TextToucher: Fine-Grained Text-to-Touch Generation

Jiahang Tu[1], **Hao Fu**[1], **Fengyu Yang**[2, 3], **Hanbin Zhao**[*1], **Chao Zhang**[1], **Hui Qian**[1]

[1]College of Computer Science and Technology, Zhejiang University
[2]College of Computer Science and Technology, Yale University
[3]UniX AI
{tujiahang, haof.pizazz, zhaohanbin, zczju, qianhui}@zju.edu.cn, fengyu.yang@yale.edu

Abstract

Tactile sensation plays a crucial role in the development of multi-modal large models and embodied intelligence. To collect tactile data with minimal cost as possible, a series of studies have attempted to generate tactile images by vision-to-touch image translation. However, compared to text modality, visual modality-driven tactile generation cannot accurately depict human tactile sensation. In this work, we analyze the characteristics of tactile images in detail from two granularities: object-level (tactile texture, tactile shape), and sensor-level (gel status). We model these granularities of information through text descriptions and propose a fine-grained Text-to-Touch generation method (TextToucher) to generate high-quality tactile samples. Specifically, we introduce a multi-modal large language model to build the text sentences about object-level tactile information and employ a set of learnable text prompts to represent the sensor-level tactile information. To better guide the tactile generation process with the built text information, we fuse the dual grains of text information and explore various dual-grain text conditioning methods within the diffusion transformer architecture. Furthermore, we propose a Contrastive Text-Touch Pre-training (CTTP) metric to precisely evaluate the quality of text-driven generated tactile data. Extensive experiments demonstrate the superiority of our TextToucher method.

Code — https://github.com/TtuHamg/TextToucher

Introduction

Tactile sensation is one of the earliest developed senses in humans (Ackerman, Nocera, and Bargh 2010). Infants begin to explore the world by touching objects, which enables them to build up their cognition of texture and shape. In the field of Multimodal Large Language Models (MLLM), researchers recognize the importance of tactile sensation in physical reasoning (Yu et al. 2024; Wang et al. 2023) and tactile sensation is regarded as an important component in multimodal learning (Fu et al. 2024; Zambelli et al. 2021; Rodriguez et al. 2024b,a; Li et al. 2024). It also plays a fundamental role in embodied intelligence to interact with the environments (Wu et al. 2024; Barreiros et al. 2022). To construct effective tactile-based multimodal models, a large volume of high-quality tactile data is required; however, obtaining tactile data is not straightforward. On the one hand, manual collection (Yang et al. 2022; Fu et al. 2024; Song et al. 2020a; Sundaram et al. 2019; Owens et al. 2016) is quite labor-intensive and time-consuming; on the other hand, robotic collection(Calandra et al. 2018; Li et al. 2019; Murali et al. 2018; Kerr et al. 2022) limits the ways of interacting with objects, lacking flexibility and diversity. Therefore, high-quality generation of tactile data(Yang et al. 2024; Yang, Zhang, and Owens 2023; Gao, Yuan, and Zhu 2023) is becoming a frontier research area.

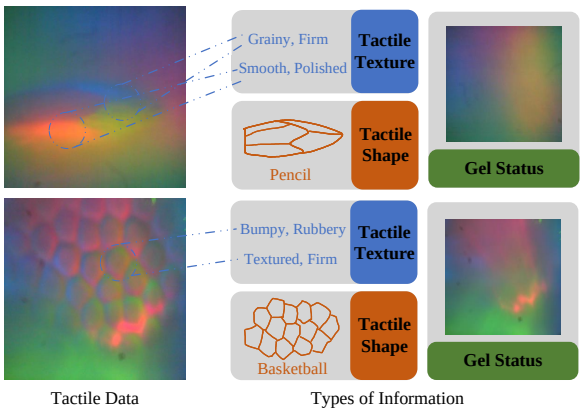

Figure 1: We present tactile images captured by sensors under different gel statuses. In our opinion, each tactile image contains three types of information: tactile texture, tactile shape, and gel status.

In recent years, as large-scale generative models (Ramesh et al. 2022) have demonstrated outstanding generative capabilities, some vision-conditioned methods attempt to utilize these models to generate tactile images by visual-to-tactile image translation (Yang, Zhang, and Owens 2023; Gao, Yuan, and Zhu 2023; Yang et al. 2024; Dou et al. 2024). Nevertheless, we believe these methods have two fatal shortcomings. **1)** Humans tend to describe tactile sensation using text rather than visual images. Consequently, text-conditioned generation methods can utilize more accurate information descriptions to produce data that align closely with tactile experiences (Obrist, Seah, and Subrama-

*Corresponding author.

nian 2013). **2)** Tactile images capture the deformation of objects on an elastomer gel (Johnson and Adelson 2009; Yuan, Dong, and Adelson 2017) which is embedded with cameras and lighting systems. This implies that setting the same object on tactile sensors under different gel conditions, including changes in lighting design, camera placement, and gel material, will produce distinctly different tactile images. To our knowledge, existing works (Yang, Zhang, and Owens 2023; Yang et al. 2024; Gao, Yuan, and Zhu 2023) on tactile generation have not considered the impact of different gels on the quality of the synthesized data.

As people usually describe tactile sensations in terms of the smoothness or softness, a straightforward approach is to use texture descriptions (Picard et al. 2003) to guide tactile image generation. However, we carefully analyze the characteristics of tactile images and divide them into two granularities: **1) Object-level**. Tactile texture (Hollins et al. 1993) and tactile shape (Johnson and Adelson 2009) are the two types of information related to the object in tactile images. Tactile texture is the primary attribute associated with tactile perception, such as smoothness and softness; tactile shape refers to the shape of the contact surface of the object and is manifested through changes in color and brightness. **2) Sensor-level**. Gel status includes the sensor information about the position of cameras, light sources, and gel material. In Fig. 1, we extract texture and shape information from tactile images, and present two different gel statuses, which are reflected in the tactile images collected without contact. The gel status affects the color variations of both tactile texture and shape, making this information crucial for the generation of tactile images.

