TactiLDM for VLA: Learning Unified Visual-Text-Tactile Representations to Empower Vision-Language-Action Policies

Zhiyan Li*
ACM Class
Shanghai Jiao Tong University
Shanghai, China
lzy001@sjtu.edu.cn

Hantao Jiang*
ACM Class
Shanghai Jiao Tong University
Shanghai, China
eric_jht@sjtu.edu.cn

*These authors contributed equally to this work.

***Abstract*—Integrating tactile perception into Vision-Language-Action (VLA) models is essential for achieving human-level dexterity and robust physical interaction in robotics. In this paper, we propose TactiLDM, a novel framework designed to learn unified visual-text-tactile representations within a Latent Diffusion Model (LDM) framework. Our approach employs a two-stage training strategy:1) align tactile embeddings with vision and text modalities through contrastive learning;2) utilize a conditional latent diffusion process, capturing both high-level semantics and intrinsic physical attributes. On Touch and Go (TAG) dataset, TactiLDM achieves 82.84% accuracy in tactile-only material classification, outperforming prior tactile encoders by over 15%. In *zero-shot* tactile task environment, π_0 policy with tactile modality integration achieves 4x increase in success rate and a 60% improvement in interaction stability. These results pave the way for more capable multimodal robotic systems. Code of our framework can be found at https://github.com/lzy001Yuki/TactiLDM**

***Index Terms*—Tactile Sensing, Multimodal Representation Learning, Vision-Language-Action Models**

I. Introduction

Learning robust and semantic-rich representations from multimodal sensory inputs is a cornerstone of embodied intelligence, enabling robots with information-rich interaction with the physical world and precise control on dexterous contact-rich manipulation tasks effectively [1]. Recent advances in vision-language-action (VLA) models have pushed this direction forward by unifying perception, instruction understanding and action generation within a single architectural framework [2], [3]. Built upon this unified formulation, current-stage VLAs typically make their progress on emphasizing language ability and capability [2], rearranging VLA system centering a SOTA core model [4], or focusing on CoT reasoning as a reinforcement learning goal [5], etc.. Despite their architectural differences, they overwhelmingly remain vision-dominant, with VLMs serving as the finetuning model base [4], [6] and visual input serving as the primary perceptual modality for grounding actions.

Yet as VLAs evolve, there is a central challenge for contact-rich manipulations: visual input alone cannot fully support contact-rich manipulation [**?**], [7]. Among the available sensory channels in robot action tasks, vision and touch offer highly complementary information: vision providesobject geometry and semantic identity, whereas touch perception is capable of capturing material properties, frictional forces [8], and interaction dynamics such as vibrations.

Current tactile integration in action models remains constrained by sparsity and scarcity of data. This is because tactile and action signals are intrinsically temporally sparse [9]. Recent works in VLA models have attempted to bridge vision and touch via contrastive learning [10], masked modeling [11], or joint embedding alignment [12]. However, tactile signals—though commonly represented as images—are inherently low in high-level semantic content compared to natural vision. As a result, when processed by off-the-shelf visual encoders (e.g. ResNet [13]), tactile inputs yield representations that underperform significantly on semantic tasks compared to vision or text inputs.

Due to the current scarcity of tactile-action data, learning effective touch-based policies remains challenging. To address this, our work introduces TactiLDM, a tactile-specific encoder within latent diffusion model framework and brings enlightenment for next-generation vision-tactile-language-action (VTLA) models that can fully exploit rich tactile feedback.

Our main contribution is threefold:

- TactiLDM, a tactile encoder training framework with extensive model structure, highly adaptive alignment method and strong semantic ability.
- The tactile encoder of TactiLDM is easily swapable, and it integrates seamlessly with diverse vision or text encoders in contrastive learning with negligible performance loss, demonstrating robust implicit physical encoding capability.
- A low-cost, zero-shot tactile extension for VLA models via TactiLDM, which enables visuo-tactile fusion and improves action reasoning using only vision-tactile synchronized data—without requiring tactile-action supervision or reinforcement learning.

II. Related Work

A. Touch Representation Encoding

Touch representation (ordinarily image-based, from GelSight or similar sensors [14]) encoding methodology determines the recognition and classification ability of an encoder. Embodied intelligence, such as VLAs, use *touch-as-vision* paradigm in majority, treat tactile data as visual inputs by directly applying pretrained vision encoders [6], whose choices include ResNet [13] and SigLIP [15], or pretrained vision-language models (VLMs). This paradigm makes the greatest convenience but lacks specific tactile understanding without adaptation refining e.g. LoRA. Variants [16] leverage tactile-language models (TLMs) such as Octopi [17] and LLaMa [18] to generate language descriptions of tactile input.

