VQ-Touch: A Data-Efficient Tactile Generation Framework Across Sensors and Scenarios

***Abstract*—Tactile image generation significantly reduces the dependency on expensive and wear-prone sensors by synthesizing high-fidelity tactile data, offering an efficient solution for tactile information acquisition in robotic perception and human-machine interaction systems. However, existing methods depend on large-scale, diverse datasets from specific sensors and lack efficient data utilization and robust generalization capabilities, struggling in vision-limited environments. To address this, we introduce *VQ-Touch*, a tactile generation framework that supports both cross-sensor and multi-scenario applications. Specifically, to efficiently extract complex deformation and texture features from the data, we propose DM-VQGAN, an effective tactile representation learner. Furthermore, we introduce a discrete diffusion decoder with a unified conditioning interface, supporting multimodal generation tasks such as images and labels, and enhances the model's generalization capability through few-shot mixed training, thus achieving compatibility with current mainstream sensors and their variants. Experiments show that *VQ-Touch* surpasses state-of-the-art methods in multiple tasks.**

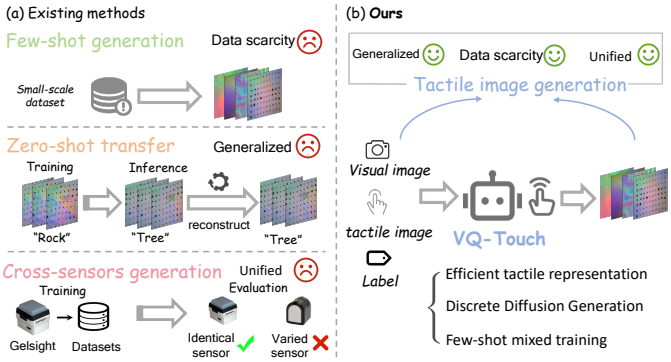

Fig. 1: **The illustration of tactile generation tasks and comparison of different methods.** Leveraging data-efficient tactile representations and a unified discrete diffusion model, ***VQ-Touch*** supports few-shot, zero-shot transfer, and cross-sensor tactile generation.

I. Introduction

Touch is a critical modality for human interaction with the environment, providing rich information about geometry, texture, and contact forces, and is therefore increasingly exploited in robotic manipulation [2, 20]. Vision-based tactile sensors (VBTS) [13], which use a camera to capture fine-grained elastomer deformations at high resolution and low cost, have emerged as the dominant tactile sensing paradigm. However, their limited durability and the laborious real-world acquisition pipeline render tactile data substantially scarcer and noisier than visual data [26, 16]. This bottleneck has spurred growing interest in synthesizing high-fidelity tactile images as a scalable surrogate for physical collection, lowering the training cost of downstream tasks such as robotic manipulation [3] and texture recognition [1, 17, 21, 15].

The fundamental paradigm of tactile image generation is to first extract diverse tactile feature representations from existing data and then synthesize tactile images that satisfy specific scene requirements. Although previous work has made significant progress [14], two challenges remain: ❶ the need for large-scale and diverse datasets for pretraining each sensor type [22, 27], which results in significantly degraded performance in environments where tactile data acquisition is difficult; and ❷ the inability of learned features to generalize to image generation tasks involving other sensors. This naturally leads to the following question: ***Can we learn a data-efficient tactile representation that facilitates tactile generation across sensors and scenarios?***

To address this challenge, we introduce ***VQ-Touch***, a tactile generation framework that supports cross-sensor and multi-scenario applications. First, to model the low-variance latent space of tactile images while preserving their rich informational content, we adopt a VQGAN-based architecture, incorporating deformable convolutions and a multi-scale fusion module to extract both macro-level deformation patterns and micro-level texture features from tactile images. This enhances its applicability to vision-based tactile sensors (VBTS). Additionally, we propose a few-shot mixed training approach that enables the transfer of learned tactile representations from existing sensor data to image generation tasks for new sensors, facilitating data-efficient cross-sensor generalization. Based on these learned representations, we further train a discrete diffusion model to support multiple input modalities and perform conditional generation of corresponding tactile images. Our contributions are summarized as follows:

- We introduce **DM-VQGAN**, an efficient tactile representation learner that captures generalizable, information-rich features tailored to tactile data.
- We propose a **unified multimodal conditional generation framework** that supports multiple input modalities, including tactile images, visual images, and semantic labels, to address diverse tactile generation requirements.
- We model tactile sensors at the sensor-family level and use clustering and **few-shot mixed training** to transfer features from one sensor to its family, eliminating full multi-dataset training.
- ***VQ-Touch*** outperforms state-of-the-art models in tactile image reconstruction and generation, excelling in vision-limited scenarios.