In this paper, we introduce TextToucher, the first method specially designed for the text-to-touch generation task. We analyze the characteristics of tactile images in detail from two granularities: object-level (tactile texture, tactile shape), and sensor-level (gel status), which are modeled through text descriptions. For object-level conditions, we employ a multimodal large language model to build text sentences and design a question template tailored for the tactile collection situation to improve the accuracy of model responses. For sensor-level conditions, a set of learnable text prompts is defined to represent gel statuses. We further employ a pre-trained text model to encode text sentences and propose a time-adaptive strategy to fuse the tactile information. In the diffusion transformer architecture, various dual-grain text conditioning methods are explored to better control the tactile generation. Additionally, we introduce the Contrastive Text-Touch Pre-training (CTTP) metric, akin to CLIP (Radford et al. 2021), for evaluating the alignment between the generated tactile images and the text conditions. Extensive results demonstrate that through fine-grained textual conditions, TextToucher can effectively generate high-quality tactile images.

In summary, our contributions can be outlined as follows:

- We are the first to explore the text-to-touch generation task and demonstrate text-conditioned methods are more suitable for tactile generation than vision-conditioned methods.
- We conduct an in-depth analysis of tactile images, identifying two granularities of tactile images: object-level (tactile texture, tactile shape) and sensor-level (gel status). With fine-grained textual conditions, TextToucher can effectively synthesize high-quality tactile images.
- We introduce the CTTP metric, a new measure for evaluating the alignment between generated tactile images and textual descriptions.

Related Work

Tactile Sensor

Recent years have witnessed the development of different tactile sensors in many robotic applications, including sliding detection, texture recognition, object pushing, insertion, and tightening. Initial tactile sensors were designed to measure force, vibration and temperature by capturing simple, low-dimensional sensory signals. Lately, vision-based tactile sensors (e.g., GelSight (Yuan, Dong, and Adelson 2017; Johnson and Adelson 2009), GelTip (Gomes, Lin, and Luo 2020), TacTip (Ward-Cherrier et al. 2018), DIGIT (Lambeta et al. 2020)) have been proposed and utilize the deformation of an illuminated membrane to provide detailed information about shape and material properties. Compared to traditional single-point tactile sensors and tactile arrays, these vision-based tactile sensors offer higher-resolution tactile data. With the development of tactile sensors, a variety of tactile datasets are created by simulation-based (Gao et al. 2021, 2022), generation model-based (Yang, Zhang, and Owens 2023; Yang et al. 2024; Gao, Yuan, and Zhu 2023), human-collected methods (Fu et al. 2024; Kerr et al. 2022; Yang et al. 2022). Our approach belongs to the generation model-based methods and focuses on generating high-resolution tactile data with diffusion generative models.

Cross-Modal Synthesis with Generative Models

An emerging line of work has addressed the challenges of learning from cross-modal synthesis with generative models (Dong et al. 2024; Zhang et al. 2024; Sun et al. 2024). ImageBind (Girdhar et al. 2023) bridges multi-modality within a joint embedding space, it aligns the encoders of audio, video, depth, and text with the image encoder, thereby subtly integrating all five modalities and facilitating downstream cross-modal tasks, such as tactile image generation. Building on ImageBind, ImageBind-LLM (Han et al. 2023) incorporates a Large Language Model (LLM), enhancing the model's multi-modal understanding and reasoning capabilities. ImageBind-LLM employs a visual encoder to connect LLM with all other encoders; consequently, by leveraging ImageBind-LLM, the images can be generated using LLM or other modalities.

In the field of touch modality, Vision2Touch. (Li et al. 2019) introduces the VisGel dataset, which consists of tactile-visual paired images, and employs conditional GANs to achieve cross-modal image synthesis between GelSight tactile images and visual images. GVST (Yang et al. 2022) proposes the Touch and Go dataset, which encompasses multiple scenarios of visuo-tactile paired images, and utilizes diffusion models to accomplish cross-modal syn-

