross-modal collaboration in intermediate representations, TLV-CoRe incorporates a *Unified Bridging Adapter* (UBA) into three modality encoders. UBA consists of modality-specific projection layers to learn individual representations and a shared feature space mapping layer to facilitate tri-modal alignment.

For evaluation, we propose a **RSS** evaluation framework, designed to analyze the **R**obustness, **S**ynergy, and **S**tability of various tactile representation learning methods. We define three evaluation protocols—intra-sensor evaluation, cross-sensor generalization, and multi-sensor generalization—to assess the **robustness** of different methods. Given that **multimodal alignment should enhance rather than compromise individual modality performance** (Wang et al. 2022; Jiang et al. 2025b; Wu et al. 2024; Li et al. 2025c; Dufumier et al. 2024), we introduce modal cross-evaluagtion tasks (especially between tactile and vision modalities) to assess **synergy** of various modal encoders. We also investigate the impact of batch size on model **stability**, as varying batch sizes affect the number of negative samples in contrastive learning. We encourage future research to adopt the RSS framework for comprehensive comparisons of different multimodal tactile methods based on CLIP.

Our key contributions are as follows:

- We design a *Sensor-Aware Modulator* that enables flexible learning of unified tactile representations across multiple sensors and introduce tactile-irrelevant decoupled learning to effectively disentangle tactile-irrelevant features.
- We propose a novel *Unified Bridging Adapter*, which includes separate projection layers for tactile, language, and vision encoders, as well as a shared projection to better align their representations.
- We provide a rigorous theoretical analysis of robustness, synergy and stability of our proposed method, providing valuable insights to guide the design of future tactile representation methods.
- We introduce a fair and comprehensive RSS evaluation framework to systematically analyze the robustness, synergy, and stability of other tactile representation learning methods, and verify the effectiveness of the proposed TLV-CoRe.

Note that **the proposed RSS evaluation framework requires consistency in the base model and batch size to ensure that the evaluation focuses more on the differences in the design of the tactile representation methods**.

2 Related Works

In this section, we summarize previous research on multimodal alignment and tactile representation learning. Detailed discussions and comprehensive comparisons of related works are provided in the extended version.

3 Methodology

We introduce TLV-CoRe, a method for learning collaborative representations that align tactile (T), visual (V), and language (L) modalities in a shared latent space (see Fig. 2). TLV-CoRe includes modality-specific encoders and two key modules: the *Sensor-Aware Modulator* (SAM) in the tactile branch and the *Unified Bridging Adapter* (UBA) for cross-modal alignment. SAM removes sensor-specific biases (Zhao et al. 2024; Yang et al. 2024), ensuring invariant tactile representations across sensors. UBA projects features from each modality into a common latent space for alignment. The model is trained with symmetric contrastive losses and a sensor-invariance loss for consistency across sensors. Theoretical guarantees for TLV-CoRe's properties are provided, with detailed proofs in the extended version.

3.1 Tactile Encoder and Sensor-Aware Modulator

The tactile encoder $\mathcal{E}_T$ processes raw tactile inputs x^T and produces a d-dimensional latent representation. Specifically,