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

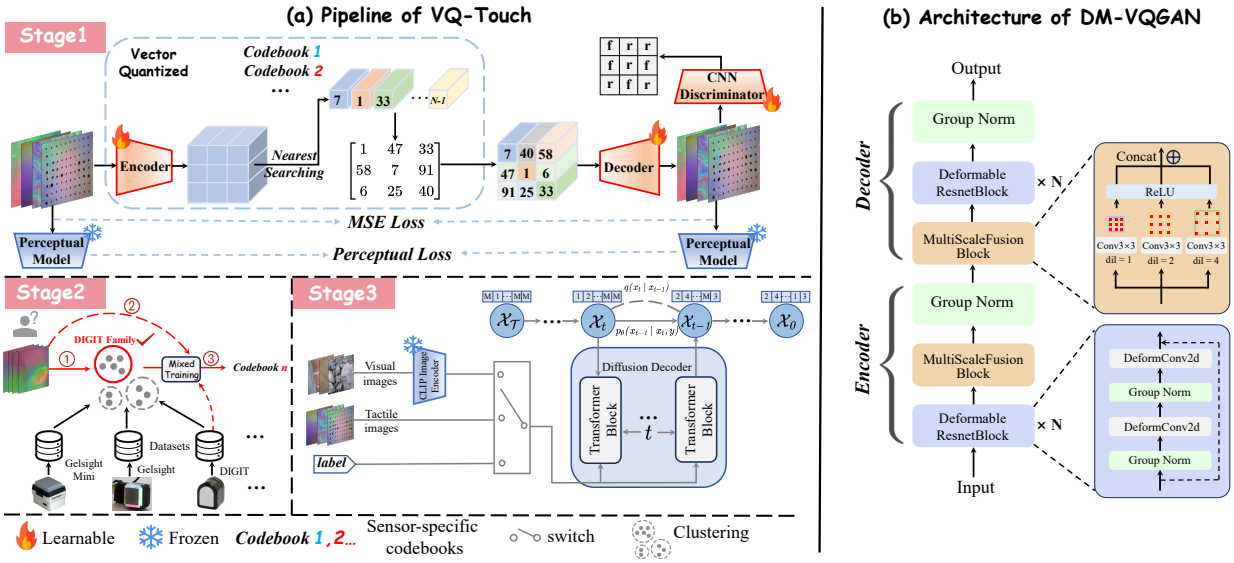

Fig. 2: **Overview of the proposed *VQ-Touch* framework.** The framework consists of three stages: **(1)** learning efficient tactile features using the DM-VQGAN encoder-decoder architecture, which is based on discrete codebooks for representation, **(2)** acquiring sensor-family-specific codebooks through clustering and few-shot mixed training method, and **(3)** by leveraging the discrete latent space and encoder learned by DM-VQGAN, a discrete diffusion decoder is employed to generate high-quality tactile images conditioned on unimodal inputs including visual images, tactile images or semantic labels.

II. Methodology

Overview. In this section, we elaborate on ***VQ-Touch***, whose pipeline is shown in Fig.2. It consists of three components: **(i) DM-VQGAN**, a tactile-specific encoder that models and amplifies geometric variations to produce generation-friendly latent representations (**Sec.II-A**); **(ii)** a **Few-shot Mixed Training** scheme that transfers these representations across sensor families for data-efficient cross-sensor generalization (**Sec.II-A**); and **(iii)** a unified **discrete-diffusion** generative framework that conditions on multimodal inputs and uses the pretrained encoder to synthesize tactile images (**Sec.II-B**).