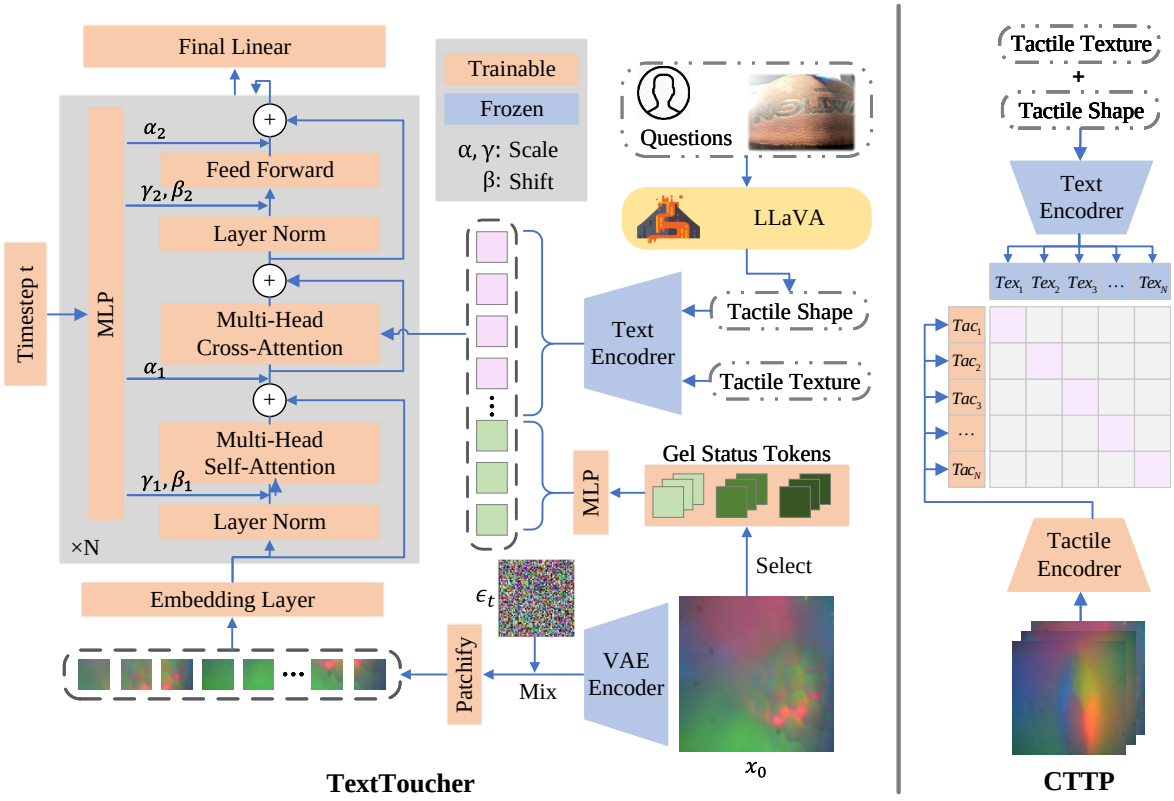

Figure 2: **Left:** Our proposed TextToucher utilizes text modality to obtain tactile texture, tactile shape and gel status information. We employ LLaVA, a vision-language large model, to caption the shape information in tactile images. Combining with texture descriptions from tactile datasets, we encode them with a text encoder. Additionally, we define a set of special word tokens to represent gel status information. **Right:** We train a tactile encoder using a contrastive loss function. In the shared space of text and tactile modalities, we propose a metric called CTTP, which uses cosine similarity to represent the relationship between tactile images and text descriptions. Our metric aims to effectively evaluate the quality of text-conditioned tactile image generation.

thesis tasks from tactile images to visual images. UniTouch (Yang et al. 2024) also integrates the tactile modality into ImageBind-LLM, enabling the model to generate tacile images from various tactile sensors such as GelSight, DIGIT. However, few works directly establish a bridge between text and tactile modalities, addressing the text-to-touch generation task.

Methodology

Texture, shape, and gel status are three significant attributes of tactile representations. The well-explored modality, text, with its rich semantics, is a wise choice for expressing the aforementioned attributes. In light of this, we aim to translate text to tactile images using a generative model. we utilize the Diffusion Transformer (DiT) (Peebles and Xie 2023) to conduct this task.

Preliminaries

Before introducing our method for tactile generation, we briefly review the fundamental of diffusion probabilistic models (Ho, Jain, and Abbeel 2020; Song et al. 2020b) (DPMs). Like most generative models (Kingma and Welling 2013; Goodfellow et al. 2020; Zhu et al. 2024), DPMs need to learn a mapping from a simple distribution, such as a Gaussian distribution, to the distribution of datasets. Given a data distribution $x_0 \sim q_{data}(x)$, we define that the noising process iteratively adds Gaussian noise $\epsilon_t \sim \mathcal{N}(0, I)$ to the sample data x_0 until x_T. This process can be described as follows:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}), \quad (1)$$

where β_t denotes the noise intensity at each timestep t. As highlighted by DDPM (Ho, Jain, and Abbeel 2020), the Eq. 1 can be simplified with $\alpha_t = 1-\beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$:

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \quad (2)$$

where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$. DPMs adopt a neural network to represent $p_\theta(x_{t-1}|x_t)$ and approximate the posterior $q(x_{t-1}|x_t, x_0)$ with it. Then the loss function can be written as follows:

$$\mathcal{L}_{\text{simple}} = E_{t,x_0,\epsilon_t}\left[||\epsilon - \epsilon_\theta(x_t, t)||^2\right]. \quad (3)$$

where $\epsilon_\theta(x_t, t)$ is the predicted noise by diffusion model at time t.

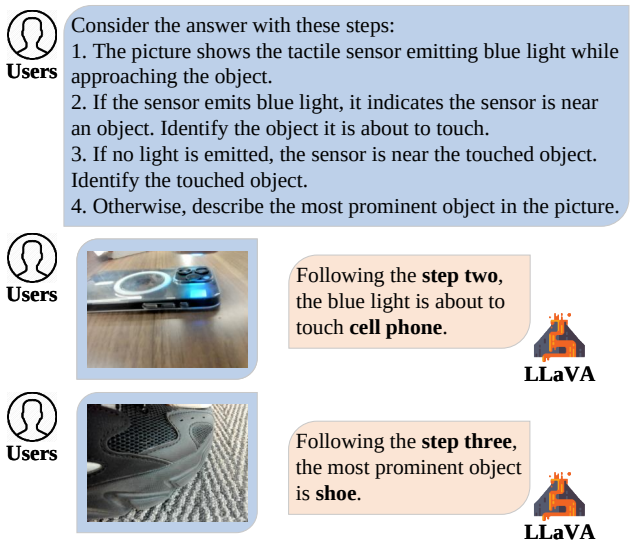

Figure 3: LLaVA engages in a step-by-step reasoning process based on carefully designed questions to achieve accurate data annotation.

To make image generation more controllable, conditional diffusion models incorporate additional inputs, such as textural descriptions (Rombach et al. 2022; Chen et al. 2023; Wang et al. 2024a; Tu et al. 2024; Wang et al. 2024b) and segmentation map (Zhang, Rao, and Agrawala 2023). Classifier-free guidance (Ho and Salimans 2022) aims to find a x that maximizes $\log p(c|x)$. With Bayes theorem, the model $\tilde{\epsilon}_\theta(x_t, t, c)$ can be modified as follows:

$$\tilde{\epsilon}_\theta(x_t, t, c) \propto s \cdot \epsilon_\theta(x_t, t, c) + (1 - s) \cdot \epsilon_\theta(x_t, t, \emptyset), \quad (4)$$

where s represents the scale of the guidance and $c = \emptyset$ indicates DPMs generate data with non-conditions. Since classifier-free guidance can significantly improve the quality of vision images, we also adopt the technique in tactile image generation.