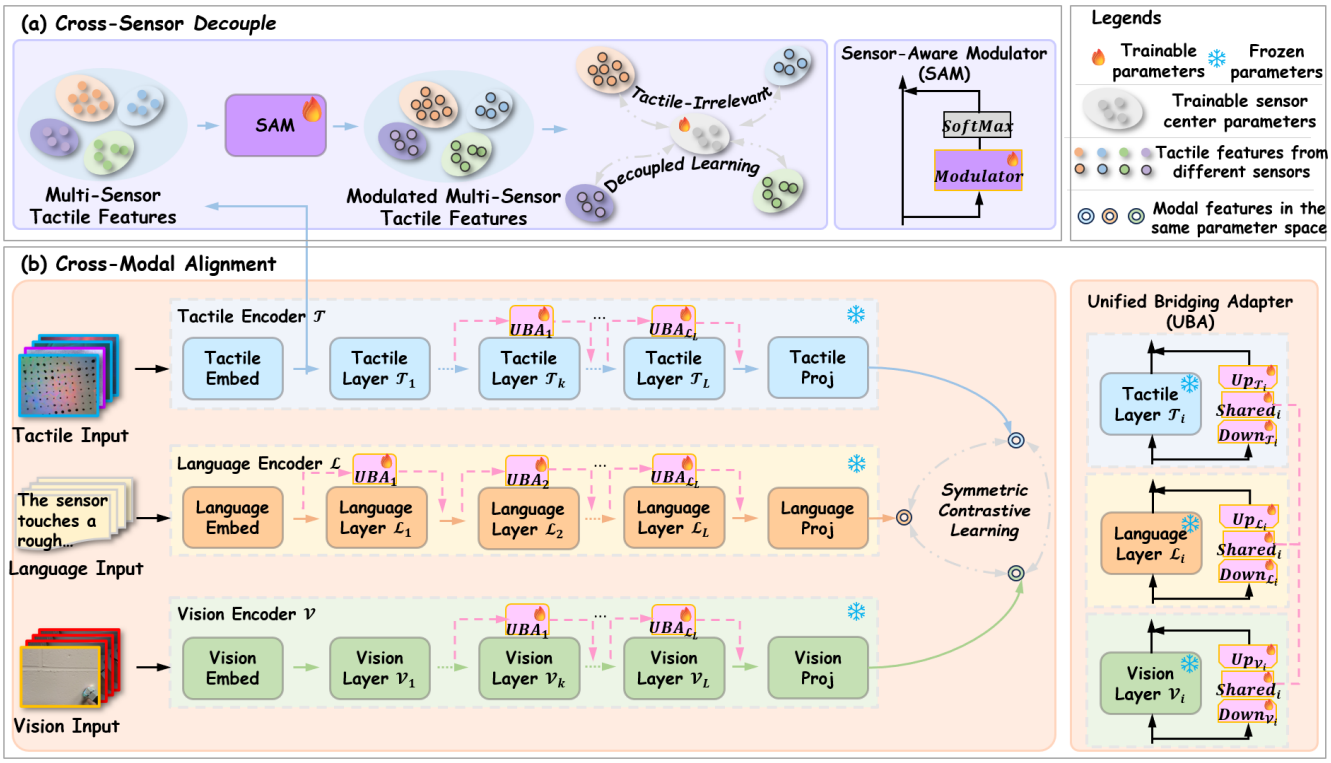

Figure 2: Overview of TLV-CoRe, which consists of modality-specific encoders for tactile, visual, and language modalities inputs, a *Sensor-Aware Modulator* to remove sensor-specific biases, and a *Unified Bridging Adapter* for alignment.

we implement $\mathcal{E}_T$ as a Vision Transformer (ViT) (Dosovitskiy et al. 2020). The tactile image is divided into non-overlapping patches, each patch is linearly embedded, and the resulting sequence is processed through standard transformer blocks. This design enables the encoder to capture spatial patterns and tactile features.

To **address sensor-specific biases**, we incorporate a *Sensor-Aware Modulator* (SAM) into the tactile branch. Let $s \in \{1, \ldots, S\}$ index the sensor that captured a given sample. Given a tactile feature $h^T = \mathcal{E}_T(x^T) \in \mathbb{R}^d$, the SAM computes a routing-weight vector $r(h^T) \in \mathbb{R}^S$ via a learnable linear mapping and softmax: $r(h^T) = \mathrm{softmax}(W_r h^T)$, where $W_r \in \mathbb{R}^{S \times d}$. The s-th component $[r(h^T)]_s$ indicates the importance of sensor s. We then modulate the feature by

$$h^T_{\mathrm{mod}} = h^T + [r(h^T)]_s \, h^T, \tag{1}$$

which scales the feature according to the predicted sensor weight.

However, **tactile images captured under similar conditions can exhibit similar patterns across different sensors** (Ou, Chen, and Luo 2024). In such cases, the **SAM inadvertently clusters features by sensor identity rather than by underlying tactile content**. To address this, we employ **tactile-irrelevant decoupled learning that explicitly removes redundant information from the tactile representation**.

Specifically, we adversarially train the tactile encoder with a sensor-classification objective. We introduce a set of learnable sensor centroids $\{c_s\}_{s=1}^S \subset \mathbb{R}^d$. For a feature h^T, we compute its similarity to each centroid and define

$$p(s \mid h^T) = \frac{\exp\big(\langle h^T, c_s \rangle / \tau\big)}{\sum_{s'=1}^{S} \exp\big(\langle h^T, c_{s'} \rangle / \tau\big)}, \tag{2}$$

where $\langle \cdot, \cdot \rangle$ denotes cosine similarity and $\tau > 0$ (*e.g.*, 0.05) is a temperature. We then minimize the expected negative log-likelihood,

$$\mathcal{L}_{\mathrm{DL}} = -\mathbb{E}_{(x^T, s)} \big[\log p(s \mid h^T)\big], \tag{3}$$

and apply **a gradient reversal layer so that $\mathcal{E}_T$ learns to confuse the sensor classifier**. This adversarial training removes redundant information from h^T. **Combined with the SAM, this encourages the tactile features to capture intrinsic object properties rather than sensor artifacts**.

3.2 Unified Bridging Adapter for Cross-Modal Alignment

To **enable direct interactions between modalities**, we introduce a *Unified Bridging Adapter* (UBA) in each branch. The UBA is a lightweight module that projects modality-specific features into a shared latent space. Formally, for each modality $m \in \{\mathrm{T}, \mathrm{V}, \mathrm{L}\}$, we define two projection matrices: $W^m_{\downarrow} \in \mathbb{R}^{r \times d}$ to down-project from d to a lower dimension r ($r \ll d$), and $W^m_{\uparrow} \in \mathbb{R}^{d \times r}$ to up-project back to d. We also introduce a shared transformation $W_{\mathrm{sh}} \in \mathbb{R}^{r \times r}$ common to all modalities: given a feature $h^m \in \mathbb{R}^d$ from

modality m, we project it into the shared space, apply the shared transform, and project it back:

$$z^m_{\text{shared}} = W_{\text{sh}}\big(W^m_{\downarrow} h^m\big), \qquad \Delta h^m = W^m_{\uparrow} z^m_{\text{shared}}. \quad (4)$$

We then form the aligned feature by adding this residual: $h^m_{\text{aligned}} = h^m + \Delta h^m$.

Thus, h^m_{aligned} is a coordinated version of h^m that has passed through the common latent space. By sharing W_{sh} across modalities, we ensure that the transformations at the bottleneck are identical for tactile, vision, and language streams, while the modality-specific matrices $W^m_{\downarrow}$ and $W^m_{\uparrow}$ allow each modality to interface with this shared space.

Practical UBA placement. In our implementation we attach *exactly* $L = 12$ UBA blocks to *each* modality so that every sample—regardless of its source—passes through the *same number* of shared transformations before the final projection. Concretely, the language encoders comprise 12 transformer layers; we attach one UBA per layer. The vision/tactile branch is deeper (CLIP design (Cherti et al. 2023) with 24 layers), so we leave **the first** 12 **layers *unbridged* to preserve low-level visual/tactile primitives and attach UBAs only to the upper** 12 **layers where high-level semantics emerge**.