A. DM-VQGAN

Vision-based tactile images discard color while preserving essential cues such as morphology, geometry, and force distribution [12]. Exploiting this low intrinsic variance together with VQGAN's strong latent-space modeling [5], we adopt a discrete codebook to obtain structured representations under limited data. Vanilla VQGAN, however, relies on fixed convolutions and thus struggles to capture the non-rigid deformations and multi-scale patterns characteristic of VBTS [12, 6]. We therefore propose **DM-VQGAN**, which integrates deformable and multi-scale dilated convolutions to strengthen feature extraction and representation, as shown in Fig. 2(b).

Training Process of DM-VQGAN. Similar to the classic VQGAN [7], DM-VQGAN consists of an encoder E, a decoder G, a patch-based discriminator D, and a codebook $Z = \{z_k\}_{k=1}^{K}$, containing K discrete codes, as shown in Stage1 of Fig. 2(a). Given an input image $x \in \mathbb{R}^{H\times W\times 3}$, the encoder E first extracts a latent feature representation $\hat{z} \in \mathbb{R}^{H_f\times W_f\times n_z}$, where n_z and f represent the dimensionality of the latent features and the spatial compression ratio, respectively. Next, the feature vector at each spatial location (i, j) is quantized to the closest discrete code from the codebook via nearest neighbor search:

$$z_q = q(\hat{z}) := \left(\arg\min_{z_k\in Z} \|\hat{z}_{ij} - z_k\|\right) \in \mathbb{R}^{H_f\times W_f\times n_z}, \quad (1)$$

The decoder G then decodes the quantized features back into the image space, i.e., $\hat{x} = G(z_q)$. The training objective of DM-VQGAN is to minimize the difference between the input image and the reconstructed image, and the corresponding loss function can be expressed as:

$$L_{\text{total}} = \underbrace{\|x - \hat{x}\|_1 + \|\text{sg}[E(x)] - z_q\|_2^2 + \beta\|\text{sg}[z_q] - E(x)\|_2^2}_{L_{\text{vq}}} + L_{\text{per}} + L_{\text{adv}}, \quad (2)$$

where L_{per} and L_{adv} represent the perceptual loss and the adversarial loss. In the reconstruction loss L_{vq}, $\text{sg}[\cdot]$ denotes the stop-gradient operation, and $\beta\|\text{sg}[z_q] - E(x)\|_2^2$ is known as the commitment loss, where the commitment weight β is set to 0.25 [9, 24].

Deformation Pattern Extraction: In each encoder and decoder, we introduce deformable convolutions [6] to adaptively learn dynamic spatial offsets for convolution kernels, enabling the model to better capture local, complex, and irregular deformation patterns present in tactile images. The structure of the deformable convolution layer, denoted as:

$$y(p_0) = \sum_{p_k\in\mathcal{G}} w_k \cdot x\left(p_0 + p_k + \Delta p_k\right), \quad (3)$$

where $y(p_0)$ is the output feature vector at position p_0, and $p_k \in \mathbb{R}^2$, $\Delta p_k \in \mathbb{R}^2$ represent the pre-defined offset and

learned offset for the k-th sampling point. This flexible convolutional sampling mechanism facilitates effective learning of dynamic spatial relationships and feature variations, thereby enhancing sensitivity to localized deformations during tactile representation learning.

Multi-scale Fusion: To effectively capture the rich, multi-scale features inherent in tactile images, we design a multi-scale fusion module that employs parallel dilated convolutions with varying dilation rates (i.e., 1, 2, and 4), denoted as:

$$y = \mathcal{F}_{\text{conv3x3}}\left(\text{Concat}\left(\left\{\sigma\left(\mathcal{F}_{\text{dilate}}^{(d_i)}(x)\right)\right\}_{d_i\in\{1,2,4\}}\right)\right). \quad (4)$$

This design expands the receptive field, enabling the simultaneous extraction of macro-level global deformations and micro-level textural features. Consequently, the DM-VQGAN effectively addresses the limitations of traditional convolutional structures in representing complex dynamic deformations and multi-scale features of tactile images, enhancing tactile representation and improving tactile image generation.

B. Few-shot Mixed Training for Cross-sensor Generalization

Many studies [23, 19] have observed substantial discrepancies among tactile sensors, which hinder the generalization of tactile representation models. Existing methods, such as UniTouch [23] and T3 [27], typically learn sensor-specific features from multiple large-scale datasets. Although effective, such a strategy is data-intensive and potentially redundant.