Multimodal Large Language Model Annotation

This subsection discusses how to construct the object-level conditions corresponding to a tactile image. Existing tactile datasets (Fu et al. 2024; Kerr et al. 2022) already include tactile images along with texture descriptions and visual images. For simplicity, we derive the tactile shape based on these datasets.

As demonstrated in Fig. 1, we plan to use the text modality to describe the tactile shape. For example, a tactile image collected from the surface of a basketball is composed of interlocking pentagons; however, only a few tactile images can be precisely described the shape in words, as most shape in tactile images is irregular. TextToucher proposes to use the object that is contacted as a surrogate for tactile shape. Our motivation for this approach stems from the human ability to associate objects with their corresponding shape. When we mention a pencil, we can always imagine its shape. TextToucher aims to incorporate this ability into the generation of tactile images, allowing the model to learn the relationship between objects and their tactile shape.

Since manually labeling tactile images is time-consuming, we adopt LLaVA (Liu et al. 2024), a large language-vision model, to caption objects in the corresponding vision images automatically. As shown in Fig. 3, we immerse the model in the process of tactile data acquisition. By explicitly providing descriptions of the stages "about to touch", "in contact", and "special situations", we guide the model to complete reasoning step by step (Zhang et al. 2022), thereby enhancing LLaVA's understanding of the acquisition scenarios. Additionally, we check the annotation results of step four to ensure the accuracy of the model's annotations.

Gel Status Prompts

For sensor-level conditions, the gel status encompasses the placement of cameras and light sources, as well as the material properties of gels in vision-based tactile sensors. This status is evident in the tactile images captured by the sensor when it is not in contact with any object. The sensor-level conditions significantly influence the tactile images when the sensor is in contact with objects. Therefore, in addition to object-level conditions, we consider the gel status to be a critical aspect of tactile images. To represent different gel statuses, TextToucher employs special words. Specifically, we define a set of learnable prompts $c_i^{sen} = (c_i^1, c_i^2, \ldots, c_i^{n_{gs}})$ to denote the i-th gel status, where each status consists of n_{gs} tokens. These prompts capture the unique characteristics of each gel status, enabling the model to accurately reflect the influence of the gel properties on the generated tactile images.

Dual-Grain Text Conditioning Design

The information of tactile images is divided into two granularities: the object level and the sensor level. For object-level conditions, we specially design a text template P that includes both tactile texture p_t and tactile shape p_s: "the touch of $[p_s]$ is $[p_t]$". In the text-to-touch generation task, T5 language model (Raffel et al. 2020) is applied to translate tactile texture and shape conditions to embeddings $c^{obj} = (c_1, c_2, \ldots, c_l) \in \mathcal{R}^{l \times d_c}$. To refine condition fusion, we propose a time-adaptive condition method. Specifically, we first concatenate dual grains of text conditions $\tilde{c} = (c^{sen}, c^{obj})$ to generate a rough image containing gel information when the sampling timestep t is larger than a certain threshold θ_t. After that, object-level conditions are applied to focus on generating the tactile details.

We further design three different conditioning mechanisms and explore how to apply texture, shape and gel status to guide the generation of tactile images.

Text Modulation. Similar to AdaLN in DiT (Peebles and Xie 2023; Chen et al. 2023), we explore modulating the output of layer norm layers, self-attention layers and MLP layers. Rather than directly translate class embeddings to scale and shift parameters γ and β, we employ learnable MLP layers to fuse the sequence information in condition $\tilde{c}$ and add the result to timestep t.

Joint Attention. We concatenate the text tokens $\tilde{c}$ with the input tokens $x \in \mathcal{R}^{n \times d_c}$ of transformer blocks. The joint sequence $\tilde{x} = [c^{sen}, c^{obj}, x_1, x_2, \ldots, x_n]$ is then fed into the self-attention layer. To ensure the scalability of transformer blocks, the output discards tokens related to the text conditions before being fed into subsequent layers.

Cross Attention. We insert a multi-head cross attention layer after the multi-head self-attention layer within the transformer block. Each tactile token calculates the attention scores with text tokens that include texture, shape, and gel status. As shown in Fig. 2, we retain the modulation part of timestep embeddings t as employed in DiT, which is crucial for incorporating temporal information into the model.

Contrastive Text-Touch Pre-training (CTTP)

In the text-to-touch generation task, it is difficult to evaluate the correlation between a generated tactile image and its text description by visual inspection. Following other works in cross-modal synthesis (Yang, Zhang, and Owens 2023), we propose a new metric, Contrastive Text-Touch Pre-training (CTTP), to measure the alignment between the tactile images and text descriptions.

We refer to the instances from the same tactile-textual record $\{tac_i, tex_i\}$ as positive pairs, and instances from different tactile-textual record $\{tac_i, tex_j\}$ as negative pairs. Our goal is to minimize the embedding distance between positive sample pairs and maximize the embedding distance between negative sample pairs. Inspired by the CLIP (Radford et al. 2021) training method, we use InfoNCE (Oord, Li, and Vinyals 2018) to maximize the probability of positive sample pairs in each training batch B in Fig. 2:

$$\mathcal{L}_i^{tac,tex} = -\log \frac{exp(E_{tac}(tac_i) \cdot E_{tex}(tex_i)/\tau)}{\sum_{j=1}^{B} exp(E_{tac}(tac_i) \cdot E_{tex}(tex_j)/\tau)}, \tag{5}$$

where E_{tac} and E_{tex} are corresponding encoders, and τ is a temperature parameter. Similarly, we obtain the symmetric objective $\mathcal{L}_i^{tex,tac}$ and minimize it:

$$\mathcal{L} = \mathcal{L}^{tac,tex} + \mathcal{L}^{tex,tac}. \tag{6}$$

Once the tactile encoder E_{tac} and text encoder E_{tex} have been trained, we use Eq. 7 to evaluate the similarity between the tactile images and text prompts:

$$CTTP(tac_i, tex_i) = \frac{E_{tac}(tac_i) \cdot E_{tex}(tex_i)}{\|E_{tac}(tac_i)\|_2 \cdot \|E_{tex}(tex_i)\|_2}. \tag{7}$$

