TOUCH BEYOND VISION: A SURVEY OF VISION-TACTILE-LANGUAGE MODELS IN EMBODIED INTELLIGENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Embodied intelligence increasingly leverages multimodal perception—particularly vision and language—to support rich interaction with the physical world. Yet the tactile modality remains under-explored, despite its essential role in human perception and manipulation. In this survey, we systematically review research at the intersection of vision, tactile sensing, and language, which we refer to as Vision-Tactile-Language (VTL) models. We provide (i) a historical context tracing the shift from vision-centric embodied systems to multisensory agents, (ii) foundational aspects of tactile sensing and representation, (iii) methods for integrating vision and touch, (iv) emerging architectures that incorporate language alongside vision and touch, (v) applications in embodied robotics, (vi) current challenges and open problems, and (vii) a forward-looking outlook toward tactile foundation models. We conclude by arguing that touch closes a key gap in embodied AI, enabling truly grounded perception, reasoning and action.

1 INTRODUCTION

Embodied artificial intelligence (AI) aims to endow agents with the ability to perceive, reason, and act in physical environments. Historically, embodied learning has been dominated by vision-centric paradigms, relying on visual perception for navigation, manipulation, and object understanding, and later extended to include language for high-level semantic reasoning and instruction following (Ma et al., 2024; Driess et al., 2023). However, in contrast to human perception—which integrates vision, audition, and touch—most AI systems still lack tactile feedback, a crucial channel for understanding contact, force, and texture (Pirozzi, 2020).

Humans rely heavily on tactile cues to perceive material properties, shape compliance, and microslip during manipulation. Without such feedback, robots must infer contact indirectly through vision, which often fails under occlusion, lighting variation, or complex object geometries (Yuan et al., 2023). The absence of touch limits the robustness of manipulation, sim-to-real transfer, and adaptability in contact-rich tasks. Recent reviews of tactile sensing for robotics emphasize that, despite decades of progress in tactile hardware and signal processing, the integration of tactile sensing with multimodal learning remains at an early stage (Zhang & Luo, 2025; Pirozzi, 2020).

In parallel, vision-language foundation models such as CLIP (Radford et al., 2021), Flamingo (Alayrac et al., 2022), and PaLM-E (Driess et al., 2023) have demonstrated remarkable abilities to align perception with linguistic reasoning. These successes have inspired a new question: *if models can learn from vision and language jointly, can we extend this framework to include touch?* The emerging research direction of Vision–Tactile–Language (VTL) models explores precisely this—combining tactile perception with visual and linguistic understanding to create more grounded and physically aware agents.

This paper presents a comprehensive survey of recent advances in vision–tactile–language modeling for embodied intelligence. Our goals are three-fold:

1. To systematically review tactile sensing technologies, datasets, and representation methods in the context of embodied AI.

2. To analyse how vision and touch are integrated, and how language supervision further enhances representation learning.
3. To identify major challenges—including data scarcity, sensor heterogeneity, cross-modal alignment, and benchmarking—and to outline future directions toward tactile foundation models.

Concretely, our contributions are:

- We trace the historical evolution from vision-centric to multisensory embodied systems.
- We summarise tactile sensing fundamentals and discuss existing tactile datasets and representation paradigms.
- We review recent architectures that align or fuse vision and touch, as well as VTL models integrating language grounding.
- We discuss open challenges and propose potential directions for scalable and generalisable multisensory AI.

The remainder of this paper is structured as follows. Section 2 outlines the historical trajectory from vision-dominant perception to multisensory embodiment. Section 3 discusses the foundations of tactile sensing. Section 4 reviews vision–tactile integration. Section 5 surveys emerging VTL architectures. Section 6 highlights applications in robotic manipulation and multimodal learning. Section 7 analyses challenges and open problems, and Section 8 concludes with future outlooks.

2 Historical Context: From Vision-Centric to Multisensory Embodiment

The trajectory of embodied intelligence has historically been shaped by advances in visual perception. Early research in robotics and AI focused primarily on vision-based sensing for navigation, recognition, and scene understanding (Anderson et al., 2018; Savva et al., 2019). Visual sensors such as RGB and depth cameras provided rich environmental information, enabling breakthroughs in object detection, visual navigation, and embodied instruction following. The dominance of visual data was partly due to its accessibility and the existence of large-scale datasets such as ImageNet and COCO, which catalyzed deep learning in computer vision. However, visual modalities alone lack direct physical grounding: they perceive geometry and appearance but not the forces, compliance, or friction that govern interaction with objects.

Emergence of Language in Embodied AI. The integration of language understanding marked the second major milestone. With the advent of large-scale pretraining, models such as CLIP (Radford et al., 2021), Flamingo (Alayrac et al., 2022), and PaLM-E (Driess et al., 2023) connected vision to natural language supervision, bridging low-level perception with high-level semantics. This gave rise to embodied vision–language–action frameworks (Ma et al., 2024; Sharma et al., 2022), where agents follow textual commands ("pick up the red cup") or answer visual questions about their surroundings. Yet even with such advances, these systems remain predominantly exteroceptive—they perceive the world visually and linguistically, but cannot "feel" it.

Early Exploration of Tactile Sensing. Parallel to the evolution of visual perception, tactile sensing research developed largely within the robotics community. Early tactile arrays and force sensors provided contact and pressure information for robotic hands, with notable examples such as BioTac, and DIGIT sensors (Lambeta et al., 2020). These technologies enabled fine-grained measurements of contact geometry, shear forces, and slip detection, offering insight into material and shape properties. Reviews on tactile sensing (Pirozzi, 2020; Zhang & Luo, 2025) highlight significant progress in sensor fabrication, data collection, and representation learning. However, the tactile modality long remained siloed from mainstream machine learning due to the lack of large-scale datasets and standardized formats.