Individual Tactile Encoder paradigm is popular as well. Our zero-shot extension experiment also use this paradigm. Typical structures usage include MLP [19], CNN [14], convolutional autoencoders [20] and ViTs [21]. On novel methodology and systems, 3D-ViTac [22] proposes a real-world point cloud rebuilding system, while AnyTouch [23] employed a two-stage training framework to unify static and dynamic features. Encoders that aim on generalization on heterogeneous data, most famously T3 [14], routes specific encoders and project the features into a shared attention chunk. The model structure we choose and the training methodology design are distinguishing points, as open source tactile datasets are relatively scarce but comprehensive on material classes.

B. Modality Extension of VLA

VLA frameworks have traditionally unified vision and language for robotics [24], and recent works seek to extend this unification with additional physical sensory modalities such as audio [25] and touch. Recent research that evolves VLA to vision-tactile-language-action models (VTLAs) converges on several methodological directions:

- *Semantic Alignment* between tactile signals and pretrained vision-language representations, typically through contrastive or language-grounding objectives, allowing inheriting advanced semantic priors and few-shot transfer [16].Our experiments give a zero-shot proof for such effectiveness, and TactiLDM provides a low-cost adaptive semantic alignment framework.
- *Cross-modal, Multi-level and late-fusion* methods add transformer-based fusion blocks or cross-attention layers to the VLA policy core, updating the shared latent space [19]. The additional modules require finetuning with tactile-action trajectories, yet improve the performance better on tasks [26] specified by the training data [27].

III. Method

A. Stage 1: Tactile encoder with Multimodal Contrastive Learning

In this stage, to align the tactile modality with vision and language, we employ contrastive learning between the tactile embedding $\mathbf{z}_t$ and representations from other modalities:

- Vision: A pre-trained SigLIP [15] model encodes a vision image input $\mathbf{x}_v$ into $\mathbf{z}_v = \text{SigLIP}(\mathbf{x}_v) \in \mathbb{R}^d$.
- Text: A pre-trained T5 [28] encoder processes a label text input $\mathbf{c}$ into $\mathbf{z}_c = \text{T5}(\mathbf{c}) \in \mathbb{R}^d$.
- Tactile: An encoder encodes a tactile image input $\mathbf{x}_t \in \mathbb{R}^{H\times W\times 3}$ into a latent representation via $\mathbf{z}_t = E_t(\mathbf{x}_t) \in \mathbb{R}^d$, which is then projected to the shared embedding space by a learnable tactile projector $P_t(\cdot)$, yielding $\tilde{\mathbf{z}}_t = P_t(E_t(\mathbf{x}_t)) \in \mathbb{R}^d$.We consider two variants: $E_t^{CNN}(\cdot)$, employs a convolutional neural network and $(E_t^{AE}(\cdot), D_t^{AE}(\cdot))$, employs a convolutional autoencoder.

In both variants, we aim to minimize the InfoNCE loss [29] across all modality pairs:

$$\begin{aligned}\mathcal{L}_{\text{infoNCE}}(\mathbf{z}_t, \mathbf{z}_c) = &- \log \frac{\exp(\mathbf{z}_t^\top \mathbf{z}_v/\tau)}{\sum_{\mathbf{z}'\in\mathcal{N}_v} \exp(\mathbf{z}_t^\top \mathbf{z}'/\tau)} \\ &- \log \frac{\exp(\mathbf{z}_t^\top \mathbf{z}_c/\tau)}{\sum_{\mathbf{z}'\in\mathcal{N}_c} \exp(\mathbf{z}_t^\top \mathbf{z}'/\tau)}.\end{aligned} \tag{1}$$

where τ is a temperature hyperparameter, and $\mathcal{N}_v$, $\mathcal{N}_c$ denote negative samples from the vision and text modalities, respectively.All projected embeddings are L2-normalized: $\hat{\mathbf{z}} = \mathbf{z}/\|\mathbf{z}\|_2$.