Experiment

Experiment Settings

Datasets We conduct experiments on two representative datasets. HCT (Fu et al. 2024) comprises visual-tactile data collected using a handheld 3D-printed data collection device. This dataset includes 43741 pairs of in-contact frames with tactile texture descriptions. Each data record contains the process of approaching, touching, sliding, and withdrawing from an object using the tactile sensor. Another dataset, SSVTP (Kerr et al. 2022), utilizes a UR5 robotic arm equipped with an RGB camera and a tactile sensor to collect data from various deformable surface environments, such as clothing seams, buttons, and zippers.

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**

Table 1: Quantitative results on the tactile generation task are presented. Our method can achieve the best results on each metric.

Metrics We adopt CTTP, LPIPS, SSIM and PSNR metrics to quantitatively evaluate the generated results. CTTP is employed to assess how well the tactile images align with the tactile descriptions, ensuring that the generated images accurately reflect the tactile sensations. LPIPS (Zhang et al. 2018) evaluates feature-level similarity between generated and real samples. Following GVSR (Yang, Zhang, and Owens 2023), we use SSIM and PSNR to evaluate the consistency at the pixel level.

Main Results

We categorize the comparison methods into two groups: 1) Vision-conditioned methods, where GVST (Yang, Zhang, and Owens 2023) and UniTouch (Yang et al. 2024) utilize diffusion models with a Unet architecture to generate tactile images from visual scenes. 2) Text-conditioned methods, where PixArt-α (Chen et al. 2023) translates text descriptions into visual images. We employ PixArt-α to complete the text-to-touch generation task, as both tasks involve processing text-based conditions.

Quantitative Evaluation. The quantitative results are presented in Tab. 1. TextToucher consistently achieves superior performance across all metrics on both HCT and SSVTP datasets. It is observed that TextToucher significantly outperforms GVST and UniTouch, confirming our hypothesis that the text modality can more accurately describe tactile sensations compared to the vision modality. Additionally, our method distinctly surpasses PixArt-α, demonstrating a +0.063 improvement in CTTP and a +0.077 enhancement in LPIPS on the HCT dataset. It also performs comparably on the SSVTP dataset, suggesting that traditional text-to-image methods are inadequate for tactile generation tasks.

Qualitative Evaluation. Fig. 4 shows the qualitative comparisons with alternative methods. Provided with text descriptions, our method can generate results that are more consistent with the reference tactile images. In contrast to vision-conditioned methods, TextToucher can effectively captures the contours of the contact objects through the tactile shape conditions. Furthermore, we observe that our method produces fewer artifacts in the generated images compared to PixArt-α. This improvement can be attributed to the dual-grain text conditions, which accurately model the intrinsic characteristics of tactile images, thereby enhancing the overall fidelity of tactile image synthesis.

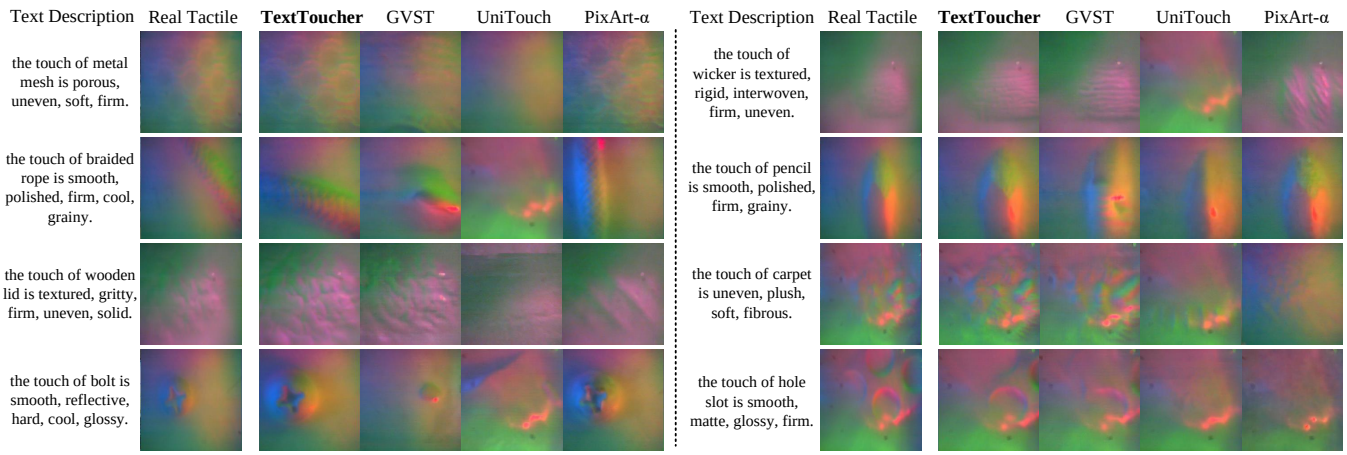

Figure 4: We compare our approach with other representative methods. TextToucher can produce tactile images with fewer artifacts and higher quality, closely aligning with the provided text descriptions.

Method	Text Conditions			CTTP ↑	LPIPS ↓	SSIM ↑	PSNR ↑
	Texture	Shape	Gel				
TC-T	✓			0.198	0.504	0.876	20.26
TC-TS	✓	✓		0.236	0.445	0.887	22.55
TextToucher	✓	✓	✓	**0.261**	**0.427**	**0.904**	**22.70**

Table 2: Comparison of different text condition types.

Conditioning Mechanism	CTTP ↑	LPIPS ↓	SSIM ↑	PSNR ↑
Text Modulation	0.151	0.476	0.899	21.91
Joint Attention	0.213	0.453	0.898	22.67
Cross Attention	**0.261**	**0.427**	**0.904**	**22.70**

Table 3: Comparison of different conditioning mechanisms.