By analyzing mainstream tactile datasets [22, 8], we identify two major sources of variation in tactile images: intrinsic sensor properties and environmental factors. The former are determined by sensor design and manufacturing and remain relatively stable, while the latter, including illumination, calibration, and material degradation, introduce appearance shifts that may obscure consistent tactile patterns. We therefore hypothesize that sensors A_1 and A_2 with shared intrinsic properties but different environmental conditions belong to the same sensor family A, whose representations can be jointly learned. Consequently, a model trained on A_1 can be adapted to A_2 with only minimal additional data.

To identify the family of an unseen sensor, we propose an unsupervised clustering strategy. Specifically, DM-VQGAN is used to extract tactile features from images of multiple sensor types, followed by cross-sensor clustering. For each cluster, we compute its prototype vector by iteratively estimating feature centroids, and define the inter-cluster distance as

$$D(C_i, C_j) = \left\| \frac{1}{|C_i|}\sum_{x\in C_i} x - \frac{1}{|C_j|}\sum_{y\in C_j} y \right\|, \quad i,j \in \mathbb{N}, \quad (5)$$

where $|C_i|$ and $|C_j|$ denote the numbers of samples in clusters C_i and C_j, respectively, and x and y are their corresponding feature vectors. Given an unseen sensor, we compute its prototype vector and assign it to the nearest sensor family. We then construct a small-sample mixed dataset to train DM-VQGAN for tactile representation learning, avoiding full retraining on all datasets.

TABLE I: **Tactile image reconstruction.** We compare tactile image reconstruction across different methods and sample sizes, evaluated by FID (↓) and SSIM (↑).

Method	FID(↓)			SSIM(↑)			Mask Ratio
Sample Size	100	1K	10K	100	1K	10K	
MAE	160.11	129.45	117.42	0.680	0.695	0.712	**0.25**
	205.32	209.22	186.11	0.561	0.635	0.645	**0.5**
	270.04	235.10	220.86	0.480	0.582	0.598	**0.75**
VQGAN	46.76	18.34	16.54	0.823	0.911	0.923	\
DM-VQGAN	46.02	17.25	**15.62**	0.831	0.917	**0.936**	\

C. Multimodal Conditional Generation Framework

Although DM-VQGAN learns effective tactile representations, its original decoder cannot flexibly incorporate external prompts such as images. We therefore adopt discrete diffusion [10], which naturally operates in discrete latent spaces, to build a unified multimodal conditional generation framework. As shown in Stage 3 of Fig. 2(a), the diffusion decoder is conditioned on a single modality at each training step:

$$\boldsymbol{c} = f_m(y_m), \quad m \in \{\textit{visual}, \textit{tactile}, \textit{label}\}, \quad (6)$$

where $f_m(\cdot)$ denotes the encoder for modality m, and $\boldsymbol{c}$ is the corresponding conditional latent feature. This design enables flexible tactile image generation from visual, tactile, or label conditions according to task requirements. For visual conditioning, we use a frozen pretrained CLIP encoder [18], i.e., $\boldsymbol{c}_{\text{visual}} = f_V^{\text{CLIP}}(y_{\text{visual}})$. For tactile conditioning, the input y_{tactile} is mapped to discrete indices using the sensor-specific codebook $\mathcal{Z} = \{z_k\}_{k=1}^{K}$ learned by DM-VQGAN:

$$\boldsymbol{c}_{\text{tactile}} = q\left(f_{\text{DM-VQGAN}}(y_{\text{tactile}}),\, \mathcal{Z}\right). \quad (7)$$

The selected conditional feature $\boldsymbol{c}$ then interacts with the noisy latent representation $\boldsymbol{z}_t$ through transformer modules in the diffusion decoder, enabling condition-specific tactile synthesis.

III. Experiments

A. Dataset and Baselines

We train and evaluate the proposed model on four public tactile datasets, namely Touch and Go [22], FabricVST [4], VisGel [25], and HCT [8], covering GelSight, GelSight Mini, and DIGIT sensors. We compare our method with representative tactile generation approaches [23] across multiple downstream tasks. For fairness, all experiments are repeated three times, and the average results are reported.