At layer ℓ, the outputs of each modality's transformer block are fed via its UBA and merged via residual addition. This multi-level UBA design ensures alignment occurs progressively at different semantic levels. After the final layer, we obtain aligned $h^T_{\text{aligned}}, h^V_{\text{aligned}}, h^L_{\text{aligned}}$, which we L_2-normalize to yield final embeddings $z^T, z^V, z^L \in \mathbb{R}^d$ for cross-modal comparisons.

3.3 Cross-Modal Contrastive Learning

We train TLV-CoRe with symmetric contrastive losses for each pair of modalities and the sensor-invariance loss. For each pair $(X, Y) \in \{(\mathrm{T},\mathrm{V}), (\mathrm{T},\mathrm{L}), (\mathrm{V},\mathrm{L})\}$, we use a symmetric InfoNCE contrastive loss (Oord, Li, and Vinyals 2018). Given a batch of N aligned triplets $\{(z^T_i, z^V_i, z^L_i)\}_{i=1}^N$, the loss for the tactile–vision pair (as an example) is:

$$\begin{aligned}\mathcal{L}_{T\text{-}V} \;=\; -\frac{1}{2N}\sum_{i=1}^{N}\Big[&\log\frac{\exp\big(\langle z^T_i, z^V_i\rangle/\tau\big)}{\sum_{j=1}^{N}\exp\big(\langle z^T_i, z^V_j\rangle/\tau\big)} \\ &+\; \log\frac{\exp\big(\langle z^V_i, z^T_i\rangle/\tau\big)}{\sum_{j=1}^{N}\exp\big(\langle z^V_i, z^T_j\rangle/\tau\big)}\Big],\end{aligned} \quad (5)$$

where $\tau > 0$ (*e.g.*, 0.05) is a temperature parameter. The first term aligns each z^T_i with its paired z^V_i, and the second term does the reverse, making the loss symmetric. We define analogous losses $\mathcal{L}_{T\text{-}L}$ and $\mathcal{L}_{V\text{-}L}$ for the tactile–language and vision–language pairs. The total alignment loss is $\mathcal{L}_{\text{SCL}} = \mathcal{L}_{T\text{-}V} + \mathcal{L}_{T\text{-}L} + \mathcal{L}_{V\text{-}L}$.

The overall training objective combines the alignment loss with the sensor-invariance loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{SCL}} + \lambda_{\text{DL}}\,\mathcal{L}_{\text{DL}}, \quad (6)$$

where $\lambda_{\text{DL}} > 0$ **balances the influence of sensor invariance**. We minimize $\mathcal{L}_{\text{total}}$ end-to-end (applying a gradient reversal layer to $\mathcal{L}_{\text{DL}}$). This objective trains the model to produce modality-agnostic representations: in the shared embedding space, matching tactile, visual, and textual inputs are embedded closely, while non-matching inputs are far apart. The sensor-invariance loss further ensures that tactile embeddings remain consistent across different sensors.

3.4 Theoretical Guarantees

We analyze TLV-CoRe's properties under standard optimization assumptions (Patel, Zhang, and Tian 2022; Xu et al. 2024; Lei et al. 2019; Pham et al. 2020). Let $\mathcal{L}(\Theta)$ denote the training objective. We make the following assumptions to facilitate analysis:

Assumption 3.1 (Smoothness)**.** The gradient of $\mathcal{L}(\Theta)$ is L-Lipschitz continuous. Formally, for all parameters Θ, Θ', $\|\nabla\mathcal{L}(\Theta) - \nabla\mathcal{L}(\Theta')\| \le L\,\|\Theta - \Theta'\|$.

Assumption 3.2 (Polyak–Łojasiewicz (PL) Condition (Karimi, Nutini, and Schmidt 2016))**.** In a neighborhood of a local optimum Θ^*, the loss satisfies

$$\mathcal{L}(\Theta) - \mathcal{L}(\Theta^*) \;\ge\; \frac{\mu}{2}\,\|\Theta - \Theta^*\|^2 \quad \text{for all } \Theta \text{ near } \Theta^*.$$

Assumption 3.3 (Bounded Gradient Variance)**.** The stochastic gradient has bounded variance. Specifically, $\mathbb{E}\big\|\nabla\mathcal{L}_{\mathcal{B}}(\Theta) - \nabla\mathcal{L}(\Theta)\big\|^2 \le \sigma^2$, where $\nabla\mathcal{L}_{\mathcal{B}}(\Theta)$ denotes the gradient on a mini-batch $\mathcal{B}$.

Convergence Analysis Under Assumptions 3.1–3.3, we obtain:

Theorem 3.1 (Convergence Rate)**.** *Suppose Assumptions 3.1–3.3 hold, and let Θ^* be a local minimizer satisfying the PL condition. Running SGD with step size $\eta < 2/L$ gives:*

$$\mathbb{E}\|\Theta_t - \Theta^*\|^2 \;\le\; (1 - \eta\mu\beta)^t\|\Theta_0 - \Theta^*\|^2 \;+\; \frac{\eta\sigma^2}{\mu\beta}, \quad (7)$$

where $\beta = 1/(1 + \kappa(W_{\text{sh}}))$ and $\kappa(W_{\text{sh}})$ is the condition number of the shared UBA matrix.

Theorem 3.1 shows that sharing the UBA across modalities accelerates convergence by improving the effective condition number.