Ablation Studies

Text Condition Types. In Tab. 2, we explore various combinations of text condition types in the text-to-touch generation task. Compared to employing only the tactile texture condition, generating tactile images with additional descriptions of tactile shapes and gel statuses contributes to more accurately forming the contours of contact objects and presenting them on specific gels. This approach achieves significant improvements across four metrics, underscoring the effectiveness of fine-grained text conditions in enhancing the quality of tactile image generation.

Conditioning mechanisms. In Tab. 3, we observe the Cross Attention conditioning mechanism is more effective for the text-to-touch generation task. In the Text Modulation approach, compressing sequential text tokens into inputs for modulation leads to a substantial loss of information. Besides, the imbalance in the number of tactile (1024 tokens) and text tokens (17.3 tokens on average) in the Joint Attention approach hampers the model's ability to establish connections between the two modalities.

Gel status prompts in different layers. We investigate the effects of applying gel status prompts to different layers of the model via two designs: gradually adding them from shallow to deep and from deep to shallow. The results in Tab. 4 show that employing gel status prompts within the first 14 layers enhances the quality of generated tactile images, and the best results are achieved when it is applied up to 28 layers. Adding gel status prompts from deeper to shallower layers can degrade the model's performance. This indicates that gel state encoding is crucial for shallow layer image representation and the optimal setting is adopted as the final setting for our approach.

Layer	CTTP ↑	LPIPS ↓	SSIM ↑	PSNR ↑
1-14	0.258	0.431	0.900	22.14
1-28	**0.261**	**0.427**	**0.904**	**22.70**
14-28	0.241	0.460	0.893	21.74
7-28	0.249	0.458	0.902	21.84

Table 4: Comparison of gel status prompts inserted in different layers.

n_{gs}	CTTP ↑	LPIPS ↓	SSIM ↑	PSNR ↑
1	0.248	0.452	0.900	21.74
2	0.255	0.452	0.895	21.93
4	**0.261**	**0.427**	**0.904**	**22.70**
6	0.250	0.458	0.899	21.56
8	0.252	0.453	0.899	21.60

Table 5: Comparisons of token counts representing the gel status prompts.

Token number n_{gs}. We study the impact of varying the number of tokens n_{gs} in gel status embeddings. As shown in Tab. 5, we find that using $n_{gs} = 4$ tokens effectively represents different gel status and surpasses other settings across various metrics by a margin. We hypothesize that fewer tokens ($n_{gs} = 1, 2$) are insufficient to capture the gel information, while a larger number of tokens ($n_{gs} = 6, 8, 10$) introduces redundant information, preventing the model from allocating weights in the cross attention layers.

θ_t	CTTP ↑	LPIPS ↓	SSIM ↑	PSNR ↑
800	0.258	**0.427**	0.903	22.69
600	**0.261**	**0.427**	**0.904**	**22.70**
400	0.247	0.434	0.897	22.54
200	0.235	0.435	0.892	22.26
0	0.234	0.434	0.899	22.12

Table 6: The Effect of timestep threshold θ_t employing different tactile text conditions.

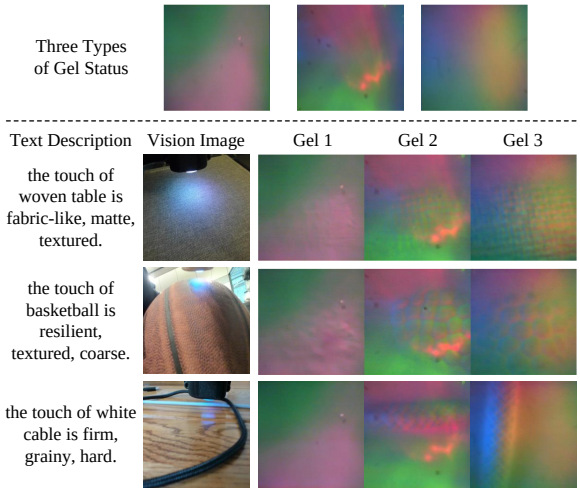

Figure 5: The first row displays the gel statuses contained in HCT dataset. We generate the same object under different gel statuses in the remaining rows.

Timestep threshold θ_t. We study the effect of varying tactile text conditions at different sampling timesteps for tactile generation. We set $\tilde{c} = \{c^{sen}, c^{obj}\}$ (three types of tactile information) for early timesteps before θ_t, and use tactile texture and shape description c afterwards. The results in Tab. 6 show that this approach with time-varying tactile text conditions improves model performance, achieving the best results at θ_t of 600 or 800 timesteps. This improvement occurs because using $\tilde{c}$ in the early sampling steps helps form an initial tactile image with gel information. Subsequently, the diffusion model with the conditional encoding c is able to refine the generation of tactile texture and shape.

Variation in Gel Statuses

In this section, we utilize different gel states to control the tactile generation conditioned on the same text sentences, enhancing the diversity of tactile images. We use three sets of prompts to represent the three gel statuses in the HCT dataset. Specifically, we use the gel state prompts to control the generation of tactile images in inference stage. The results in Fig. 5 indicate that the gel status conditions effectively control the presentation of tactile shape and texture. However, we also observe that the quality of images generated for gel state 1 is poor. We analyze the dataset and find that only 2.4% of the tactile images are collected using gel state 1. This limited sample size likely makes it difficult for the model to understand the relationship between gel state 1 and the corresponding texture and shape. We also conduct more tasks related to various gels in the Appendix.