B. Results

Effectiveness Analysis of DM-VQGAN for Tactile Representation Learning. To evaluate the effectiveness of DM-VQGAN for tactile representation learning, we conduct reconstruction experiments on the FabricVST dataset [4] under varying training-sample scales, using MAE [11] and VQGAN [7] as baselines. As summarized in Table I, DM-VQGAN consistently achieves superior reconstruction quality, preserving fine tactile structures across all data scales. In contrast, MAE fails to recover detailed textures (Fig. 3(b)), and its performance

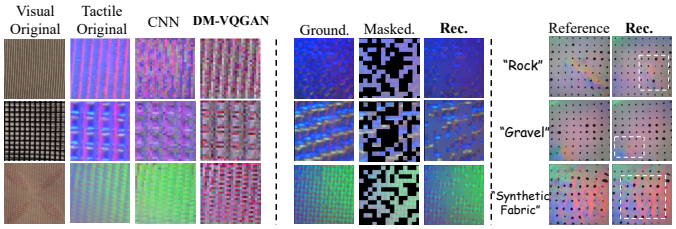

Fig. 3: **Latent space visualization and reconstruction comparison. (a)** shows RGB latent features from three encoders. **(b)** presents the reconstruction results of MAE with a ViT-Base backbone. **(c)** shows zero-shot reconstruction by ***VQ-Touch***, demonstrating generalization to unseen objects.

TABLE II: **Quantitative Results on Tactile Generation Task.** Our method achieves the best results on most metrics.

Method	HCT				SSVTP			
	CTTP ↑	LPIPS ↓	SSIM ↑	PSNR ↑	CTTP ↑	LPIPS ↓	SSIM ↑	PSNR ↑
GVST	-	0.573	0.881	19.45	-	0.502	0.918	21.15
UniTouch	0.156	0.528	0.902	19.84	0.127	0.555	0.824	12.42
PixArt-α	0.198	0.504	0.876	20.26	0.125	0.497	0.916	22.35
TextToucher	0.261	0.427	0.904	22.70	0.152	0.465	0.930	22.43
Ours	**0.278**	**0.216**	**0.910**	**24.35**	**0.158**	**0.231**	**0.932**	**25.56**

with $N = 10$k remains inferior to DM-VQGAN trained with only $N = 100$ samples. Furthermore, latent-space visualization (Fig. 3(a)) shows that, unlike CNN autoencoders that mainly retain color distributions, DM-VQGAN learns more structured and compact representations that emphasize salient tactile features while suppressing background variations. This demonstrates its effectiveness in learning discriminative and low-variance tactile representations.

Evaluation Across Multiple Tasks and Applications. We evaluate the proposed framework across data-limited learning, multimodal generation, and downstream recognition tasks. Under limited samples and categories, our method demonstrates superior few-shot representation learning and zero-shot generalization to unseen objects, consistently outperforming competing approaches (Tab. IV, Fig. 3(c)). We further assess vision- and label-conditioned tactile image generation using real-world scene images and semantic labels. As shown in Fig. 4 and Tab. II, ***VQ-Touch*** produces tactile images that maintain strong cross-modal consistency, structural fidelity, and semantic alignment. Finally, classifiers trained on generated tactile images achieve higher object recognition accuracy than baseline methods (Tab. III), demonstrating the practical value of the generated data for downstream applications.

Validation of Transferability and Generalizability across Datasets and Sensors. As discussed in Sec. II-B, we improve the generalizability of ***VQ-Touch*** across different sensors and datasets within the same sensor family through few-shot mixed training. Specifically, a small number of tactile images from unseen sensors are combined with existing datasets from their corresponding sensor families to train DM-VQGAN. As illustrated in Fig. 5, the learned codebook can be directly transferred to tactile image reconstruction for unseen sensors, demonstrating the effectiveness of our approach in enabling cross-sensor generalization under limited-data settings.

TABLE III: Performance of Tactile Images Generated by VQ-Touch on Object Recognition Tasks Across Different Datasets.