Robustness via Sensor-Invariance We next examine how removing sensor-specific information via the SAM affects training robustness.

Lemma 3.2 (Gradient Variance Reduction)**.** *Let $\mathcal{I}(h^T; s)$ denote the mutual information between the tactile representation h^T and the sensor identity s. As the model removes sensor-specific features, the variance of the stochastic gradient satisfies*

$$\mathrm{Var}[\nabla\mathcal{L}(\Theta)] \;\le\; \sigma_0^2 \;-\; \gamma\,\mathcal{I}(h^T; s), \quad (8)$$

for constants $\sigma_0^2, \gamma > 0$. Hence, reducing $\mathcal{I}(h^T; s)$ (via the SAM) lowers the gradient variance.

Proposition 3.3 (Optimization Robustness)**.** *Since the SAM drives $\mathcal{I}(h^T; s) \to \varepsilon$ (with $\varepsilon \ge 0$ small), the asymptotic gradient variance is bounded by*

$$\limsup_{t\to\infty} \mathrm{Var}[\nabla\mathcal{L}(\Theta_t)] \;\le\; \sigma_0^2 - \gamma(1 - \varepsilon). \quad (9)$$

Thus, as sensor-specific information is eliminated, the training gradients become more stable.

Lemma 3.2 and Proposition 3.3 explain ours training robustness: removing sensor-specific signals via the SAM reduces stochastic gradient noise and leads to more stable optimization.

Cross-Modal Synergy The UBA also enables information transfer across modalities, under the following assumption:

Assumption 3.4 (Shared and Unique Information)**.** Each modality $m \in \{T, V, L\}$ encodes information about the task label Y, with components unique to m and components shared across modalities.

Theorem 3.4 (Cross-Modal Information Transfer)**.** *Under Assumption 3.4, aligning modality m with modality m' via the UBA increases its label information. Formally,*

$$\begin{aligned} &\mathcal{I}\big(h^m_{\text{aligned}}; Y\big) \;\geq\; \mathcal{I}\big(h^m; Y\big) \\ &\quad+\; \alpha \,\min\big\{\, r,\; \mathcal{I}\big(h^{m'}; Y\big) - \mathcal{I}\big(h^m; Y\big)\,\big\}, \end{aligned} \tag{10}$$

where r is the dimension of the UBA's shared subspace and $\alpha \in (0, 1)$ is a constant. Thus, h^m_{aligned} can gain up to αr bits of information that modality m' has but m lacks.

Corollary 3.5 (Cross-Modal Performance)**.** *Let $\mathcal{A}_m^{\text{task}_{m'}}$ be the accuracy of encoder m on tasks of modality m'. Then under Theorem 3.4,*

$$\mathcal{A}_m^{\text{task}_{m'}} \;\geq\; \mathcal{A}_{m'}^{\text{task}_{m'}} \;-\; \Delta_{m,m'} \;-\; \frac{C}{r}, \tag{11}$$

where $\Delta_{m,m'}$ is a small modality-gap term and C is a task-dependent constant. Hence, as the shared dimension r grows, the cross-modal performance of modality m approaches that of the best modality m' up to a small gap.

These results imply that multi-level UBA alignment allows each modality to absorb useful information from the others, thereby improving its performance on cross-modal tasks. In practice, we indeed observe consistent performance gains in modal cross-evaluation (see Sec. 5.2).

Batch-Size Stability We consider the effect of batch size in training. In a batch of size N, an anchor typically encounters $\mathbb{E}[N_{\text{sim}}] = (N-1)p_{\text{sim}}$ semantically similar negatives on average, where p_{sim} is the probability of semantic overlap. Thus there is a trade-off: **Small batches:** fewer negatives (weaker contrastive signal) but emphasize fine-grained distinctions. **Large batches:** many negatives (stronger alignment signal) but may bias toward coarser features.

Theorem 3.6 (Batch-Size Stability)**.** *Let ϵ_N be the expected task error when using batch size N. Under sensor-invariance decoupling, the error gap between any two batch sizes satisfies*

$$|\epsilon_N - \epsilon_{N'}| \;\leq\; \frac{C_1}{1 + C_2\big(1 - \mathcal{I}(h^T; s)\big)}, \tag{12}$$

for constants $C_1, C_2 > 0$. As $\mathcal{I}(h^T; s)$ decreases (via the SAM), this bound shrinks, making performance less sensitive to the choice of N.

Proposition 3.7 (Representation Enhancement)**.** *After UBA alignment, each representation satisfies*

$$\begin{aligned} &\mathcal{I}\big(h^m_{\text{aligned}}; Y\big) \;\geq\; \mathcal{I}\big(h^m; Y\big) \\ &\quad+\; \max_{m' \neq m} \big[\mathcal{I}\big(h^{m'}; Y\big) \;-\; \mathcal{I}\big(h^m; Y\big)\big]_+, \end{aligned} \tag{13}$$

so it gains any extra information present in the most informative other modality.

Theorem 3.6 and Proposition 3.7 together explain TLV-CoRe's stability:

- **Tactile Decoupling** ensures the tactile encoder focuses on invariant, task-relevant features, reducing spurious correlations and enabling stable learning even with small batches.
- **UBA Sharing** ensures each encoder learns fine-grained and global features by absorbing information from other modalities, making the learned representations robust to batch-size variations.

In summary, our theoretical analysis demonstrates that sensor-aware decoupling and shared adapters yield benefits in convergence, cross-modal transfer, and training stability. They align with our empirical performance findings (see Sec. 5.3).