For any two modalities A and B with normalized embeddings $\{\hat{\mathbf{z}}_i^A\}_{i=1}^N$ and $\{\hat{\mathbf{z}}_i^B\}_{i=1}^N$ from a batch of N aligned samples, we define the contrastive loss as:

$$\mathcal{L}_{A\leftrightarrow B} = \frac{1}{2}\left(\frac{1}{N}\sum_{i=1}^{N} \mathcal{L}_{\text{infoNCE}}(\hat{\mathbf{z}}_i^A, \hat{\mathbf{z}}_i^B)\right), \tag{2}$$

The total contrastive loss differs between the two tactile encoder variants.

For CNN-encoder $E_t^{CNN}(\cdot)$, we leverage all three pairwise alignments to maximize cross-modal consistency

$$\mathcal{L}_{\text{cont}} = \mathcal{L}_{\text{tactile}\leftrightarrow\text{vision}} + \mathcal{L}_{\text{tactile}\leftrightarrow\text{text}} + \mathcal{L}_{\text{vision}\leftrightarrow\text{text}}. \tag{3}$$

For autoencoder $(E_t^{AE}(\cdot), D_t^{AE}(\cdot))$, we combine restruction loss of the autoencoder with $\mathcal{L}_{\text{tactile}\leftrightarrow\text{vision}}$ and $\mathcal{L}_{\text{tactile}\leftrightarrow\text{text}}$.

$(E_t^{AE}(\cdot), D_t^{AE}(\cdot))$ is trained end-to-end with a standard reconstruction loss $\mathcal{L}_{\text{recon}} = \|\mathbf{x}_t - D_t^{AE}(E_t^{AE}(\mathbf{x}_t))\|_2^2$, yielding the total loss function:

$$\mathcal{L}_1 = \mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{tactile}\leftrightarrow\text{vision}} + \mathcal{L}_{\text{tactile}\leftrightarrow\text{text}}, \tag{4}$$

. Upon convergence, we freeze the encoder $E_t^{AE}(\cdot)$ and discard the decoder for subsequent stages.

B. Stage 2: Latent Diffusion Model with Multimodal Conditioning

In the second stage, we adopt and freeze the encoder $E_t(\cdot)$ from Stage 1 to embed tactile inputs into a latent space suitable for diffusion modeling. Given a tactile image $\mathbf{x}_t$, we obtain its latent representation,

$$\mathbf{z}_0 = E_t(\mathbf{x}_t). \tag{5}$$

We then apply the standard forward diffusion process over T timesteps:

$$\mathbf{z}_t = \sqrt{\bar{\alpha}_t}\mathbf{z}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{6}$$

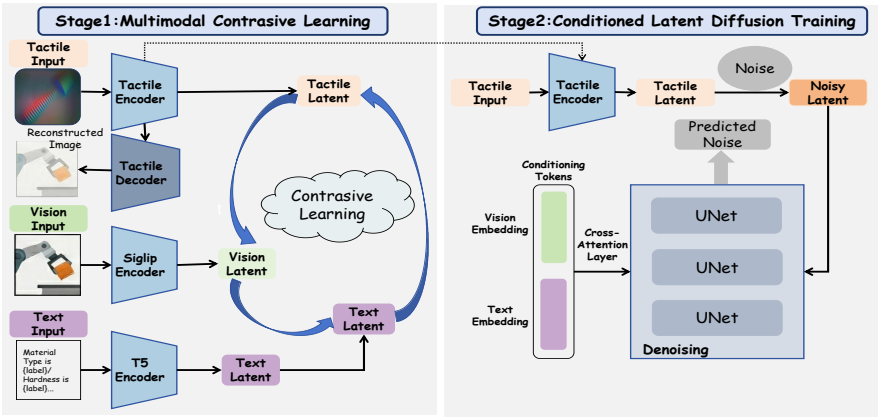

Fig. 1. Training process for auto-TactiLDM. For TactiLDM, discard tactile decoder design at Stage 1.

where $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$ controls noise scheduling.

