UniTac: A Unified Multimodal Model for Cross-Sensor Tactile Understanding and Generation

Jiahang Tu*[1], Fengyu Yang*[2,6], Chenyang Ma[3], Xihang Yu[4], Ziyao Zeng[2], Shaokai Wu[5], Hanbin Zhao†[1], Zhi Tao[6], Chao Zhang[1], Hui Qian[1], and Alex Wong[2]

[1] Zhejiang University
[2] Yale University
[3] University of Oxford
[4] MIT
[5] Shanghai Jiaotong University
[6] UNIX AI

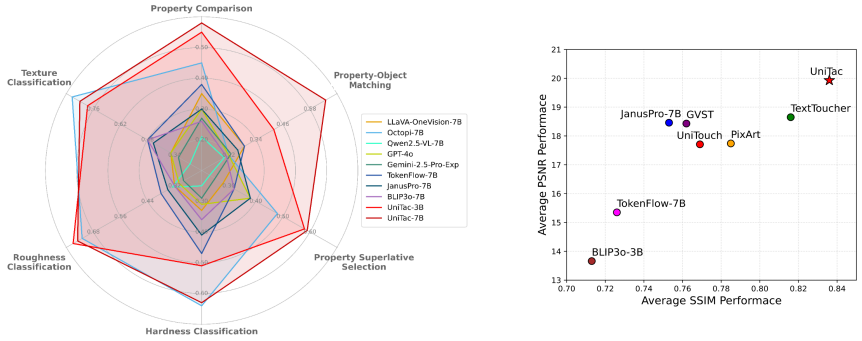
Abstract. Unified multimodal models (UMMs) have shown great promise in integrating understanding and generation across diverse modalities. However, existing research rarely extends this paradigm to the tactile domain, where both object-level semantics and sensor-level configurations jointly determine the meaning of touch. To address this gap, we propose UniTac, the first UMM designed for tactile understanding and generation. UniTac models the tactile process as a transition from non-contact to contact, capturing the physical interaction between sensors and objects through a dual-level representation that encodes both sensor and object attributes. For tactile understanding, UniTac introduces two tasks, object property description and sensor identification, to enhance reasoning over physical and cross-sensor information. For tactile generation, we design a two-stage training paradigm consisting of reconstruction and alignment, together with a sensor-prior-based sampling strategy that simulates realistic tactile contact. Trained on large-scale multi-sensor datasets, UniTac achieves state-of-the-art performance in tactile understanding and generates realistic tactile signals across sensors.

Keywords: Tactile Understanding · Tactile Generation · Unified Multimodal Model

1 Introduction

Unified multimodal models (UMMs) have recently gained significant attention for integrating perception and generation within a single architecture [22, 31, 49]. By unifying diverse modalities, UMMs enhance the adaptability and scalability of multimodal systems, laying the foundation for developing interactive

*: Equal contribution. †: Corresponding author.



(a) Average Performance on PHYSICLEAR-Test Benchmark.

(b) Average Performance on Tactile Generation Task.

Fig. 1: Quantitative evaluation of UniTac on tactile understanding and generation tasks. (a) Average results on the PHYSICLEAR-Test benchmark across six tactile understanding tasks, where UniTac-7B achieves the strongest overall performance. (b) Average SSIM and PSNR on the tactile generation task, showing that UniTac provides superior generation quality compared with existing baselines.

and physically grounded intelligent agents [5, 28]. Within this framework, tactile understanding and generation are essential for embodied intelligence and robotic perception, as touch provides a direct means of interaction with the physical world. Accurate tactile understanding enables fine-grained manipulation [8, 47], material and surface recognition [2, 17], and geometric inference beyond vision [33]. High-fidelity tactile generation facilitates simulation-based training [26], strengthens cross-modal learning [35, 44], and enables realistic tactile feedback in virtual environments [10, 14, 32]. Together, these capabilities strengthen multimodal perception and sensor-level grounding in embodied robotic systems.