Note that **we provide the proposed TLV-CoRe's empirical validation of convergence analysis and more detailed theoretical analysis in the extended version**.

4 Experiments

Our experiments, under the RSS framework, focus on tactile representations using a linear probing approach (Cheng et al. 2024; Feng et al. 2025; Lei et al. 2024). We evaluate tactile representations across three protocols: intra-sensor, cross-sensor, and multi-sensor generalization, for material property identification and robot grasping prediction.

5 Experimental Setup

In this section, we evaluate the effectiveness of TLV-CoRe across multiple datasets and tasks, comparing against state-of-the-art tactile representation learning approaches. Detailed information on the datasets, baselines, and implementation details is provided in the extended version.

5.1 Main Results ⇒ Robustness Analysis

We systematically compare various methods using three evaluation protocols, as shown in Table 1. Overall, among these CLIP-based approaches, the proposed **TLV-CoRe consistently outperforms existing methods with significantly fewer trainable parameters**. The following observations emerge:

- Across the three evaluation protocols, the overall performance ranks as: intra-sensor evaluation $>$ cross-sensor generalization $\approx$ multi-sensor generalization. This trend primarily stems from the fact that intra-sensor evaluation is an in-distribution test, where the patterns learned by the model closely match those in test set, leading to better generalization. In contrast, in the out-of-distribution settings of cross- and multi-sensor generalization, the model

Training Data	Method	%Param	TAG			OF 1.0	OF 2.0	Feel
			Material	Roughness	Hardness	Material	Material	Grasp
	CLIP (Cherti et al. 2023)	-	52.73	82.16	85.32	41.15	72.97	72.52
TAG	TLV-Link† (Cheng et al. 2024)	1.23	53.26	84.80	85.94	43.75	74.12	76.01
	AnyTouch (Feng et al. 2025)	1.31	61.48	86.31	85.32	43.88	75.20	80.53
	VIT-LENS-2 (Lei et al. 2024)	7.00	**65.99**	87.16	91.08	37.00	75.85	-
	TLV-CoRe	0.30	65.44	**88.81**	**92.65**	**49.12**	**76.28**	**81.28**
SSVTP	TLV-Link	1.23	55.52	84.63	86.32	36.38	75.45	74.88
	AnyTouch	1.31	62.49	67.19	73.93	40.12	71.46	68.26
	VIT-LENS-2	7.00	48.95	**86.91**	83.75	35.38	75.00	-
	TLV-CoRe	0.30	**63.25**	85.39	**86.78**	**48.50**	**75.74**	**75.39**
TVL	TLV-Link	1.23	51.14	80.00	84.33	40.50	75.38	76.06
	AnyTouch	1.31	46.18	84.39	73.74	41.88	75.41	77.57
	VIT-LENS-2	7.00	52.64	82.16	80.53	38.26	76.24	-
	TLV-CoRe	0.30	**54.47**	**84.54**	**84.47**	**45.13**	**77.89**	**77.95**
Octopi*	TLV-Link	1.23	48.72	79.55	81.97	47.12	73.58	74.99
	AnyTouch	1.31	44.39	**86.36**	81.13	38.12	73.11	79.07
	VIT-LENS-2	7.00	48.11	82.02	84.36	39.62	75.13	-
	TLV-CoRe	0.30	**52.65**	85.83	**86.43**	**48.88**	**75.86**	**80.63**
TacQuad	TLV-Link	1.23	56.60	83.93	87.37	37.25	**76.53**	76.12
	AnyTouch	1.31	45.14	84.12	80.61	41.62	74.29	80.42
	VIT-LENS-2	7.00	47.50	85.94	84.44	39.62	75.25	-
	TLV-CoRe	0.30	**58.37**	**86.80**	**87.52**	**42.25**	75.91	**80.77**
TAG,SSVTP, Octopi, TVL, TacQuad	TLV-Link	1.23	54.82	84.53	86.78	42.64	75.58	76.39
	AnyTouch	1.31	56.43	85.72	84.31	44.12	76.50	79.24
	VIT-LENS-2	7.00	57.16	84.58	84.69	42.63	76.62	-
	TLV-CoRe	0.30	**60.26**	**86.53**	**87.13**	**47.25**	**76.87**	**79.35**

Table 1: Performance (%) comparison of different methods under three evaluation protocols: intra-sensor evaluation, cross-sensor generalization, and multi-sensor generalization. † Note that we follow the default configuration of the TLV-Link repository, applying LoRA (Hu et al. 2022) to fine-tune the tactile and vision encoders while keeping the language encoder frozen. *For methods that cannot handle missing modalities, the tactile modality is used as a substitute for the missing vision modality in the Octopi dataset.

may struggle to interpret previously unseen features effectively, resulting in performance drops that cannot be easily mitigated by simply increasing the amount of training data.

- Although Octopi and TAG use the same type of GelSight sensors, models trained on Octopi generally underperform on the three TAG sub-tests. We attribute this to Octopi's lack of visual modality and limited data size, which leads to insufficient training.
- The performance differences among existing methods remain relatively minor. VIT-LENS-2 performs better in some cases, possibly due to its larger number of trainable parameters. However, in general, these methods still lack robust performance across all three evaluation protocols.
- In contrast, the proposed TLV-CoRe consistently achieves leading results across all tasks, excelling particularly in material classification. While it may not always lead in roughness binary classification—likely due to the higher randomness of simpler tasks—it reliably ranks within the top two, with only a minimal margin from the best. This further validates the robustness and reliability of TLV-CoRe.