The core of Stage 2 is a U-Net denoising network $\mathcal{U}_\theta$ that predicts the noise $\boldsymbol{\epsilon}$ given a noisy latent $\mathbf{z}_t$, timestep t, and multimodal conditioning tokens. Specifically, we form the conditioning vector by concatenating the frozen visual and textual embeddings:

$$\mathbf{c}_{\text{cond}} = [\mathbf{z}_v; \mathbf{z}_c] \in \mathbb{R}^{2d}. \tag{7}$$

During denoising, $\mathbf{c}_{\text{cond}}$ is added into cross-attention layers. The training objective is the standard diffusion loss:

$$\mathcal{L}_2 = \mathbb{E}_{\mathbf{z}_0,t,\boldsymbol{\epsilon}} \left[\|\boldsymbol{\epsilon} - \mathcal{U}_\theta(\mathbf{z}_t, t, \mathbf{c}_{\text{cond}})\|_2^2 \right]. \tag{8}$$

Depending on the architecture of the Stage-1 tactile encoder, we distinguish two variants of this model: **TactiLDM**, which uses a CNN backbone $E_t^{CNN}(\cdot)$, and **auto-TactiLDM**, which adopts an autoencoder-based encoder $E_t^{AE}(\cdot)$. Both yield a multimodal-aligned tactile representation that enables high-fidelity generation and facilitates downstream tasks.

IV. Experiment

A. Linear Probe

Datasets. We conduct our experiments on the *Touch and Go* (TAG) dataset [30], a large-scale, human-collected benchmark including paired visual frames and GelSight tactile images, along with object category annotations.

Objective. To evaluate the quality of the tactile representation, we adopt linear probing method: freeze pretrained **TactilLDM** or **auto-TactiLDM** and only train a lightweight linear classifier on top of its fixed representations. High performance on a downstream classification task under this setting strongly suggests that $E_t(\cdot)$ has captured high-level.

To prevent data leakage of ground-truth labels in classification tasks, we mask out text modality in the conditional token during inference.

Baselines. To demonstrate the effectiveness of our **TactiLDM** and **auto-TactiLDM**, we compare the performance of our learned tactile representation with a vision-based encoder **Siglip** [15], and three tactile encoder **UniTouch** [31], **ViT-Lens-2** [32],**OmniBind** [33].

Experiment Results. We evaluate the quality of our learned representations under two distinct settings: (1) a *Tactile-Only* setting, to measure the intrinsic discriminative ability of the tactile encoder, and (2) a *Tactile-Vision Fused* setting, to assess how tactile features complement visual information when combined.

*Tactile-Only Performance.*Table I summarizes the linear probing accuracy using tactile image inputs only. We observe that both variants of our proposed framework, **TactiLDM** and **auto-TactiLDM**, significantly outperform existing baselines on downstream classification tasks, where for example, in the detailed material classification task, **auto-TactiLDM** achieves an accuracy of 82.84%, surpassing the strong vision-language baseline SigLIP by over 15% and the unified tactile baseline UniTouch by over 20%.

This highlights the advantages of our method in bridging the modality gap. While SigLIP possesses strong semantic priors, it struggles to generalize to the tactile modality without explicit alignment. Additionally, while UniTouch and ViT-Lens-2 are designed for tactile data, our results suggest that the integration of Latent Diffusion Model creates a more semantically dense representation than contrastive learning alone.

*Tactile-Vision Fused Performance.*To investigate whether our learned tactile features provide information complementary to vision, we conduct an evaluation using both modalities (Table II). We concatenate the frozen embeddings from our tactile encoder with the visual embeddings from SigLIP, and for Siglip baseline, we directly concatenate both tactile and vision embeddings from itself.

While SigLIP performs strongly due to the rich semantic context, the addition of our tactile representation pushes performance to near-saturation. **TactiLDM** achieves 97.44% on material classification and 99.34% on hardness estimation. This improvement demonstrates that our model captures physical properties that are imperceptible to vision alone, thereby effectively bridging the modality gap.

Implicit Physical Encoding. Beyond explicit classification benchmarks, we further investigate the intrinsic semantic ability of our model. We observe that when **TactiLDM** is only trained with 'full' labels which help classify material types, it achieves a success rate of **92.34%** on hardness estimation (in *Tactile Input Only Setting*). This demonstrates the model's ability to *implicitly* encode fundamental material-level priors. This finding underscores the effectiveness of our Latent Diffusion framework in transferable physical representations that go beyond simple pattern matching.

TABLE I
LINEAR PROBING ACCURACY ON *TAG* (TACTILE INPUT ONLY)

Encoder	Touch and Go		
	Material	***Rough***	***Hard***
Siglip	67.19%	71.22%	73.75%
UniTouch	62.77%	64.23%	70.12%
ViT-Lens-2	64.09%	68.56%	-
OmniBind	67.45%	-	-
TactiLDM	81.18%	86.97%	**94.68%**
auto-TactiLDM	**82.84%**	**87.03%**	94.28%

Bold font denotes the best performance.