From Vision–Tactile Fusion to Multisensory Embodiment. Recent works have started to bridge this divide by combining tactile feedback with visual perception (Yuan et al., 2023; Gao et al.,

2023). More recently, the rise of multimodal foundation models (Alayrac et al., 2022; Driess et al., 2023; Chen et al., 2023) has inspired a shift from pairwise modality fusion to unified multisensory representation learning.

The evolution of embodied AI has thus progressed from vision-only perception (Dosovitskiy et al., 2021; He et al., 2016) to vision–language understanding (Radford et al., 2021; Li et al., 2022) and now toward multisensory embodiment, where tactile feedback provides the missing channel of physical grounding (Huang et al., 2023; Li et al., 2024). This historical trajectory motivates the remainder of this paper: exploring how tactile sensing can be systematically incorporated into modern multimodal learning frameworks to achieve more robust, human-like embodied intelligence.

3 Foundations of Tactile Perception in AI

The tactile modality provides embodied agents with a means to directly sense the physical properties of their environment. Unlike vision, which captures the appearance of objects, or language, which conveys semantic abstraction, touch delivers *grounded contact information*—forces, textures, stiffness, temperature, and deformation. This section surveys the foundations of tactile sensing in artificial intelligence, covering sensor technologies, datasets, and learned representations that enable tactile perception to interface with machine learning frameworks.

3.1 Tactile Sensors and Data Acquisition

Modern tactile sensing technologies have evolved considerably over the past decade. Early tactile sensors were resistive or capacitive arrays that measured local pressure changes, such as the Takktile and Weiss sensors. More recent designs exploit optical or vision-based principles, where a camera observes deformations in an elastomeric surface illuminated by LEDs (Lambeta et al., 2020; Donlon et al., 2018). These *visuotactile sensors*, exemplified by DIGIT (Lambeta et al., 2020), and OmniTact (Padmanabha et al., 2020), provide high-resolution "tactile images" of contact geometry and can infer surface normals, slip, and local force distributions. Such sensors have enabled a transition from analog contact signals to structured visual-like data that are well-suited for convolutional and transformer-based neural architectures.

Beyond optical methods, other designs include piezoresistive (Pirozzi, 2020), capacitive, magnetic, and triboelectric sensors. Hybrid sensors combine multiple modalities (e.g., force and vibration) to capture both static and dynamic contact cues. Recent reviews emphasize the emergence of compact, inexpensive, and soft tactile sensors that can be mounted on robotic grippers and humanoid hands (Zhang & Luo, 2025; Pirozzi, 2020). However, despite rapid hardware progress, tactile data remain challenging to collect at scale due to calibration requirements, sensor fragility, and limited throughput compared to visual data collection.

3.2 Tactile Datasets

Tactile datasets have historically been limited in both scale and diversity compared with vision and language corpora, but recent efforts are rapidly expanding their scope. Early tactile and visuotactile datasets such as Tacto (Murali et al., 2021) and TouchSim (Saal & Bensmaia, 2017) focused primarily on low-level perception and control, generating synthetic tactile data or simulating skin mechanics for tasks like grasp stability and slip detection (Calandra et al., 2017). However, tactile data remain highly dependent on physical embodiment—the same object can yield distinct signals under varying sensor geometries, contact forces, or material properties—making cross-platform generalization difficult.

To address these limitations, emerging large-scale multimodal corpora integrate tactile, visual, and linguistic modalities to connect low-level sensations with high-level semantics. Touch100k (Cheng et al., 2024) represents a major milestone in this direction, comprising over 100,000 touch–language–vision triplets with fine-grained natural-language annotations and a curriculum-based pretraining strategy (TLV-Link) that aligns tactile, visual, and linguistic representations. Such resources bridge the gap between sensory grounding and semantic reasoning, enabling scalable pretraining, zero-shot tactile understanding, and laying the foundation for next-generation embodied multimodal models.

3.3 Tactile Representations and Learning Methods

A central challenge in tactile perception for AI is how to represent high-dimensional and heterogeneous signals in a way compatible with learning architectures. Existing approaches can be broadly categorized as follows:

Tactile-as-Image Representations. Visuotactile sensors yield image-like data capturing surface deformations, which can be processed using convolutional neural networks (CNNs) or vision transformers (ViTs) (Lambeta et al., 2020). These methods treat tactile frames analogously to visual images, allowing the use of pretrained backbones for feature extraction. However, unlike visual images, tactile images encode local geometry and contact intensity rather than global appearance.

Signal and Time-Series Representations. For non-visual tactile arrays, the data are often treated as temporal signals—pressure, vibration, or force sequences. Recurrent or temporal convolutional networks are used to model temporal dynamics during contact events (Calandra et al., 2017). Such methods excel in recognizing slip, compliance, and dynamic interactions.

Latent Embeddings and Cross-Modal Learning. Recent research employs self-supervised and contrastive learning to align tactile embeddings with visual or linguistic spaces (Gomes et al., 2022; Zheng & Yuan, 2021; Gao et al., 2023). Early work demonstrated that self-supervised tactile representation learning can capture material and geometric properties without manual labels (Yuan et al., 2017; Calandra et al., 2018). More recent visuotactile frameworks exploit cross-modal contrastive objectives to map tactile and visual signals into a shared latent space, enabling bidirectional prediction and multimodal reasoning (Gomes et al., 2022; Gao et al., 2023).

Tactile perception in AI has thus evolved from simple pressure sensing to high-dimensional visuotactile imaging and multimodal representation learning (Dong et al., 2021; Li et al., 2024). Although data collection and standardization remain bottlenecks, the combination of tactile sensing with deep representation learning has paved the way for large-scale, transferable tactile understanding, forming the basis for the vision–tactile–language models discussed in Section 5.

4 Vision–Tactile Integration: Bridging Perception Modalities

While tactile sensing provides local physical interaction information, vision supplies global spatial context and object appearance. Integrating these two modalities allows embodied agents to reason about both what an object *looks like* and what it *feels like*. This section reviews the main approaches for fusing vision and touch, covering architectural designs, learning objectives, and representative applications.