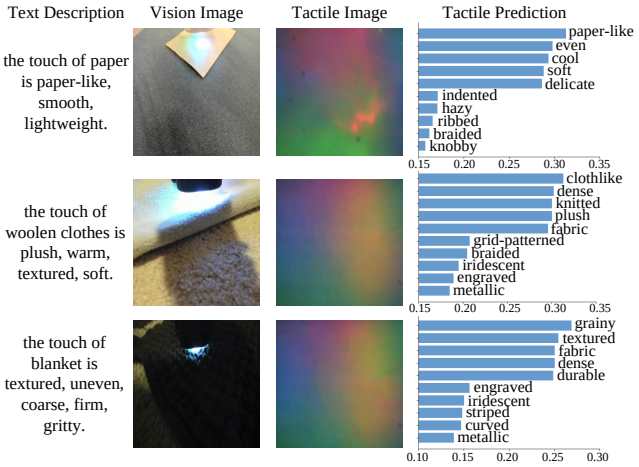

Figure 6: With the pre-trained tactile encoder, we extract features from tactile images and calculate the CTTP with all texture descriptions, predicting the top five most likely tactile texture descriptions and the most irrelevant ones.

Tactile Prediction

Similar to how CLIP is used for image classification, we employ the trained tactile image encoder for tactile predictions to verify the validity of CTTP metric. We extract all adjectives describing tactile texture from the HCT datasets. Using the tactile image encoder alongside the CLIP text encoder, we can compute the CTTP metric between the tactile images and all tactile descriptions $P_i = (p_s, p_{t_i})$. In Fig. 6, we present the prediction results, including the top five highest-scoring tactile texture descriptions and the most irrelevant descriptions. For example, in the second row, terms like "clothlike", "dense", and "knitted" accurately describe woolen clothes, whereas "metallic" does not relate to clothes at all. The experimental results demonstrate that the tactile image encoder trained with contrastive learning effectively aligns with the text encoder.

Conclusion

In this paper, we are the first to propose the text-to-touch generation task and specifically analyze tactile images at two levels: object-level (tactile texture, tactile shape), and sensor-level (gel status). Our proposed TextToucher leverages tactile text descriptions to generate high-quality tactile images. We extend the tactile datasets with texture descriptions and employ LLaVA to label shape information. For gel status, we define learnable prompts to present different gel statuses, which can be selectively added to the text conditions based on sampling timestep. Three text conditioning mechanisms are explored to guide image generation. Additionally, we introduce the CTTP metric to assess the alignment between tactile images and their corresponding text conditions. Experimental results and the extensive visualizations demonstrate that our method outperforms other methods and effectively generates high-quality tactile images.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China under Grant 62402430, 62206248 ,62476238, and Aeronautical Science Foundation of China 20240048076001.

References

Ackerman, J. M.; Nocera, C. C.; and Bargh, J. A. 2010. Incidental haptic sensations influence social judgments and decisions. *Science*, 328(5986): 1712–1715.

Barreiros, J. A.; Xu, A.; Pugach, S.; Iyengar, N.; Troxell, G.; Cornwell, A.; Hong, S.; Selman, B.; and Shepherd, R. F. 2022. Haptic perception using optoelectronic robotic flesh for embodied artificially intelligent agents. *Science Robotics*, 7(67): eabi6745.

Calandra, R.; Owens, A.; Jayaraman, D.; Lin, J.; Yuan, W.; Malik, J.; Adelson, E. H.; and Levine, S. 2018. More than a feeling: Learning to grasp and regrasp using vision and touch. *IEEE Robotics and Automation Letters*, 3(4): 3300–3307.

Chen, J.; Yu, J.; Ge, C.; Yao, L.; Xie, E.; Wu, Y.; Wang, Z.; Kwok, J.; Luo, P.; Lu, H.; et al. 2023. PixArt-α: Fast Training of Diffusion Transformer for Photorealistic Text-to-Image Synthesis. *arXiv preprint arXiv:2310.00426*.

Dong, J.; Liang, W.; Li, H.; Zhang, D.; Cao, M.; Ding, H.; Khan, S.; and Khan, F. S. 2024. How to Continually Adapt Text-to-Image Diffusion Models for Flexible Customization? *NeurIPS 2024*.

Dou, Y.; Yang, F.; Liu, Y.; Loquercio, A.; and Owens, A. 2024. Tactile-Augmented Radiance Fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Fu, L.; Datta, G.; Huang, H.; Panitch, W. C.-H.; Drake, J.; Ortiz, J.; Mukadam, M.; Lambeta, M.; Calandra, R.; and Goldberg, K. 2024. A Touch, Vision, and Language Dataset for Multimodal Alignment. *arXiv preprint arXiv:2402.13232*.

Gao, R.; Chang, Y.-Y.; Mall, S.; Fei-Fei, L.; and Wu, J. 2021. Objectfolder: A dataset of objects with implicit visual, auditory, and tactile representations. *arXiv preprint arXiv:2109.07991*.

Gao, R.; Si, Z.; Chang, Y.-Y.; Clarke, S.; Bohg, J.; Fei-Fei, L.; Yuan, W.; and Wu, J. 2022. Objectfolder 2.0: A multisensory object dataset for sim2real transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10598–10608.

Gao, R.; Yuan, W.; and Zhu, J.-Y. 2023. Controllable visual-tactile synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7040–7052.

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

Gomes, D. F.; Lin, Z.; and Luo, S. 2020. GelTip: A finger-shaped optical tactile sensor for robotic manipulation. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 9903–9909. IEEE.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.

Han, J.; Zhang, R.; Shao, W.; Gao, P.; Xu, P.; Xiao, H.; Zhang, K.; Liu, C.; Wen, S.; Guo, Z.; et al. 2023. Imagebind-llm: Multi-modality instruction tuning. *arXiv preprint arXiv:2309.03905*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Ho, J.; and Salimans, T. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Hollins, M.; Faldowski, R.; Rao, S.; and Young, F. 1993. Individual differences in perceptual space for tactile textures: evidence from multidimensional scaling analysis. *Perception and Psychophysics*, 54(6): 697–705.

Johnson, M. K.; and Adelson, E. H. 2009. Retrographic sensing for the measurement of surface texture and shape. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 1070–1077. IEEE.