5.2 Tactile × Vision ⇒ Synergy Analysis

The **goal of cross-modal alignment is to achieve information complementarity between modalities, rather than sacrificing the representation quality of each modality** (Wang et al. 2022; Dufumier et al. 2024). In other words, the aim is to enable synergy across modalities. To evaluate how well different methods achieve this synergy between tactile and vision modalities, we design a modal cross-evaluation. We select models trained on the TAG, Octopi, and TacQuad datasets in Sec. 5.1 as the evaluation subjects. For tactile tasks, we choose three subtasks from the TAG dataset and evaluate them using the corresponding models' vision encoders. For vision tasks, we use three image classification datasets—CIFAR-10, CIFAR-100 (Krizhevsky and Hinton 2009), and ImgNetDogs (Cukierski 2017)—and evaluate them with the tactile encoders. Table 2 shows the following observations: (1) In tactile tasks, vision encoders

Training Data	Method	TAG			CIFAR-10	CIFAR-100	ImgNetDogs
		Material	Roughness	Hardness	Image CLS	Image CLS	Image CLS
TAG	TLV-Link	52.91	82.69	85.17	32.39	10.88	25.09
	AnyTouch	53.64	84.52	85.19	40.90	18.18	25.97
	VIT-LENS-2	53.32	85.94	86.13	44.38	19.97	28.90
	TLV-CoRe	**53.86**	**87.39**	**88.62**	**68.15**	**34.22**	**30.07**
Octopi*	TLV-Link	48.72	79.55	81.97	59.35	31.22	28.11
	AnyTouch	44.39	**86.36**	81.13	38.64	13.88	24.11
	VIT-LENS-2	48.11	82.02	84.36	43.23	17.62	26.75
	TLV-CoRe	**52.65**	85.83	**86.43**	**70.46**	**37.41**	**28.25**
TacQuad	TLV-Link	54.15	84.62	85.97	76.77	52.39	29.27
	AnyTouch	50.37	84.50	82.19	50.67	25.29	26.26
	VIT-LENS-2	51.29	85.26	84.72	43.18	18.76	26.56
	TLV-CoRe	**56.52**	**85.97**	**86.47**	**78.90**	**52.70**	**31.47**

Table 2: Performance (%) comparison of different methods in modal cross-evaluation. *Since the visual input is replaced by tactile images, the vision encoder is equivalent to the tactile encoder.

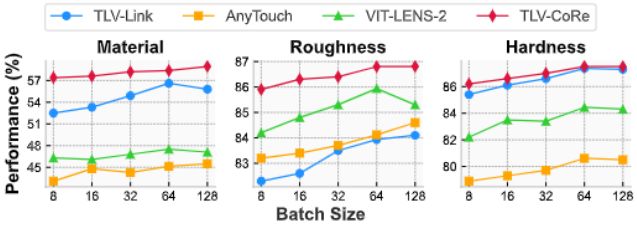

Figure 3: Performance (%) comparison of different methods across various batch sizes.

generally perform worse than tactile encoders, which is expected given their modality focus. (2) In vision tasks, most methods demonstrate limited performance, possibly due to the setting of only 50 linear probing training epochs, which may indicate that a model with only a learnable classification head may not have sufficiently learned the new visual information. (3) Notably, the proposed TLV-CoRe achieves the best performance across both types of tasks. We attribute this to the introduction of the *UBA* module, which bridges the modality branches via parameter sharing. This facilitates a unified feature space, enabling smooth gradient flow and efficient information transfer and complementarity across modalities, while also accelerating convergence. These results further validate the analysis in Sec. 3.4.

5.3 Different Batch Sizes ⇒ Stability Analysis

In CLIP-based contrastive methods, different batch sizes lead to significant variations in the number of negative samples. As illustrated in Fig.3, batch size significantly affects the performance of various methods trained by TacQuad dataset, evaluated across three subsets of TAG dataset. Notably, larger batch sizes tend to yield better performance, which aligns with observations from previous studies (Chen et al. 2020; Kerr et al. 2022b). However, we also observe that when the batch size reaches 128, the performance of TLV-Link and VIT-LENS-2 no longer improves and may even decline. This is because a larger batch contains more data from different sensors, making it harder for single-sensor methods

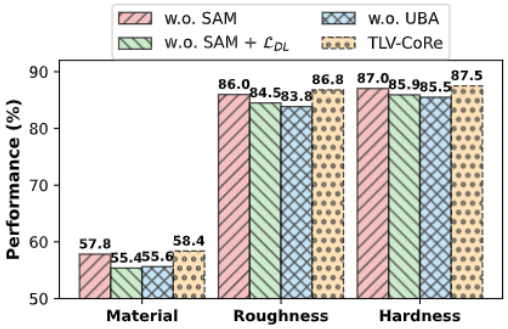

Figure 4: Ablation experiments on various components.

to handle the increased heterogeneity in tactile data. Overall, the proposed TLV-CoRe exhibits a smoother performance curve and demonstrates superior stability. This observation is consistent with the analysis presented in Sec. 3.4.

5.4 Ablation Study

We perform an ablation study on the components of TLV-CoRe trained by TacQuad. As shown in Fig. 4, removing the SAM leads to a performance drop, and eliminating the decoupled loss $\mathcal{L}_{DL}$ causes an even greater decline, highlighting the crucial role of decoupled learning in handling multi-sensor data. Additionally, removing the UBA module results in the worst performance, underscoring its critical role in the overall architecture.