4.1 Alignment Approaches

The central challenge of vision–tactile integration lies in bridging heterogeneous sensory streams: tactile signals are typically localized, high-frequency, and contact-dependent, while visual data are global and spatially dense. Several approaches have been proposed to align these modalities at different levels of abstraction:

Feature-level alignment. Contrastive learning methods train vision and tactile encoders such that embeddings of corresponding contact events are close in latent space, while mismatched pairs are pushed apart (Calandra et al., 2017). This objective encourages cross-modal correspondence without explicit supervision, analogous to CLIP-style visual–text alignment (Radford et al., 2021).

Cross-attention and multimodal fusion. Beyond alignment, multimodal transformers and cross-attention architectures explicitly fuse visual and tactile tokens. Each modality is processed by its own encoder (e.g., a Vision Transformer for images, a CNN for tactile data), and information is integrated via cross-attention layers that learn inter-modal dependencies (Gao et al., 2023). Such architectures allow mutual conditioning—visual cues guide tactile feature weighting and vice versa—improving material recognition and manipulation stability.

Geometry- and contact-based alignment. For robotic manipulation, accurate geometric correspondence between visual and tactile frames is critical. Some works align the two modalities by estimating contact poses in the visual frame (Yuan et al., 2023; Gomes et al., 2022). Others use 3D reconstruction to integrate tactile depth into object models, effectively "completing" occluded visual surfaces using touch. These approaches bridge spatial grounding between external and contact-based sensing.

4.2 Cross-Modal Learning and Representations

Vision–tactile integration not only enhances perception but also enables knowledge transfer across modalities. Existing approaches can be broadly grouped into three paradigms. First, *vision supervision for tactile representation* leverages large-scale visual data to regularize tactile encoders and compensate for the scarcity of tactile datasets. Second, *tactile augmentation for visual recognition* uses touch to disambiguate visually similar yet physically distinct objects (e.g., sponge vs. rubber block); when fused with vision, tactile cues help correct uncertainty under occlusion or lighting variation and improve robustness in manipulation (Zhang & Luo, 2025). Third, several works pursue *shared latent spaces and generative modeling*, where vision and touch are jointly embedded in a unified representation that supports bidirectional generation—predicting tactile signals from vision ("imagine what it feels like") or reconstructing appearance from touch ("see by feeling") (Gao et al., 2023; Gomes et al., 2022). Such cross-modal representations lay the foundation for tactile simulation, data augmentation, and more general embodied multimodal learning.

4.3 Representative Works

A variety of systems illustrate the effectiveness of vision–tactile integration across different architectures and objectives. Early efforts focused on explicit model-based fusion, where tactile and visual signals are combined through spatial or feature-level alignment. Robot Synesthesia (Yuan et al., 2023) introduces visuotactile sensing for in-hand manipulation using point-cloud-based fusion of tactile and visual inputs, while the VisuoTactile Transformer (Gao et al., 2023) employs cross-attention mechanisms to integrate modalities for material classification and slip detection.

Building on these approaches, recent studies pursue *unified tactile representation learning* that generalizes across sensors, materials, and interaction types. UniTouch (Yang et al., 2024) aligns tactile signals from heterogeneous vision-based sensors with visual, linguistic, and even auditory modalities through shared encoder alignment, establishing one of the first unified tactile representation frameworks. AnyTouch (Feng et al., 2025) further bridges static (image) and dynamic (video) tactile data, introducing a unified static–dynamic representation that enhances cross-sensor generalization and robust visuotactile reasoning.

4.4 Challenges in Vision–Tactile Integration

Despite significant progress, several challenges remain open. First, *data imbalance* between visual and tactile samples leads to biased models, since tactile data are harder to collect. Second, *temporal synchronization* between asynchronous visual and tactile streams can degrade alignment during dynamic interactions. Third, *domain heterogeneity* across sensors and lighting conditions limits model transferability. Finally, there is no widely adopted benchmark for evaluating vision–tactile fusion, making quantitative comparison difficult (Zhang & Luo, 2025). These issues motivate the development of unified multimodal benchmarks and pretraining strategies that we discuss further in Section 5.

5 Vision–Tactile–Language Models (VTL Models)

Building upon the success of vision–language foundation models, recent research has begun exploring the inclusion of tactile sensing as a third modality to achieve multisensory embodiment. Vision–Tactile–Language (VTL) models aim to jointly learn representations that integrate visual appearance, tactile feedback, and linguistic semantics, enabling agents to perceive, reason, and act with physical grounding.

5.1 Motivation and Conceptual Overview

Language provides a natural supervisory signal for describing physical interactions: phrases such as "soft fabric," "rough metal," or "the object slipped" correspond directly to tactile events. VTL models thus seek to map tactile signals and visual observations into semantic spaces where linguistic concepts can act as anchors. Incorporating tactile input allows grounding of language in the material and force-based properties of the physical world—an essential component of embodied understanding.

Conceptually, a VTL model extends the standard vision–language paradigm by introducing a tactile encoder that processes visuotactile images, pressure maps, or force sequences, and aligns its latent representations with those of vision and text encoders through joint pretraining objectives. This multimodal alignment allows queries such as "what does this surface feel like?" or "find an object that feels smooth and looks shiny" to be answered through joint reasoning.

5.2 Model Architectures

Most existing VTL systems adopt a three-branch architecture: a vision encoder, a tactile encoder, and a language encoder, each producing modality-specific embeddings that are fused via cross-modal transformers or contrastive alignment layers. Representative architectural patterns include:

Contrastive Tri-Modal Alignment. Inspired by CLIP (Radford et al., 2021), models such as tactile, and language embeddings using triplet or multi-modal contrastive losses. Given paired visual frames, tactile readings, and natural language descriptions, the model learns a shared embedding space where semantically related samples are close together. This design supports zero-shot retrieval across modalities (e.g., "find the tactile reading corresponding to this caption").