TABLE II
LINEAR PROBING ACCURACY ON *TAG* (TACTILE-VISION INPUT)

Encoder	Touch and Go		
	Material	***Rough***	***Hard***
Siglip	90.44%	92.58%	96.69%
TactiLDM	97.44%	98.89%	99.34%

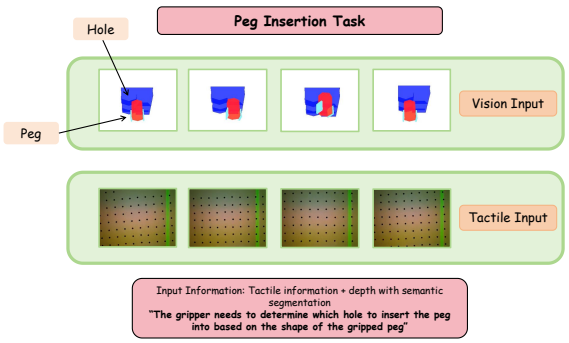

Fig. 2. Illustration for Peg Insertion Task on Maniskill-ViTac2025

B. VTLA Performance On Maniskill-Vitac 2025

Platform. We conduct our experiments on the **ManiSkill-ViTac 2025** challenge [34] simulation environment.

Our experimental setup employs the challenge's focus on *Tactile-Vision Fusion Manipulation* (Track 2). An illustration of the simulation setup is shown in Figure 2, which depicts the peg-insertion task along with visual and tactile observations.

Baseline. We adopt π_0 [35] as baseline, which employs a novel flow matching architecture built upon a pre-trained Vision-Language Model, demonstrating strong zero-shot generalization and the ability to perform dexterous tasks.

Challenges and Approach. Current tactile simulation capabilities are limited. Most existing tactile datasets are collected from real-world hardware. In simulation environments including **ManiSkill-ViTac 2025** [34], there is a scarcity of trajectory demonstrations, where VLA models are confined to interacting with the environment only through Reinforcement Learning. This restricts the application of imitation learning.

Because of these challenges, we bypass the need for fine-tuning π_0. Instead, we directly adopt π_0 inference by concatenating tactile observations into its input modality stream and evaluate it in a *zero-shot* setting.

Metrics. We employ the following metrics to quantify model's performance:

- *Success Rate (SR):* The percentage of evaluation episodes that are completed successfully according to the task-specific success criteria.
- *Average Truncation Step (ATS):* The average timestep at which an episode is terminated. The platform simulates a soft, deformable tactile sensor. If the policy causes excessive deformation during action chunk, a `tactile_movement_too_large` error is triggered, causing immediate episode truncation. A higher ATS indicates the model effectively interprets tactile images and modulates its actions to maintain safe contact.

Experimental Results.

TABLE III
ZERO-SHOT PERFORMANCE ON THE MANISKILL-VITAC2025 PEG INSERTION TASK. THE MAXIMUM EPISODE LENGTH IS 500 STEPS.

Metrics	π_0 (Baseline)	π_0 (+ Tactile Input)
SR ↑	5%	**20%**
ATS ↑	155.71	**249.98**

The experimental results in Table III demonstrate that adding tactile observations into the input stream of π_0 without any fine-tuning leads to a substantial improvement in both success rate and interaction stability.

The success rate increases from 5% to 20%, representing a **4**× improvement in zero-shot task completion. This highlights the critical role of tactile feedback in tasks requiring precise physical interaction.

Moreover, the Average Truncation Step (ATS) rises from 155.71 to 249.98 steps. This +60% increase in interaction duration demonstrates that the integration of tactile data allows the robot to modulate its actions more effectively, maintaining safer contact and avoiding immediate simulation truncation.

V. CONCLUSION

In this work, we introduced **TactiLDM**, a framework designed to bridge the gap between tactile, visual, and textual modalities through Latent Diffusion Models. By employing multimodal contrastive alignment followed by conditional latent diffusion, we successfully developed a tactile encoder

capable of capturing both semantic categories and physical attributes.

Our experiments demonstrate that these representations excel in downstream classification tasks in both tactile-only setting (≈**20%** improvement) and tactile-vision fused setting (≈5% improvement). Furthermore, our zero-shot evaluation on the ManiSkill-ViTac 2025 challenge underscores the practical value of these representations for Vision-Language-Action models, leading to a **4x** increase in success rates for peg-insertion tasks.TactiLDM bridges the gap between generative modeling and embodied perception. Future work will explore the scaling of TactiLDM to larger, more diverse tactile datasets and its integration into end-to-end control policies.