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

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Lambeta, M.; Chou, P.-W.; Tian, S.; Yang, B.; Maloon, B.; Most, V. R.; Stroud, D.; Santos, R.; Byagowi, A.; Kammerer, G.; et al. 2020. Digit: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation. *IEEE Robotics and Automation Letters*, 5(3): 3838–3845.

Li, S.; Rodriguez, S.; Dou, Y.; Owens, A.; and Fazeli, N. 2024. Tactile Functasets: Neural Implicit Representations of Tactile Datasets. *arXiv preprint arXiv:2409.14592*.

Li, Y.; Zhu, J.-Y.; Tedrake, R.; and Torralba, A. 2019. Connecting touch and vision via cross-modal prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10609–10618.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Murali, A.; Li, Y.; Gandhi, D.; and Gupta, A. 2018. Learning to grasp without seeing. In *International Symposium on Experimental Robotics*, 375–386. Springer.

Obrist, M.; Seah, S. A.; and Subramanian, S. 2013. Talking about tactile experiences. In *Proceedings of the SIGCHI conference on human factors in computing systems*, 1659–1668.

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

Owens, A.; Isola, P.; McDermott, J.; Torralba, A.; Adelson, E. H.; and Freeman, W. T. 2016. Visually indicated sounds. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2405–2413.

Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4195–4205.

Picard, D.; Dacremont, C.; Valentin, D.; and Giboreau, A. 2003. Perceptual dimensions of tactile textures. *Acta psychologica*, 114(2): 165–184.

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

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140): 1–67.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. arXiv 2022. *arXiv preprint arXiv:2204.06125*.

Rodriguez, S.; Dou, Y.; Bogert, W. v. d.; Oller, M.; So, K.; Owens, A.; and Fazeli, N. 2024a. Contrastive Touch-to-Touch Pretraining. *arXiv preprint arXiv:2410.11834*.

Rodriguez, S.; Dou, Y.; Oller, M.; Owens, A.; and Fazeli, N. 2024b. Touch2touch: Cross-modal tactile generation for object manipulation. *arXiv preprint arXiv:2409.08269*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Song, S.; Zeng, A.; Lee, J.; and Funkhouser, T. 2020a. Grasping in the wild: Learning 6dof closed-loop grasping from low-cost demonstrations. *IEEE Robotics and Automation Letters*, 5(3): 4978–4985.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2020b. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

Sun, G.; Liang, W.; Dong, J.; Li, J.; Ding, Z.; and Cong, Y. 2024. Create your world: Lifelong text-to-image diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Sundaram, S.; Kellnhofer, P.; Li, Y.; Zhu, J.-Y.; Torralba, A.; and Matusik, W. 2019. Learning the signatures of the human grasp using a scalable tactile glove. *Nature*, 569(7758): 698–702.

Tu, J.; Ji, W.; Zhao, H.; Zhang, C.; Zimmermann, R.; and Qian, H. 2024. Driveditfit: Fine-tuning diffusion transformers for autonomous driving. *arXiv preprint arXiv:2407.15661*.

Wang, B.; Yang, F.; Yu, X.; Zhang, C.; and Zhao, H. 2024a. APISR: Anime Production Inspired Real-World Anime Super-Resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 25574–25584.

Wang, F.; Yin, H.; Dong, Y.-J.; Zhu, H.; Zhang, C.; Zhao, H.; Qian, H.; and Li, C. 2024b. BELM: Bidirectional Explicit Linear Multi-step Sampler for Exact Inversion in Diffusion Models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Wang, Y. R.; Duan, J.; Fox, D.; and Srinivasa, S. 2023. NEWTON: Are Large Language Models Capable of Physical Reasoning? *arXiv preprint arXiv:2310.07018*.

Ward-Cherrier, B.; Pestell, N.; Cramphorn, L.; Winstone, B.; Giannaccini, M. E.; Rossiter, J.; and Lepora, N. F. 2018. The tactip family: Soft optical tactile sensors with 3d-printed biomimetic morphologies. *Soft robotics*, 5(2): 216–227.

Wu, T.; Dong, Y.; Liu, X.; Han, X.; Xiao, Y.; Wei, J.; Wan, F.; and Song, C. 2024. Vision-based tactile intelligence with soft robotic metamaterial. *Materials & Design*, 238: 112629.

Yang, F.; Feng, C.; Chen, Z.; Park, H.; Wang, D.; Dou, Y.; Zeng, Z.; Chen, X.; Gangopadhyay, R.; Owens, A.; et al. 2024. Binding touch to everything: Learning unified multimodal tactile representations. *arXiv preprint arXiv:2401.18084*.

Yang, F.; Ma, C.; Zhang, J.; Zhu, J.; Yuan, W.; and Owens, A. 2022. Touch and go: Learning from human-collected vision and touch. *arXiv preprint arXiv:2211.12498*.

Yang, F.; Zhang, J.; and Owens, A. 2023. Generating visual scenes from touch. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22070–22080.

Yu, S.; Lin, K.; Xiao, A.; Duan, J.; and Soh, H. 2024. Octopi: Object Property Reasoning with Large Tactile-Language Models. *arXiv preprint arXiv:2405.02794*.

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

Zambelli, M.; Aytar, Y.; Visin, F.; Zhou, Y.; and Hadsell, R. 2021. Learning rich touch representations through cross-modal self-supervision. In *Conference on Robot Learning*, 1415–1425. PMLR.

Zhang, D.; Yu, Y.; Dong, J.; Li, C.; Su, D.; Chu, C.; and Yu, D. 2024. Mm-llms: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601*.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3836–3847.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhang, Z.; Zhang, A.; Li, M.; and Smola, A. 2022. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*.

Zhu, H.; Wang, F.; Ding, T.; Qu, Q.; and Zhu, Z. 2024. Analyzing and Improving Model Collapse in Rectified Flow Models. *arXiv preprint arXiv:2412.08175*.