Cross-Attention Fusion Transformers. Other architectures employ cross-attention layers to allow modalities to attend to each other dynamically. For instance, the VisuoTactile Transformer (Gao et al., 2023) extends to tri-modal settings, where tactile embeddings influence both vision and language representations for multimodal reasoning.

Unified Multisensory Encoders. Recent research explores unified multimodal encoders that process both visual and tactile inputs as spatial tokens. Such encoders—often transformer-based—learn inter-modality dependencies implicitly, leveraging large-scale pretraining and masked modeling objectives (Xie & Correll, 2025). This direction mirrors the trend toward "foundation models" in robotics and perception.

5.3 Training Strategies

Training VTL models requires balancing heterogeneous data sources and objectives. Several strategies have emerged to effectively align visual, tactile, and linguistic modalities.

Joint Pretraining. Paired visual, tactile, and textual samples are used for multimodal pretraining with objectives such as contrastive alignment, masked modeling, and captioning. Contrastive losses (e.g., InfoNCE) encourage cross-modal correspondence (Radford et al., 2021; Li et al., 2024), while masked token prediction and language-conditioned decoding leverage textual supervision (Alayrac et al., 2022; Driess et al., 2023). Such joint objectives enable models to predict linguistic descriptions of tactile sensations or infer contact states from text-based cues, as demonstrated in TouchGPT (Huang et al., 2023).

Two-Stage Training. When tactile data are limited, a two-stage paradigm is often adopted: first, vision–tactile encoders are pretrained using self-supervised contrastive learning (Zheng & Yuan, 2021; Gao et al., 2023), then a language module is introduced through instruction tuning or captioning datasets (Li et al., 2022; Tu et al., 2025). This decoupled approach stabilizes optimization and improves transferability across domains.

Multitask and Reinforcement Learning. For embodied and robotic applications, multimodal pretraining can be integrated with reinforcement learning, where tactile and visual feedback jointly

inform policy optimization. Language serves as a high-level goal specification, while touch provides fine-grained corrective signals for grasping or insertion tasks (Calandra et al., 2018; Lee et al., 2019; Driess et al., 2023). This combination bridges high-level reasoning with low-level control, paving the way for tactile-informed embodied agents.

5.4 Benchmarks and Evaluation

Compared with vision–language evaluation tasks such as visual question answering or image captioning, benchmarks for vision–tactile–language (VTL) systems remain nascent. Existing studies typically evaluate four categories of tasks: (i) *multimodal retrieval*, where models retrieve vision–tactile pairs given a language query (e.g., "a soft rubber object"); (ii) *haptic captioning*, which generates natural language descriptions of tactile experiences; (iii) *cross-modal prediction*, involving the synthesis of tactile readings from visual input or vice versa (Gao et al., 2023); and (iv) *manipulation policy learning*, which integrates language-guided touch feedback to predict or optimize grasp success (Calandra et al., 2017).

Despite these efforts, standardized large-scale benchmarks for VTL remain unavailable. Data scarcity, heterogeneous sensor modalities, and inconsistent linguistic annotations hinder reproducibility and comparison. Ongoing community initiatives aim to construct unified visuotactile–language corpora and open-source evaluation suites—akin to COCO or LVIS—for advancing large-scale and standardized assessment in this emerging field (Zhang & Luo, 2025; Xie & Correll, 2025).

5.5 Representative VTL Systems

Recent systems demonstrate the expanding scope of vision–tactile–language (VTL) modeling, ranging from core tri-modal architectures to generative and reasoning-oriented frameworks. ForceFM (Xie & Correll, 2025) provides one of the earliest overviews and prototypes of touch-aware foundation models, emphasizing large-scale tactile–language data alignment and scalable pre-training strategies. Building on alignment-based foundations, newer studies explore generation and higher-level reasoning. TextToucher (Tu et al., 2025) enables fine-grained *text-to-touch* generation for controllable tactile synthesis conditioned on linguistic prompts, while Octopi (Yu et al., 2024) grounds large tactile–language models on tactile videos to infer object properties such as material, stiffness, and contact dynamics. The latter also releases the PhysiCLeAR dataset and a standardized language-evaluation benchmark, supporting quantitative assessment of tactile understanding.

5.6 Discussion

The emergence of VTL models marks a paradigm shift from purely visual-language perception to multisensory foundation models. By unifying the three major perceptual axes—vision, touch, and language—these systems can achieve more grounded reasoning, semantic generalization, and physical adaptability. Nonetheless, significant research gaps remain in dataset scale, modality synchronization, and open benchmarking. Bridging these challenges will be key to realizing tactile foundation models capable of zero-shot tactile reasoning and cross-modal generation.

6 Applications and Emerging Trends

The integration of tactile sensing with vision and language has opened new frontiers for embodied intelligence. By enriching perception with contact-based feedback and semantic grounding, VTL systems enable agents to interact with their environments more safely, robustly, and intelligently. This section reviews major application areas where tactile information has proven beneficial and highlights emerging trends that leverage multimodal fusion in robotics and AI.

6.1 Robotic Manipulation and Grasping

Robotic manipulation is perhaps the most natural domain for tactile sensing. While visual perception informs object identity and geometry, tactile feedback provides crucial cues about contact state,

grip stability, and surface compliance. Integrating both modalities improves performance in grasp planning, slip detection, and in-hand manipulation.

Early work demonstrated that tactile sensing improves grasp success prediction compared to vision-only systems (Calandra et al., 2017). Using DIGIT sensors, robots can detect micro-slip and adjust grasp force in real time (Lambeta et al., 2020). Vision–tactile fusion further enhances robustness by allowing grasp corrections even under visual occlusion (Yuan et al., 2023). Recent VTL models extend these capabilities with linguistic grounding: for instance, a robot can interpret the command "grasp the softest sponge" or "pick the rough cylinder" and use tactile cues to discriminate between visually similar items.