Method	Dataset	Top-1 Acc (%)	Precision (%)	Recall	F1 score
PixArt-α	TAG	81.26	81.8	0.82	0.819
Ours	TAG	**84.53**	**85.1**	**0.84**	**0.845**
PixArt-α	FabricVST	88.79	89.4	0.90	0.897
Ours	FabricVST	**92.48**	**92.7**	**0.92**	**0.923**

TABLE IV: **Comparison of few-shot tactile image reconstruction.**

method	FID(↓)	LPIPS(↓)	SSIM(↑)	Sample Size
Unitouch [23]	72.43	0.583	0.615	N=500
	56.71	0.541	0.663	N=2k
Ours	20.71	0.421	0.863	N=500
	17.03	0.408	0.894	N=2K

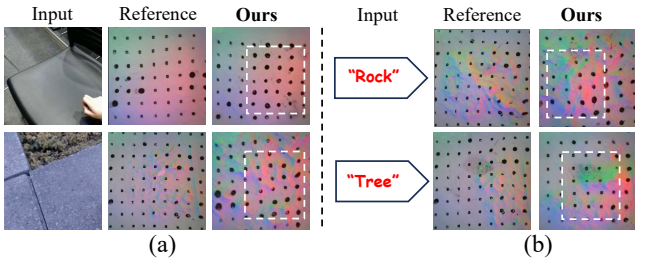

Fig. 4: **Tactile image generation under vision or label conditions.** ***VQ-Touch*** can also generate high-fidelity tactile images from **(a)** vision inputs or **(b)** label inputs.

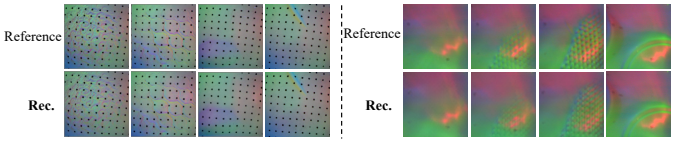

Fig. 5: **Cross-Sensor Tactile Image Reconstruction.** We evaluate reconstruction performance by training with a few-shot mixed training method. The dataset consists of a small number of unseen sensor data (50 images from GelSight Derivative Version I and DIGIT) combined with data from their respective sensor families at a 1:5 ratio.

IV. Conclusion

In this work, we propose ***VQ-Touch***, a novel and data-efficient framework for tactile image generation. ***VQ-Touch*** establishes a new paradigm for few-shot tactile generation by leveraging the representation learner DM-VQGAN and a unified conditional diffusion encoder, enabling cross-sensor compatibility and vision-free multimodal-to-tactile generation. Our framework demonstrates superior performance over SOTA methods while requiring limited training data. The proposed sensor-family clustering and mixed training strategy further enhance generalizability. These findings highlight the potential of ***VQ-Touch*** as an effective solution for scalable tactile generation and efficient tactile perception in robotics and human computer interaction applications.

References

[1] Vikas Raviraj Acharya and Vidya S Rao. Exploring modern sensor in robotics: A review. In *2024 Asia Pacific Conference on Innovation in Technology (APCIT)*, pages 1–5. IEEE, 2024.

[2] Roberto Calandra, Andrew Owens, Dinesh Jayaraman, Justin Lin, Wenzhen Yuan, Jitendra Malik, Edward H Adelson, and Sergey Levine. More than a feeling: Learning to grasp and regrasp using vision and touch. *IEEE Robotics and Automation Letters*, 3(4):3300–3307, 2018.

[3] Guanqun Cao and Shan Luo. Multimodal perception for dexterous manipulation. In *Tactile Sensing, Skill Learning, and Robotic Dexterous Manipulation*, pages 45–58. Elsevier, 2022.

[4] Guanqun Cao, Jiaqi Jiang, Danushka Bollegala, Min Li, and Shan Luo. Multimodal zero-shot learning for tactile texture recognition. *Robotics and Autonomous Systems*, 176:104688, 2024.

[5] Shiyue Cao, Yueqin Yin, Lianghua Huang, Yu Liu, Xin Zhao, Deli Zhao, and Kaigi Huang. Efficient-vqgan: Towards high-resolution image generation with efficient vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7368–7377, 2023.

[6] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017.

[7] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021.

[8] Letian Fu, Gaurav Datta, Huang Huang, William Chung-Ho Panitch, Jaimyn Drake, Joseph Ortiz, Mustafa Mukadam, Mike Lambeta, Roberto Calandra, and Ken Goldberg. A touch, vision, and language dataset for multimodal alignment. *arXiv preprint arXiv:2402.13232*, 2024.