Recent advances in visuo-tactile sensing have opened new directions for touch-centered multimodal research. Sensors such as GelSight [46], DIGIT [20], and Duragel [48] capture tactile data in image-like formats encoding both *sensor-level configurations* [35] (lighting, gel deformation, camera parameters) and *object-level semantics* [18, 20, 46] (surface geometry, hardness, roughness). However, two major challenges remain for tactile understanding. First, existing touch-language models are often trained on small, self-curated datasets, such as PHYSICLEAR [45] with only 482 touch videos. In contrast, the tactile research community has collectively released large-scale open datasets comprising over 400K video clips and 1.6 million frames, yet these datasets are typically used in isolation rather than being combined for joint training, which limits the representation of tactile semantics. Second, significant domain gaps among sensors hinder cross-sensor generalization due to differences in sensor characteristics and illumination conditions [35]. For tactile generation, prior work [30] has explored translation between specific sensor pairs but has not addressed generalizable cross-sensor synthesis across diverse sensor types. Moreover, most tactile stud-

ies treat understanding and generation as separate problems without exploring their intrinsic connection.

Unlike visual cameras that capture a natural image in a single exposure, tactile data acquisition inherently involves two stages: a *non-contact* stage capturing sensor configuration and a *contact* stage recording object-level physical properties under that configuration [18,35,46]. Without object-level information, non-contact tactile data lack semantic meaning; without sensor-level information, generative models are unaware of the sensor configuration on which tactile signals should be synthesized, while understanding models cannot interpret how sensor design translates tactile patterns into object properties. For example, GelSight [46] encodes surface roughness through marker displacement patterns that depend on its sensor configuration. Thus, a unified multimodal model should jointly model both levels of tactile information to achieve cross-sensor tactile understanding and generation.

In this work, we present UniTac, the first UMM for the touch domain, designed to jointly perform tactile understanding and generation tasks within a single framework. As illustrated in Fig. 2, UniTac unifies tactile understanding (*e.g.*, property comparison/description) and tactile generation under a shared multimodal backbone. UniTac is trained on a large composite tactile corpus that integrates diverse open datasets comprising over 400K video clips and 1.6 million frames from multiple sensor domains. The proposed framework comprises several essential modules for tactile data processing. The touch encoder extracts meaningful representations from tactile inputs, while the multimodal large language model (MLLM) serves as a shared backbone for both understanding and generation. The sensor-aware DiT (Diffusion Transformer) projector further maps the MLLM outputs into the conditional space required by the touch decoder, which synthesizes accurate tactile data. For understanding, UniTac introduces two supervised tasks that enhance comprehension of tactile information: object property description [41, 45], which captures the physical attributes of contacted objects, and sensor identification, which enables recognition of specific sensor configurations. These tasks jointly strengthen the model's understanding of both object-level and sensor-level information. For generation, the training is divided into two stages: reconstruction and alignment. In reconstruction stage, the decoder is trained using hidden tokens from the touch encoder as conditional inputs, enabling efficient parallel training without involv-

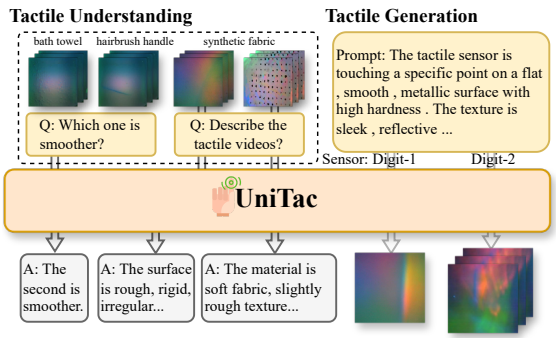


Fig. 2: Overview of UniTac for unified tactile understanding and tactile generation.

ing the MLLM backbone. The alignment stage then trains a DiT Projector to align the MLLM's output with encoder representations. To compensate for the absence of sensor-level cues in tactile descriptions, sensor-level tokens from the encoder are incorporated in the projector to refine the generated results. Moreover, we propose a sensor-prior-based sampling strategy that simulates the transition from non-contact to contact stages. By incorporating two level conditions into the sampling process, this method achieves flexible and accurate tactile signal generation aligned with real sensor behaviors. Extensive experiments on the PHYSICLEAR-Test benchmark [45] and multiple tactile datasets demonstrate strong performance of UniTac on both understanding and generation (Fig. 1), compared with existing UMMs and leading tactile models. We further deploy UniTac on real robotic platforms to verify its understanding ability and the practical effectiveness of generated cross-sensor data.

Our main contributions are summarized as follows:

- We introduce UniTac, the first UMM for the tactile domain, jointly modeling sensor-level configurations and object-level semantics for unified understanding and generation.
- We enhance tactile understanding through large-scale multi-sensor training and dual-level supervision tasks that improve object physical reasoning.
- We advance tactile generation via a two-stage training scheme and a sensor-prior-based sampling strategy that models the non-contact–to–contact transition for realistic tactile synthesis.
- We validate UniTac via extensive experiments and real-world deployment, achieving SOTA performance and demonstrating its practicality in robotic perception.