Beyond simple grasping, tactile feedback plays an essential role in fine manipulation tasks such as peg insertion, surface polishing, and fabric folding. Touch-guided policies and contact-rich reinforcement learning (Cao et al., 2022) have shown that combining tactile feedback with visual context significantly accelerates skill acquisition and reduces physical trial failures.

6.2 Simulation-to-Real Transfer

One of the persistent challenges in robotics is the sim-to-real gap: policies trained in simulation often fail in the real world due to unmodeled contact dynamics, frictional variance, and material compliance. Tactile sensing mitigates this problem by providing *direct physical grounding* during real-world interaction. By incorporating tactile feedback into policy learning, robots can adapt online to deviations between simulated and actual dynamics (Murali et al., 2021; 2022).

Moreover, simulated tactile data can augment real datasets. Frameworks such as Tacto (Murali et al., 2021) and TouchSim (Saal & Bensmaia, 2017) allow large-scale synthetic data generation for pretraining tactile encoders, similar to how synthetic imagery bootstraps visual models. Domain randomization techniques are used to bridge sensor-specific variations, improving cross-platform generalization. Combined with visual and linguistic supervision, such methods pave the way toward foundation models that generalize across domains (Xie & Correll, 2025).

Beyond simulation pretraining, recent work shows that language can serve as a unifying interface to fuse heterogeneous sensors into generalist visuomotor policies. FuSe ("Beyond Sight") fine-tunes pretrained policies to incorporate touch (and audio) via language grounding, yielding more than 20% improvement in real-world success on contact-rich manipulation tasks (Jones et al., 2025).

6.3 Interactive and Grounded Learning

VTL systems enable a new paradigm of interactive embodied learning, where agents learn through continuous feedback from both humans and the environment. Language serves as a high-level communication channel, while vision and touch provide perceptual grounding. For example, a human could instruct a robot to "press harder until you feel resistance" or "rub the surface to check if it's smooth," linking linguistic concepts with tactile experiences.

Such systems have potential applications in human–robot collaboration, haptic teleoperation, and assistive technologies. In rehabilitation or prosthetics, for instance, VTL-based interfaces could provide language-mediated tactile feedback to improve intuitiveness and user control (Dahiya & Valle, 2019).

Beyond teleoperation, MimicTouch collects multi-modal human *tactile* demonstrations to directly teach contact-rich manipulation skills, mitigating modality mismatch (latency, occlusion) common in vision-only setups and improving policy learning from grounded touch cues.

6.4 Emerging Trends

Several research trends are emerging at the intersection of vision, touch, and language:

- **Tactile Foundation Models.** Inspired by visual-language pretraining, researchers are building large-scale tactile datasets and encoders capable of general tactile reasoning (Xie & Correll, 2025).

- **Generative Touch Simulation.** Diffusion and generative adversarial models are being applied to synthesize tactile images or force patterns conditioned on visual or linguistic prompts, enhancing data diversity (Gao et al., 2023).
- **Integration with 3D and Physics Engines.** Combining tactile data with 3D scene understanding (e.g., NeRFs or physics-based rendering) enables full-scene physical reasoning (Yuan et al., 2023).
- **Multisensory Human–AI Interaction.** Haptic-language interfaces are emerging where users communicate tactile expectations through speech, and the agent responds both verbally and physically (Dahiya & Valle, 2019).

The applications of VTL systems extend far beyond robotics. By combining perception, language, and touch, embodied agents gain the capacity for nuanced interaction—grasping, exploring, describing, and learning through experience. These capabilities form a critical step toward physically grounded AI, where multimodal fusion enables both cognitive understanding and safe, adaptive manipulation of the real world.

7 Challenges and Open Problems

Despite the rapid progress of multisensory learning, significant challenges remain before Vision–Tactile–Language (VTL) models can reach the maturity and generality of vision–language systems. This section highlights the key bottlenecks and open research questions across data, model, and evaluation dimensions, as well as ethical and safety considerations for embodied tactile AI.

7.1 Data Scarcity and Sensor Diversity

A fundamental limitation in VTL research is the scarcity of large-scale, standardized tactile datasets. Unlike vision or language corpora, tactile data are inherently physical, requiring contact interactions to collect (Zhang & Luo, 2025). This process is slow, labor-intensive, and sensor-dependent. Even small variations in sensor design—such as gel stiffness, camera placement, or surface coating—can drastically alter tactile readings (Lambeta et al., 2020). As a result, tactile datasets are fragmented across research groups, each using different calibration pipelines and data formats.

To overcome this challenge, the community has begun developing simulation frameworks such as Tacto (Murali et al., 2021) and TouchSim (Saal & Bensmaia, 2017) to generate synthetic tactile data, analogous to how synthetic imagery has boosted vision research. However, bridging the *reality gap* between simulated and real tactile signals remains an open problem (Murali et al., 2022). Standardization of data collection protocols, metadata annotation (e.g., contact force, material type), and multimodal synchronization will be crucial for scalable tactile datasets and pretraining.

7.2 Cross-Modal Representation Alignment

Aligning vision, tactile, and language modalities poses significant representational and temporal challenges. Vision captures global appearance, touch captures localized contact dynamics, and language encodes symbolic abstractions. Synchronizing and fusing such heterogeneous information streams is nontrivial. Temporal misalignment is common in manipulation, where tactile signals lag or lead visual observations due to actuation delays (Yuan et al., 2023). Spatial correspondence between tactile contact regions and visual frames also depends on accurate pose estimation.

Furthermore, tactile signals differ statistically from visual data—while visual embeddings are high-dimensional and dense, tactile signals are often sparse or low-frequency. Naïvely aligning these modalities may bias the joint representation toward visual features. Recent approaches use contrastive or cross-attention architectures (Gao et al., 2023), yet robust tri-modal alignment strategies that respect both temporal and spatial consistency are still lacking. Future work may benefit from explicit geometric priors or physics-informed alignment mechanisms.