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[10] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10696–10706, 2022.

[11] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[12] Rui Li and Edward H Adelson. Sensing and recognizing surface textures using a gelsight sensor. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1241–1247, 2013.

[13] Sandra Q Liu and Edward H Adelson. Gelsight fin ray: Incorporating tactile sensing into a soft compliant robotic gripper. In *2022 IEEE 5th International Conference on Soft Robotics (RoboSoft)*, pages 925–931. IEEE, 2022.

[14] Fotios Lygerakis, Vedant Dave, and Elmar Rueckert. M2curl: Sample-efficient multimodal reinforcement learning via self-supervised representation learning for robotic manipulation. In *2024 21st International Conference on Ubiquitous Robots (UR)*, pages 490–497. IEEE, 2024.

[15] Kailin Lyu, Di Wu, Long Xiao, Jianning Zeng, Jianwei He, Chang Lin, Lianyu Hu, Lin Shu, Jie Hao, and Ce Hao. Tacreasoner: A dynamic tactile-language framework for interactive reasoning in real-world scenarios. *arXiv preprint arXiv:2607.05131*, 2026.

[16] Kailin Lyu, Di Wu, Pengwei Zhang, Yuhang Zheng, Yingxin Lai, Long Xiao, Kangyi Wu, Pengna Li, Chen Gao, Lianyu Hu, et al. Touchthinker: Scaling tactile commonsense reasoning to the open world with large-scale data and action-aware representation. *arXiv preprint arXiv:2606.11637*, 2026.

[17] Kailin Lyu, Long Xiao, Jianing Zeng, Junhao Dong, Xuexin Liu, Zhuojun Zou, Haoyue Yang, Lin Shu, and Jie Hao. Touchformer: A robust transformer-based framework for multimodal material perception. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 18496–18504, 2026.

[18] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[19] Yuhao Sun, Ning Cheng, Shixin Zhang, Wenzhuang Li, Lingyue Yang, Shaowei Cui, Huaping Liu, Fuchun Sun, Jianwei Zhang, Guo Di, et al. Tactile data generation and applications based on visuo-tactile sensors: A review. *Information Fusion*, page 103162, 2025.

[20] Neha Sunil, Shaoxiong Wang, Yu She, Edward Adelson, and Alberto Rodriguez Garcia. Visuotactile affordances for cloth manipulation with local control. In *Conference on Robot Learning*, pages 1596–1606. PMLR, 2023.

[21] Long Xiao, Kailin Lyu, Jianwei He, Jianing Zeng, Lin Shu, and Jie Hao. Tacexpert: A pseudo-temporal mixture-of-experts framework for open-set tactile object recognition. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3761–3765. IEEE, 2026.

[22] Fengyu Yang, Chenyang Ma, Jiacheng Zhang, Jing Zhu, Wenzhen Yuan, and Andrew Owens. Touch and go: Learning from human-collected vision and touch. *arXiv preprint arXiv:2211.12498*, 2022.

[23] Fengyu Yang, Chao Feng, Ziyang Chen, Hyoungseob Park, Daniel Wang, Yiming Dou, Ziyao Zeng, Xien Chen, Rit Gangopadhyay, Andrew Owens, et al. Binding touch to everything: Learning unified multimodal tactile representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26340–26353, 2024.

[24] Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved vqgan. *arXiv preprint arXiv:2110.04627*, 2021.

[25] Wenzhen Yuan, Yuchen Mo, Shaoxiong Wang, and Edward H Adelson. Active clothing material perception using tactile sensing and deep learning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4842–4849. IEEE, 2018.

[26] Shixin Zhang, Zixi Chen, Yuan Gao, Weiwei Wan, Jianhua Shan, Hongxiang Xue, Fuchun Sun, Yiyong Yang, and Bin Fang. Hardware technology of vision-based tactile sensor: A review. *IEEE Sensors Journal*, 22(22): 21410–21427, 2022.

[27] Jialiang Zhao, Yuxiang Ma, Lirui Wang, and Edward H Adelson. Transferable tactile transformers for representation learning across diverse sensors and tasks. *arXiv preprint arXiv:2406.13640*, 2024.