6 Conclusion

In this paper, we present TLV-CoRe, a collaborative representation learning method for tactile, language, and vision modalities. TLV-CoRe introduces a *Sensor-Aware Modulator* to unify tactile representations across various sensors, employs tactile-irrelevant decoupled learning to disentangle tactile-irrelevant features, and incorporates a *Unified Bridging Adapter* to enhance tri-modal interaction. To support fair evaluation, we propose the RSS framework. Experimental results show that TLV-CoRe achieves strong performance.

Acknowledgments

This research was supported by grants from the "Pioneer" and "Leading Goose" R&D Program of Zhejiang (2025C02022) and National Natural Science Foundation of China (No.62307032). Additionally, this work was partially supported by the UK Engineering and Physical Sciences Research Council (EPSRC) [EP/W524694/1].

References

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35: 23716–23736.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PmLR.

Cheng, N.; Guan, C.; Gao, J.; Wang, W.; Li, Y.; Meng, F.; Zhou, J.; Fang, B.; Xu, J.; and Han, W. 2024. Touch100k: A large-scale touch-language-vision dataset for touch-centric multimodal representation. *arXiv preprint arXiv:2406.03813*.

Cherti, M.; Beaumont, R.; Wightman, R.; Wortsman, M.; Ilharco, G.; Gordon, C.; Schuhmann, C.; Schmidt, L.; and Jitsev, J. 2023. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2818–2829.

Cukierski, W. 2017. Dog breed identification. https://kaggle.com/competitions/dog-breed-identification. Kaggle.

Dave, V.; Lygerakis, F.; and Rueckert, E. 2024. Multimodal visual-tactile representation learning through self-supervised contrastive pre-training. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 8013–8020. IEEE.

Donlon, E.; Dong, S.; Liu, M.; Li, J.; Adelson, E.; and Rodriguez, A. 2018. Gelslim: A high-resolution, compact, robust, and calibrated tactile-sensing finger. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 1927–1934. IEEE.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Dufumier, B.; Castillo-Navarro, J.; Tuia, D.; and Thiran, J.-P. 2024. What to align in multimodal contrastive learning? *arXiv preprint arXiv:2409.07402*.

Feng, R.; Hu, J.; Xia, W.; Gao, T.; Shen, A.; Sun, Y.; Fang, B.; and Hu, D. 2025. AnyTouch: Learning unified static-dynamic representation across multiple visuo-tactile sensors. arXiv:2502.12191.

Fu, L.; Datta, G.; Huang, H.; Panitch, W. C.-H.; Drake, J.; Ortiz, J.; Mukadam, M.; Lambeta, M.; Calandra, R.; and Goldberg, K. 2024. A touch, vision, and language dataset for multimodal alignment. In *Forty-first International Conference on Machine Learning*.

Girdhar, R.; El-Nouby, A.; Liu, Z.; Singh, M.; Alwala, K. V.; Joulin, A.; and Misra, I. 2023. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15180–15190.

Guo, Z.; Zhang, R.; Zhu, X.; Tang, Y.; Ma, X.; Han, J.; Chen, K.; Gao, P.; Li, X.; Li, H.; et al. 2023. Point-bind & point-llm: Aligning point cloud with multi-modality for 3d understanding, generation, and instruction following. *arXiv preprint arXiv:2309.00615*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Inc., G. 2020. GelSight Mini.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, 4904–4916. PMLR.

Jiang, Z.; Li, K.; Zhou, Y.; Liu, S.; Wang, Z.; Zhang, S.; et al. 2025a. PureKV: Plug-and-Play KV Cache Optimization with Spatial-Temporal Sparse Attention for Vision-Language Large Models. *arXiv preprint arXiv:2510.25600*.

Jiang, Z.; Xu, J.; Zhang, S.; Shen, T.; Li, J.; Kuang, K.; Cai, H.; and Wu, F. 2025b. Fedcfa: Alleviating simpson's paradox in model aggregation with counterfactual federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 17662–17670.

Karimi, H.; Nutini, J.; and Schmidt, M. 2016. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European conference on machine learning and knowledge discovery in databases*, 795–811. Springer.

Kerr, J.; Huang, H.; Wilcox, A.; Hoque, R.; Ichnowski, J.; Calandra, R.; and Goldberg, K. 2022a. Self-supervised visuo-tactile pretraining to locate and follow garment features. *arXiv preprint arXiv:2209.13042*.

Kerr, J.; Huang, H.; Wilcox, A.; Hoque, R.; Ichnowski, J.; Calandra, R.; and Goldberg, K. 2022b. Self-supervised visuo-tactile pretraining to locate and follow garment features. *arXiv preprint arXiv:2209.13042*.

Krizhevsky, A.; and Hinton, G. 2009. Learning multiple layers of features from tiny images. *Master's thesis, Department of Computer Science, University of Toronto*.

Lambeta, M.; Chou, P.-W.; Tian, S.; Yang, B.; Maloon, B.; Most, V. R.; Stroud, D.; Santos, R.; Byagowi, A.; Kammerer, G.; et al. 2020. 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.

Lei, W.; Ge, Y.; Yi, K.; Zhang, J.; Gao, D.; Sun, D.; Ge, Y.; Shan, Y.; and Shou, M. Z. 2024. Vit-lens: Towards omni-modal representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26647–26657.

Lei, Y.; Hu, T.; Li, G.; and Tang, K. 2019. Stochastic gradient descent for nonconvex learning without bounded gradient assumptions. *IEEE transactions on neural networks and learning systems*, 31(10): 4394–4400.

Li, K.; Jiang, Z.; Shen, Z.; ZhaodeWang, Z.; Lv, C.; Zhang, S.; Wu, F.; and Wu, F. 2025a. MadaKV: Adaptive Modality-Perception KV Cache Eviction for Efficient Multimodal Long-Context Inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 13306–13318.