7.3 Generalization Across Embodiments

VTL models often overfit to specific hardware configurations or robot embodiments. A model trained on a particular tactile sensor or manipulator may fail when deployed on another platform with different mechanical compliance, friction, or camera placement (Pirozzi, 2020; Zhang & Luo, 2025). Achieving cross-embodiment generalization requires disentangling sensor-specific noise from intrinsic contact representations. Potential directions include meta-learning across multiple robots, sensor-domain adaptation, or pretraining on simulated sensor ensembles (Xie & Correll, 2025).

7.4 Benchmarking and Evaluation

Unlike vision–language research, which benefits from standardized datasets (e.g., COCO, VQA), the VTL field lacks shared benchmarks. Existing works evaluate on disparate tasks—material classification, slip detection, captioning—using inconsistent metrics. This fragmentation hampers direct comparison and reproducibility. Community-wide efforts are needed to define open evaluation protocols for:

- Cross-modal retrieval and captioning tasks.
- Manipulation benchmarks integrating tactile success rates.
- Multisensory generalization tests across sensors and domains.

In addition, there is a need for unified embodied evaluation environments, where agents can interact physically and semantically with real or simulated objects under multimodal feedback (Savva et al., 2019).

7.5 Computational and Modeling Challenges

Tactile data streams are high-frequency and high-dimensional, especially for visuotactile sensors that output RGB images at 60–120 Hz. Processing these signals jointly with video and language requires substantial computational resources. Efficient architectures for multi-stream fusion—possibly through hierarchical encoders or adaptive sampling—remain an active area of research. Moreover, large-scale tri-modal pretraining demands novel data-efficient objectives to prevent overfitting given the limited tactile data availability.

7.6 Ethical, Safety, and Societal Considerations

As embodied AI systems increasingly interact with the physical world, safety and ethics become central concerns. Tactile-enabled robots can exert forces on humans or fragile objects, introducing new risks compared to purely visual systems. Unintended excessive pressure or unsafe contact could cause damage or harm. Therefore, tactile feedback must be integrated with strict safety constraints and transparent control policies (Dahiya & Valle, 2019).

From a societal perspective, the rise of VTL models also raises issues of data privacy and bias. Tactile datasets collected from human–robot interaction may inadvertently encode sensitive biometric or material information. Furthermore, linguistic supervision may import semantic biases (e.g., associating textures or shapes with subjective adjectives). Responsible development of tactile AI requires robust consent protocols, anonymization of tactile-human data, and interpretability tools for contact-based decision making.

7.7 Summary

In summary, the challenges for VTL research span multiple levels—from sensing hardware to multimodal representation and societal implications. Addressing these obstacles will require interdisciplinary collaboration across robotics, computer vision, haptics, and natural language processing. Overcoming them is essential to unlock the next generation of embodied AI agents that can truly see, feel, and reason about the physical world.

8 FUTURE OUTLOOK

Vision–Tactile–Language (VTL) modeling represents a crucial step toward building physically grounded artificial intelligence. Yet the field remains at an early stage, with most studies focusing on small-scale prototypes or domain-specific applications. Looking ahead, we envision several transformative research directions that could shape the future of multisensory embodied AI.

8.1 TOWARDS TACTILE FOUNDATION MODELS

The remarkable generalization ability of vision–language foundation models suggests the potential for analogous tactile foundation models. Such models would be pretrained on large-scale multimodal datasets encompassing visual, tactile, and linguistic modalities, enabling zero-shot tactile reasoning and open-ended physical understanding. Recent works like ForceFM (Xie & Correll, 2025) outline early frameworks for tri-modal pretraining, aligning touch with semantics and visual context. However, achieving foundation-level scale requires advances in tactile data standardization, sensor simulation, and multimodal annotation.

In the near future, self-supervised pretraining on simulated tactile data (Murali et al., 2021; Saal & Bensmaia, 2017) combined with real-world fine-tuning could yield scalable tactile representations similar to ImageNet pretraining in vision. Moreover, continual learning frameworks that allow models to update tactile knowledge from ongoing embodied experience—"feeling to learn"—will be essential for lifelong adaptation.

8.2 GENERATIVE TOUCH MODELS

Generative modeling represents an emerging frontier for tactile AI. Generative Touch Models aim to synthesize tactile signals conditioned on visual or linguistic input, effectively allowing AI systems to "imagine what something feels like." For example, given an image of a fabric and the caption "smooth silk," the model could generate corresponding tactile patterns or pressure maps (Gao et al., 2023). Such synthetic tactile data can support low-cost data augmentation, simulation training, and interactive visualization.

Diffusion models and cross-modal transformers have already demonstrated success in text-to-image and text-to-audio synthesis; extending these paradigms to touch will require modeling spatial–temporal contact dynamics and high-frequency force signals. Recent explorations of visuotactile diffusion networks and generative adversarial tactile models suggest promising results for haptic rendering and virtual reality applications (Gomes et al., 2022; Yuan et al., 2023).

8.3 INTEGRATION WITH 3D AND PHYSICS-BASED REPRESENTATIONS

Tactile sensing naturally complements 3D geometry and physics-based reasoning. While vision provides global 3D shape, touch offers fine-grained local surface geometry and material cues. Integrating tactile feedback into 3D reconstruction pipelines enables reconstruction of occluded surfaces and estimation of friction or compliance. Combining tactile data with differentiable physics simulators could enable embodied agents to predict future contact forces, energy transfer, or deformation, enhancing physical reasoning.

Future embodied systems may fuse 3D NeRF representations with tactile textures, producing unified "touch-aware" scene representations. Such models would allow queries like "what regions of this object feel rough or soft?" or "how does pressing here deform the surface?"—bridging perception, simulation, and reasoning in one cohesive framework.