2 Related Work

2.1 Unified Multimodal Models

Recent advances in unified multimodal models (UMMs) have integrated perception and generation within a single architecture [22, 28, 31, 49]. Works such as Janus-Pro [28], Chameleon [34], and Emu3 [38] employ Transformer-based backbones to jointly handle visual understanding and synthesis. Hybrid frameworks including Transfusion [49] and Show-o [40] further combine autoregressive and diffusion paradigms to exploit their complementary strengths. BLIP3-o [3] extends this line by introducing a diffusion transformer for CLIP-space generation with sequential training that balances understanding and generation. However, existing UMMs remain limited to visual–textual modalities. In contrast, our UniTac extends this paradigm to the tactile domain, enabling both tactile understanding and generation within a unified architecture trained on a large-scale composite dataset.

2.2 Tactile Understanding Models

Tactile understanding has increasingly relied on representation learning to capture contact dynamics and material properties. UniTouch [42] and TVL-Link [7] align tactile inputs with visual and linguistic modalities to build consistent multimodal embeddings within a single sensor setup. AnyTouch [12] proposes a cross-sensor representation unifying tactile features but focuses mainly on representation learning rather than downstream reasoning. Beyond representations, Octopi [45] aligns tactile videos with vision–language models for object property reasoning, while VTV-LLM [41] explores visuo-tactile video understanding but relies on a small self-collected dataset, limiting generalization. More recent efforts such as SToLa [6] and CLTP [27] introduce commonsense reasoning and contrastive language–tactile pretraining for 3D contact understanding. Despite these advances, most models still overlook sensor-level configurations and fail to unify tactile understanding and generation within a single framework.

2.3 Tactile Generative Models

Tactile generation research mainly focuses on cross-modal synthesis between visual and tactile modalities. Vision2Touch [23] establishes paired visuo-tactile data for conditional generation, and GVST [43] adopts diffusion models for high-quality bidirectional synthesis. UniTouch [42] incorporates tactile signals into large multimodal frameworks for image-conditioned generation. Recent studies have emphasized finer control, such as TextToucher [35] for text-to-touch synthesis, ControlTac [25] for force- and position-controlled augmentation, and Touch2Touch [30] for translation between limited sensor pairs. Nonetheless, most existing methods remain confined to single-sensor scenarios and lack unified cross-sensor modeling. UniTac overcomes these limitations by enabling multi-sensor tactile generation through the joint modeling of sensor-level configurations and object-level semantics within a unified multimodal framework.

3 Method

UniTac unifies tactile understanding and generation across sensors by jointly modeling sensor-level configurations and object-level semantics within a single framework. This formulation enables the model to reason about how different sensors perceive the same physical properties and to synthesize sensor-consistent tactile signals. To this end, we introduce three key components bridging understanding and generation. In Sec. 3.1, Dual-Level Mixture Comprehension enhances tactile reasoning through object property description and sensor identification, improving cross-sensor comprehension. In Sec. 3.2, Two-Stage Aligned Generation advances tactile synthesis through reconstruction and sensor-aware alignment, ensuring semantic–physical coherence. In Sec. 3.3, Sensor-Prior Sampling Strategy models the transition from non-contact to contact by embedding sensor priors during sampling, yielding realistic cross-sensor tactile generation. An overview of the full framework is shown in Fig. 3.