REFERENCES

[1] P. P. Liang, "Foundations of multisensory artificial intelligence," 2024. [Online]. Available: https://arxiv.org/abs/2404.18976

[2] S. Yang, H. Li, Y. Chen, B. Wang, Y. Tian, T. Wang, H. Wang, F. Zhao, Y. Liao, and J. Pang, "Instructvla: Vision-language-action instruction tuning from understanding to manipulation," 2025. [Online]. Available: https://arxiv.org/abs/2507.17520

[3] Y. Ma, Z. Song, Y. Zhuang, J. Hao, and I. King, "A survey on vision-language-action models for embodied ai," 2025. [Online]. Available: https://arxiv.org/abs/2405.14093

[4] H. Zhen, X. Qiu, P. Chen, J. Yang, X. Yan, Y. Du, Y. Hong, and C. Gan, "3d-vla: A 3d vision-language-action generative world model," 2024. [Online]. Available: https://arxiv.org/abs/2403.09631

[5] C. Gu, J. Liu, H. Chen, R. Huang, Q. Wuwu, Z. Liu, X. Li, Y. Li, R. Zhang, P. Jia, P.-A. Heng, and S. Zhang, "Manualvla: A unified vla model for chain-of-thought manual generation and robotic manipulation," 2025. [Online]. Available: https://arxiv.org/abs/2512.02013

[6] P. Hao, C. Zhang, D. Li, X. Cao, X. Hao, S. Cui, and S. Wang, "Tla: Tactile-language-action model for contact-rich manipulation," 2025. [Online]. Available: https://arxiv.org/abs/2503.08548

[7] M. A. Lee, Y. Zhu, P. Zachares, M. Tan, K. Srinivasan, S. Savarese, L. Fei-Fei, A. Garg, and J. Bohg, "Making sense of vision and touch: Learning multimodal representations for contact-rich tasks," *IEEE Transactions on Robotics*, vol. 36, no. 3, pp. 582–596, 2020.

[8] Íñigo Elguea-Aguinaco, A. Serrano-Muñoz, D. Chrysostomou, I. Inziarte-Hidalgo, S. Bøgh, and N. Arana-Arexolaleiba, "A review on reinforcement learning for contact-rich robotic manipulation tasks," *Robotics and Computer-Integrated Manufacturing*, vol. 81, p. 102517, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0736584522001995

[9] H. Liu, J. Qin, F. Sun, and D. Guo, "Extreme kernel sparse learning for tactile object recognition," *IEEE Transactions on Cybernetics*, vol. 47, no. 12, pp. 4509–4520, 2017.

[10] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," 2020. [Online]. Available: https://arxiv.org/abs/1906.05849

[11] G. Cao, J. Jiang, D. Bollegala, and S. Luo, "Learn from incomplete tactile data: Tactile representation learning with masked autoencoders," 2023. [Online]. Available: https://arxiv.org/abs/2307.07358

[12] G. Khurana, L. Wei, and D. Zhang, "Sarl: Spatially-aware self-supervised representation learning for visuo-tactile perception," 2025. [Online]. Available: https://arxiv.org/abs/2512.01908

[13] S. Targ, D. Almeida, and K. Lyman, "Resnet in resnet: Generalizing residual architectures," 2016. [Online]. Available: https://arxiv.org/abs/1603.08029

[14] J. Zhao, Y. Ma, L. Wang, and E. H. Adelson, "Transferable tactile transformers for representation learning across diverse sensors and tasks," 2024. [Online]. Available: https://arxiv.org/abs/2406.13640

[15] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," 2023. [Online]. Available: https://arxiv.org/abs/2303.15343

[16] J. Bi, K. Y. Ma, C. Hao, M. Z. Shou, and H. Soh, "Vla-touch: Enhancing vision-language-action models with dual-level tactile feedback," 2025. [Online]. Available: https://arxiv.org/abs/2507.17294

[17] S. Yu, K. Lin, A. Xiao, J. Duan, and H. Soh, "Octopi: Object property reasoning with large tactile-language models," 2024. [Online]. Available: https://arxiv.org/abs/2405.02794

[18] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "Llama: Open and efficient foundation language models," 2023. [Online]. Available: https://arxiv.org/abs/2302.13971

[19] J. Huang, S. Wang, F. Lin, Y. Hu, C. Wen, and Y. Gao, "Tactile-vla: Unlocking vision-language-action model's physical knowledge for tactile generalization," 2025. [Online]. Available: https://arxiv.org/abs/2507.09160

[20] M. Polic, I. Krajacic, N. Lepora, and M. Orsag, "Convolutional autoencoder for feature extraction in tactile sensing," *IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp. 3671–3678, 2019.