Li, K.; Xiong, Y.; Jiang, Z.; Zhou, Y.; Wang, Z.; Lv, C.; and Zhang, S. 2025b. FlowMM: Cross-Modal Information Flow Guided KV Cache Merging for Efficient Multimodal Context Inference. *arXiv preprint arXiv:2511.05534*.

Li, K.; Zhan, T.; Fu, K.; Zhang, S.; Kuang, K.; Li, J.; Zhao, Z.; Wu, F.; and Wu, F. 2025c. MergeNet: Knowledge Migration across Heterogeneous Models, Tasks, and Modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 4824–4832.

Li, Y.; Fan, H.; Hu, R.; Feichtenhofer, C.; and He, K. 2023. Scaling language-image pre-training via masking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 23390–23400.

Lv, X.; Chen, J.; Li, M.; Sui, Y.; Liu, Z.; and Liao, B. 2025a. Grasp the Key Takeaways from Source Domain for Few Shot Graph Domain Adaptation. In *Proceedings of the ACM on Web Conference 2025*, 2330–2340.

Lv, X.; Li, M.; Chen, J.; Dong, Z.; Han, S.; and Liao, B. 2025b. Out-of-Distribution Detection via LLM-Guided Outlier Generation for Text-attributed Graph. In *Findings of the Association for Computational Linguistics: ACL 2025*, 19544–19555.

Lv, X.; Wang, G.; Chen, J.; Su, H.; Dong, Z.; Zhu, Y.; Liao, B.; and Wu, F. 2025c. Debiased Cognition Representation Learning for Knowledge Tracing. *ACM Transactions on Information Systems*.

Lyu, Y.; Zheng, X.; Zhou, J.; and Wang, L. 2024. Unibind: Llm-augmented unified and balanced representation space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26752–26762.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Ou, N.; Chen, Z.; and Luo, S. 2024. Marker or markerless? mode-switchable optical tactile sensing for diverse robot tasks. *IEEE Robotics and Automation Letters*.

Patel, V.; Zhang, S.; and Tian, B. 2022. Global convergence and stability of stochastic gradient descent. *Advances in Neural Information Processing Systems*, 35: 36014–36025.

Pham, N. H.; Nguyen, L. M.; Phan, D. T.; and Tran-Dinh, Q. 2020. ProxSARAH: An efficient algorithmic framework for stochastic composite nonconvex optimization. *Journal of Machine Learning Research*, 21(110): 1–48.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Shi, J.; Zhang, Z.; Wu, B.; Liang, Y.; Fang, M.; Chen, L.; and Zhao, Y. 2025. Presentagent: Multimodal agent for presentation video generation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 760–773.

Su, J.; Ahmed, M.; Lu, Y.; Pan, S.; Bo, W.; and Liu, Y. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568: 127063.

Wang, D.; Zhao, T.; Yu, W.; Chawla, N. V.; and Jiang, M. 2022. Deep multimodal complementarity learning. *IEEE Transactions on Neural Networks and Learning Systems*, 34(12): 10213–10224.

Wang, Z.; Zhang, Z.; Zhang, H.; Liu, L.; Huang, R.; Cheng, X.; Zhao, H.; and Zhao, Z. 2024. Omnibind: Large-scale omni multimodal representation via binding spaces. *arXiv preprint arXiv:2407.11895*.

Wu, T.; Li, M.; Chen, J.; Ji, W.; Lin, W.; Gao, J.; Kuang, K.; Zhao, Z.; and Wu, F. 2024. Semantic alignment for multimodal large language models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 3489–3498.

Xu, M.; Ju, P.; Liu, J.; and Yang, H. 2024. PSMGD: Periodic stochastic multi-gradient descent for fast multi-objective optimization. arXiv:2412.10961.

Yang, F.; Feng, C.; Chen, Z.; Park, H.; Wang, D.; Dou, Y.; Zeng, Z.; Chen, X.; Gangopadhyay, R.; Owens, A.; et al. 2024. Binding touch to everything: Learning unified multimodal tactile representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26340–26353.

Yang, F.; Ma, C.; Zhang, J.; Zhu, J.; Yuan, W.; and Owens, A. 2022. Touch and go: learning from human-collected vision and touch. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 8081–8103.

Yu, S.; Lin, K.; Xiao, A.; Duan, J.; and Soh, H. 2024. Octopi: Object property reasoning with large tactile-language models. *arXiv preprint arXiv:2405.02794*.

Yuan, W.; Dong, S.; and Adelson, E. H. 2017. Gelsight: High-resolution robot tactile sensors for estimating geometry and force. *Sensors*, 17(12): 2762.

Zhang, S.; Yang, Y.; Sun, F.; Bao, L.; Shan, J.; Gao, Y.; and Fang, B. 2024. A compact visuo-tactile robotic skin for micron-level tactile perception. *IEEE Sensors Journal*.

Zhao, J.; Ma, Y.; Wang, L.; and Adelson, E. H. 2024. Transferable tactile transformers for representation learning across diverse sensors and tasks. *arXiv preprint arXiv:2406.13640*.

Zhao, L.; Feng, X.; Feng, X.; Zhong, W.; Xu, D.; Yang, Q.; Liu, H.; Qin, B.; and Liu, T. 2023. Length extrapolation of transformers: A survey from the perspective of positional encoding. *arXiv preprint arXiv:2312.17044*.

Zhou, Y.; Yao, C.; and Chen, J. 2025. Cola: Collaborative low-rank adaptation. *arXiv preprint arXiv:2505.15471*.