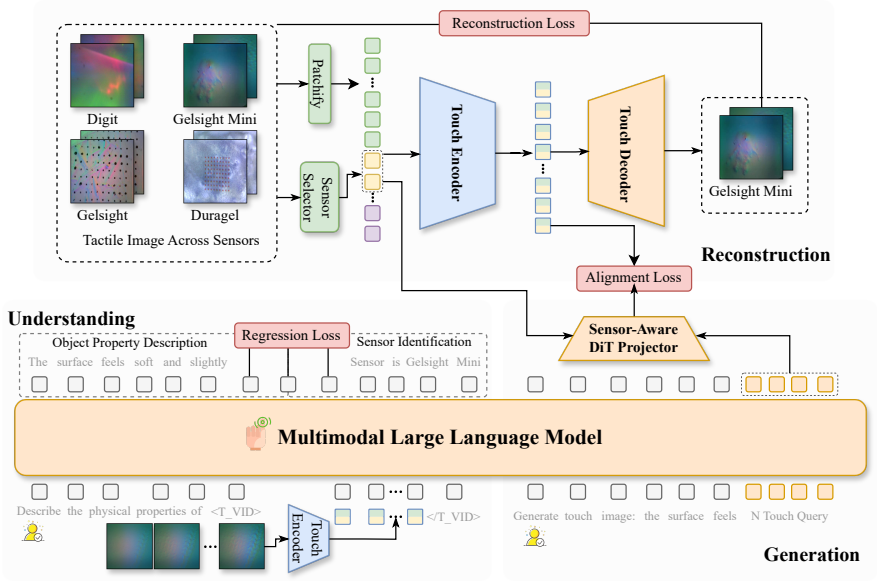


Fig. 3: Overview of the UniTac architecture. UniTac unifies tactile understanding and generation across sensors by jointly modeling sensor-level configurations and object-level semantics. The Touch Encoder extracts static and dynamic contact features, while the Multimodal Large Language Model (MLLM) integrates tactile and textual modalities for joint reasoning over object- and sensor-level information (Sec. 3.1). The Sensor-Aware DiT Projector and Touch Decoder perform two-stage tactile generation (Sec. 3.2), combining data reconstruction with sensor-aware alignment between MLLM embeddings and tactile representations. In addition, a sensor-prior-based sampling strategy (Sec. 3.3) models the transition from non-contact to contact, achieving realistic cross-sensor tactile synthesis.

3.1 Dual-Level Mixture Comprehension

To strengthen UniTac's reasoning in tactile understanding, we propose Dual-Level Mixture Comprehension, which jointly supervises the model from object-level and sensor-level perspectives using paired video–text data.

Given a visuo-tactile video V_i and an accompanying text T_i, the pretrained touch encoder produces a sequence of tactile tokens $\mathbf{Z}_i = E_{\text{touch}}(V_i) \in \mathbb{R}^{L_v \times d}$. We then splice these tokens into the MLLM text stream using two special markers `<T_VID>` and `</T_VID>`, yielding the input $X_i = [\texttt{<T\_VID>}, \mathbf{Z}_i, \texttt{</T\_VID>}, \Pi_i, T_i]$, where Π_i denotes an instructional prompt that specifies one of two comprehension objectives: *object property description* or *sensor identification*. The sequence X_i is fed into the MLLM and optimized via a standard next-token prediction objective:

$$\mathcal{L} = -\sum_{t=2+L_v+|\Pi_i|}^{|X_i|-1} \log p_\theta(x_{i,t+1} \mid x_{i,\le t}), \tag{1}$$

where each token in X_i is trained to predict its next token $x_{i,t+1}$.

Object-Level Supervision (Property Description). Following Octopi [45], we describe the contacted object using three tactile dimensions: roughness, hardness, and texture. For this task, Π_i^{prop} instructs the MLLM to generate the property description (*e.g.*, "The surface is soft, rough and big bumps."). By learning to predict these words conditioned on tactile embeddings, the model strengthens its grounding in object-level semantics, encouraging it to associate contact patterns with physical material properties.

Sensor-Level Supervision (Sensor Identification). In contrast, Π_i^{sen} guides the model to verbalize the tactile sensor identity (*e.g.*, "Captured by a Digit sensor."). The same next-token prediction loss applies, driving the MLLM to learn sensor-related variations such as lighting, gel elasticity, and imaging resolution. This supervision enforces sensitivity to sensor-level configurations, making the representation more robust across various tactile devices.

The overall comprehension loss integrates the two tasks:

$$\mathcal{L}_{\text{DLMC}} = \mathcal{L}_{\text{prop}} + \lambda_{\text{sen}}\, \mathcal{L}_{\text{sen}}, \tag{2}$$

where $\mathcal{L}_{\text{prop}}$ and $\mathcal{L}_{\text{sen}}$ denote the respective next-token prediction losses under their prompts, and $\lambda_{\text{sen}} > 0$ balance the two objectives. Through this dual-level supervision, UniTac learns to disentangle what changes with the *object* from what changes with the *sensor*, leading to more accurate tactile understanding.

3.2 Two-Stage Aligned Generation

To achieve controllable and sensor-consistent tactile generation, UniTac employs a Two-Stage Aligned Generation strategy consisting of a reconstruction stage and a sensor-aware alignment stage.