[21] C. Zhang, P. Hao, X. Cao, X. Hao, S. Cui, and S. Wang, "Vtla: Vision-tactile-language-action model with preference learning for insertion manipulation," 2025. [Online]. Available: https://arxiv.org/abs/2505.09577

[22] B. Huang, Y. Wang, X. Yang, Y. Luo, and Y. Li, "3d-vitac: Learning fine-grained manipulation with visuo-tactile sensing," 2025. [Online]. Available: https://arxiv.org/abs/2410.24091

[23] R. Feng, J. Hu, W. Xia, T. Gao, A. Shen, Y. Sun, B. Fang, and D. Hu, "Anytouch: Learning unified static-dynamic representation across multiple visuo-tactile sensors," 2025. [Online]. Available: https://arxiv.org/abs/2502.12191

[24] R. Sapkota, Y. Cao, K. I. Roumeliotis, and M. Karkee, "Vision-language-action models: Concepts, progress, applications and challenges," 2025. [Online]. Available: https://arxiv.org/abs/2505.04769

[25] X. Wei, H. Zhang, X. Cao, S. Xie, W. Ge, Y. Li, and C. Wang, "Audio-vla: Adding contact audio perception to vision-language-action model for robotic manipulation," 2025. [Online]. Available: https://arxiv.org/abs/2511.09958

[26] T. Li, C. Yu, Y. Yan, D. Song, Y. Shuai, Y. Wang, and G. Chen, "Vtlg: A vision-tactile-language grasp generation method oriented towards task," *Robotics and Computer-Integrated Manufacturing*, vol. 98, p. 103152, 2026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0736584525002066

[27] M. Ud Din, W. Akram, L. Saad Saoud, J. Rosell, and I. Hussain, "Multimodal fusion with vision-language-action models for robotic manipulation: A systematic review," *Information Fusion*, vol. 129, p. 104062, 2026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253525011248

[28] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," 2023. [Online]. Available: https://arxiv.org/abs/1910.10683

[29] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2019. [Online]. Available: https://arxiv.org/abs/1807.03748

[30] F. Yang, C. Ma, J. Zhang, J. Zhu, W. Yuan, and A. Owens, "Touch and go: Learning from human-collected vision and touch," 2022. [Online]. Available: https://arxiv.org/abs/2211.12498

[31] F. Yang, C. Feng, Z. Chen, H. Park, D. Wang, Y. Dou, Z. Zeng, X. Chen, R. Gangopadhyay, A. Owens, and A. Wong, "Binding touch to everything: Learning unified multimodal tactile representations," 2024. [Online]. Available: https://arxiv.org/abs/2401.18084

[32] W. Lei, Y. Ge, K. Yi, J. Zhang, D. Gao, D. Sun, Y. Ge, Y. Shan, and M. Z. Shou, "Vit-lens: Towards omni-modal representations," 2024. [Online]. Available: https://arxiv.org/abs/2311.16081

[33] Y. Lyu, X. Zheng, D. Kim, and L. Wang, "Omnibind: Teach to build unequal-scale modality interaction for omni-bind of all," 2024. [Online]. Available: https://arxiv.org/abs/2405.16108

[34] C. Li, R. Dang, X. Li, Z. Wu, J. Xu, H. Kasaei, R. Calandra, N. Lepora, S. Luo, H. Su, and R. Chen, "Maniskill-vitac 2025: Challenge on manipulation skill learning with vision and tactile sensing," 2024. [Online]. Available: https://arxiv.org/abs/2411.12503

[35] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter, S. Jakubczak, T. Jones, L. Ke, S. Levine, A. Li-Bell, M. Mothukuri, S. Nair, K. Pertsch, L. X. Shi, J. Tanner, Q. Vuong, A. Walling, H. Wang, and U. Zhilinsky, "π_0: A vision-language-action flow model for general robot control," 2026. [Online]. Available: https://arxiv.org/abs/2410.24164