8.4 MULTISENSORY HUMAN–AI COLLABORATION

The long-term vision for VTL models extends beyond robotics into multisensory human–AI interaction. By grounding communication in shared perceptual spaces, robots and humans could exchange not only words and visuals but also sensations. For instance, a collaborative robot could report "the surface feels slippery" or "contact pressure exceeds safe threshold," while humans could instruct using tactile language such as "press gently until you feel texture."

This capability is especially relevant in healthcare, teleoperation, and assistive robotics (Dahiya & Valle, 2019). Integrating tactile feedback with natural language dialogue and visual context could enable intuitive, trustable human–AI partnerships. Future interfaces may even provide bidirectional haptic feedback, allowing humans to "feel what the robot feels" in real time through wearable devices.

8.5 Beyond VTL: Toward VTLA and Omni-VTLA Models

Recent work extends the perception-oriented VTL paradigm toward Vision–Tactile–Language–Action (VTLA) models that couple tactile perception with language-conditioned control policies. TLA integrates tactile cues and language instructions to learn contact-rich manipulation strategies (Hao et al., 2025), while VTLA adds visual input and preference learning for insertion tasks, achieving high success rates and strong sim-to-real transfer (Zhang et al., 2025). Going further, OmniVTLA unifies diverse tactile sensors through a semantic-aligned tactile Vision Transformer and introduces the large-scale ObjTac dataset with over 135k tri-modal samples (Cheng et al., 2025). Together, these models move beyond perception to embodied decision-making, representing a step toward universal multisensory foundation models.

8.6 Ethical and Societal Considerations

As tactile AI becomes pervasive, ethical design and inclusive datasets will be critical. Touch is an intimate modality: its misuse in human–robot interaction raises privacy and safety concerns (Dahiya & Valle, 2019). Ensuring transparency, consent, and fairness in tactile data collection will be fundamental for socially acceptable deployment. Collaborations between engineers, ethicists, and policymakers will be needed to define standards for responsible tactile AI.

8.7 Summary

In summary, the future of embodied intelligence lies in bridging perception, language, and physical interaction. VTL models are paving the way toward multisensory foundation models that can both understand and generate tactile experiences. By uniting vision, touch, and language through scalable learning frameworks and ethical design, researchers can move closer to realizing physically grounded, adaptive, and communicative artificial agents that truly "see, feel, and understand" the world.

9 Conclusion

This survey provides a comprehensive overview of recent advances at the intersection of vision, tactile sensing, and language in embodied intelligence. We traced the historical trajectory from vision-centric perception to multisensory embodiment, highlighting how tactile sensing completes the perceptual triad that underpins human interaction with the physical world. We summarized the foundations of tactile sensing technologies, datasets, and representation learning, followed by an in-depth discussion of vision–tactile integration and the emergence of Vision–Tactile–Language (VTL) models. Our review reveals a growing convergence between robotics, computer vision, and natural language processing, where tactile perception is evolving from a niche sensory channel into a key component of multimodal reasoning and embodied understanding, opening new avenues for contact-rich manipulation, interactive human–robot communication, and grounded learning.

Despite this progress, several challenges persist: the scarcity and heterogeneity of tactile datasets (Zhang & Luo, 2025), difficulties in cross-modal alignment (Gao et al., 2023), limited cross-embodiment generalization (Xie & Correll, 2025), and the absence of standardized benchmarks for evaluating VTL models. Addressing these issues will require coordinated efforts across haptics, perception, and machine learning communities. Looking forward, we envision a new generation of tactile foundation models that unify visual, tactile, and linguistic modalities under scalable self-supervised frameworks. Such models could reason about and generate tactile experiences, simulate physical interactions, and communicate them through natural language, ultimately enabling robots that both perceive and *feel*. The integration of touch, vision, and language thus represents a pivotal step toward truly grounded artificial intelligence—systems that can not only see and talk about the

world but also *understand* it through direct physical experience. We hope this survey serves as both a roadmap and a call to action for advancing toward that multisensory future.

REFERENCES

Jean-Baptiste Alayrac et al. Flamingo: A visual language model for few-shot learning. In *NeurIPS*, 2022.

Peter Anderson, Qi Wu, Damien Teney, Joel Bruce, Anton van den Hengel, et al. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *CVPR*, 2018.

Roberto Calandra, Andrew Owens, Mukundan Upadhyaya, Wenzhen Yuan, Justin Lin, Edward H. Adelson, and Sergey Levine. The feeling of success: Does touch sensing help predict grasp outcomes? In *Conference on Robot Learning (CoRL)*, 2017.

Roberto Calandra, Andrew Owens, Dinesh Jayaraman, Justin Lin, Wenzhen Yuan, Jitendra Malik, Sergey Levine, and Edward H. Adelson. More than a feeling: Learning to grasp and regrasp using vision and touch. In *IEEE Robotics and Automation Letters (RA-L)*, volume 3, pp. 3300–3307, 2018. URL https://ieeexplore.ieee.org/document/8460901.

Yu Cao, Xin Li, Tong Wu, and Yuke Zhu. Contact-rich manipulation via deep reinforcement learning with tactile feedback. *IEEE Transactions on Robotics*, 2022.

Guanglu Chen, Zhe Li, Yuntao Zhang, et al. Shikra: Unifying vision-language understanding and generation with instruction-finetuned multimodal large language models. *arXiv preprint arXiv:2306.02522*, 2023.

Ning Cheng, Changhao Guan, Jing Gao, Weihao Wang, You Li, Fandong Meng, Jie Zhou, Bin Fang, Jinan Xu, and Wenjuan Han. Touch100k: A large-scale touch-language-vision dataset for touch-centric multimodal representation. *arXiv preprint arXiv:2406.03813*, 2024.

Zhengxue Cheng, Yiqian Zhang, Wenkang Zhang, Haoyu Li, Keyu Wang, Li Song, and Hengdi Zhang. Omnivtla: Vision–tactile–language–action model with semantic-aligned tactile sensing. *arXiv preprint arXiv:2508.08706*, 2025.

Ravinder S. Dahiya and Maurizio Valle. Human–robot interaction through tactile sensing: A survey. *IEEE Transactions on Robotics*, 35(3):703–724, 2019.

Siyuan Dong, Wenzhen Yuan, and Edward H. Adelson. Gelsight imagenet: Large-scale visuotactile dataset for representation learning. *arXiv preprint arXiv:2107.14152*, 2021.

Elliot Donlon, Siyuan Dong, Moju Liu, Jie Li, and Edward H. Adelson. Gelslim: A high-resolution, compact, robust, and calibrated tactile-sensing finger. In *IEEE/RSJ IROS*, 2018.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Danny Driess et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.

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 *International Conference on Learning Representations (ICLR)*, 2025.

Tianyu Gao, Wei Xu, Ruolin Chen, and Shan Luo. Visuotactile transformer: Cross-attention fusion for multimodal material recognition. In *IEEE/RSJ IROS*, 2023.

Pedro Gomes, Hugo Araujo, and Alexandre Bernardino. Learning visuotactile 3d object reconstruction via self-supervised cross-modal learning. *IEEE Robotics and Automation Letters*, 7(2): 2765–2772, 2022.

Peng Hao, Chaofan Zhang, Dingzhe Li, Xiaoge Cao, Xiaoshuai Hao, Shaowei Cui, and Shuo Wang. Tla: Tactile-language-action model for contact-rich manipulation. *arXiv preprint arXiv:2503.08548*, 2025. doi: 10.48550/arXiv.2503.08548.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Zhiyong Huang, Ziyu Gao, Wei Chen, Bin Fang, Hao Zhang, et al. Touchgpt: A vision–tactile–language model for multisensory embodied intelligence. *arXiv preprint arXiv:2308.10901*, 2023.

Joshua Jones, Oier Mees, Carmelo Sferrazza, Kyle Stachowicz, Pieter Abbeel, and Sergey Levine. Beyond sight: Finetuning generalist robot policies with heterogeneous sensors via language grounding. *arXiv preprint arXiv:2501.04693*, 2025.

Mike Lambeta, Po-Wei Chou, Stephen Tian, et al. Digit: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation. In *IEEE Robotics and Automation Letters*, volume 5, pp. 3838–3845, 2020.

Michelle A. Lee, Yuke Zhu, Krishnan Srinivasan, Silvio Savarese, Li Fei-Fei, and Animesh Garg. Touch and go: Learning to manipulate with tactile sensing and reinforcement learning. In *Proceedings of the Conference on Robot Learning (CoRL)*, 2019.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven CH Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning (ICML)*, 2022.

Yutong Li, Weihao Wang, Jing Gao, and Bin Fang. Vt-lm: Vision–tactile–language model for multimodal representation learning. *arXiv preprint arXiv:2402.04387*, 2024. URL https://arxiv.org/abs/2402.04387.

Y. Ma et al. A survey on vision-language-action models for embodied ai. *arXiv preprint arXiv:2405.14093*, 2024.

Aditya Murali, Roberto Calandra, and Wenzhen Yuan. Tactile transfer: Learning to transfer from simulated to real tactile perception. *IEEE Robotics and Automation Letters*, 7(3):6003–6010, 2022.

Aditya Murali et al. Tacto: A fast, flexible, and open-source simulator for tactile sensors. In *Conference on Robot Learning (CoRL)*, 2021.

Ananya Padmanabha, Stephen Tian, Po-Wei Chou, et al. Omnitact: A multi-directional high-resolution touch sensor. In *IEEE Robotics and Automation Letters*, volume 5, pp. 5881–5888, 2020.

Salvatore Pirozzi. Tactile sensors for robotic applications. *Sensors*, 20(24):7009, 2020.

Alec Radford et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Hannes P. Saal and Sliman J. Bensmaia. Touchsim: Simulating tactile signals from the whole hand with millisecond precision. *PNAS*, 114(28):E5693–E5702, 2017.

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, et al. Habitat: A platform for embodied ai research. In *ICCV*, 2019.

Pratyusha Sharma, Abhinav Gupta, and Saurabh Kumar. Learning generalizable vision–language–action models for embodied tasks. In *NeurIPS*, 2022.

Jiahang Tu, Hao Fu, Fengyu Yang, Hanbin Zhao, Chao Zhang, and Hui Qian. Texttoucher: Fine-grained text-to-touch generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7455–7463, 2025.

Weipeng Xie and Nikolaus Correll. Towards forceful robotic foundation models: A literature survey. *arXiv preprint arXiv:2504.11827*, 2025.

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. 26340–26353, June 2024.

Samson Yu, Kelvin Lin, Anxing Xiao, Jiafei Duan, and Harold Soh. Octopi: Object property reasoning with large tactile-language models. In *Robotics: Science and Systems (RSS)*, 2024.

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

Y. Yuan, H. Che, Y. Qin, et al. Robot synesthesia: In-hand manipulation with visuotactile sensing. *arXiv preprint arXiv:2312.01853*, 2023.

Chaofan Zhang, Peng Hao, Xiaoge Cao, Xiaoshuai Hao, Shaowei Cui, and Shuo Wang. Vtla: Vision–tactile–language–action model with preference learning for insertion manipulation. *arXiv preprint arXiv:2505.09577*, 2025.

J. Zhang and S. Luo. Recent advances and challenges of tactile sensing for robotics: From sensors to data, representation and applications. *ScienceDirect*, 2025.

Yilin Zheng and Wenzhen Yuan. Self-supervised learning for tactile representation with tactilessl. *IEEE Robotics and Automation Letters (RA-L)*, 6(4):8333–8340, 2021. URL https://ieeexplore.ieee.org/document/9507732.