Stage I: Reconstruction. The reconstruction stage focuses on learning a generative prior within the tactile domain, independent of the MLLM. Here, the latent representation $\mathbf{Z}_i$ extracted from the touch encoder inherently contains two types of information: $\mathbf{Z}_i = [\mathbf{Z}_i^{\text{obj}}, \mathbf{Z}_i^{\text{sen}}]$, where $\mathbf{Z}_i^{\text{obj}}$ encodes *object-level semantics* such as hardness and roughness, and $\mathbf{Z}_i^{\text{sen}}$ represents *sensor-level configurations* including illumination and gel properties. These latent tokens are fed into the touch decoder D_{touch} to reconstruct the tactile signal. As this stage does not involve the MLLM backbone, it can be trained in parallel with the Dual-Level Mixture Comprehension task to improve training efficiency.

Stage II: Sensor-Aware Alignment. After learning tactile priors, we align the MLLM's textual generation space with the tactile latent space. Given a tactile description T_i and N touch queries, the MLLM produces output query embeddings: $\hat{\mathbf{Q}}_i = E_{\text{MLLM}}(T_i, \mathbf{Q}_i) \in \mathbb{R}^{N\times d}$, which primarily encode *object-level semantics* consistent with the described physical attributes of the surface, but lack explicit sensor cues. To recover sensor awareness, we concatenate these embeddings with the specific *sensor tokens* $\mathbf{S}$ obtained from the pretrained touch encoder, forming the conditional representation: $\mathbf{F}_i = [\hat{\mathbf{Q}}_i; \mathbf{S}]$.

The sensor-aware DiT projector learns a conditional vector field $v_\theta(\cdot|t, \mathbf{F}_i)$ that transforms a Gaussian prior toward the tactile latent representation $\mathbf{Z}_i$,

under the guidance of $\mathbf{F}_i$. Following the rectified flow formulation [11, 24], we define a linear interpolation path between a Gaussian noise sample $\mathbf{z} \sim \mathcal{N}(0, I)$ and the target tactile latent $\mathbf{Z}_i$ as:

$$\mathbf{x}_t = (1-t)\,\mathbf{z} + t\,\mathbf{Z}_i, \qquad t \in [0,1] \tag{3}$$

whose target velocity is a constant vector

$$\mathbf{u}_t = \frac{d\mathbf{x}_t}{dt} = \mathbf{Z}_i - \mathbf{z}. \tag{4}$$

The sensor-aware DiT projector is trained by minimizing the conditional rectified flow matching loss:

$$\mathcal{L}_{\text{align}}^{\text{RF}} = \mathbb{E}_{t\sim\mathcal{U}(0,1),\mathbf{z}\sim\mathcal{N}(0,I),\,\mathbf{Z}_i}\big\|v_\theta(\mathbf{x}_t|t,\mathbf{F}_i) - (\mathbf{Z}_i - \mathbf{z})\big\|_2^2. \tag{5}$$

This alignment bridges the MLLM semantic output and the touch encoder representations, enabling the model to generate tactile signals that are both semantically faithful and sensor-consistent.

3.3 Sensor-Prior Sampling Strategy

In real tactile sensing, data acquisition follows a sequential process from a non-contact stage to a contact stage. The non-contact stage captures the intrinsic sensor state, while the contact stage records the physical interaction between the sensor and the object surface. Inspired by this process, we propose a Sensor-Prior Sampling Strategy (SPSS) that incorporates sensor priors into the sampling procedure, thereby simulating the gradual transition from non-contact to contact during tactile generation.

The denoised velocity field is typically estimated using classifier-free guidance (CFG) [15]:

$$\hat{v}_\theta(\mathbf{x}_t, c) = v_\theta(\mathbf{x}_t|t,\varnothing) + s\big[v_\theta(\mathbf{x}_t|t,c) - v_\theta(\mathbf{x}_t|t,\varnothing)\big], \tag{6}$$

where c denotes the conditional input and s is the guidance scale. However, in tactile generation, the unconditional prior $v_\theta(\mathbf{x}_t|t,\varnothing)$ does not accurately represent the non-contact initialization, as tactile signals are inherently dependent on the sensor configuration.

To better align with the tactile data acquisition process, we replace the unconditional branch with a sensor-conditioned prior that explicitly encodes the non-contact state. The sampling process is redefined as:

$$\begin{aligned}\hat{v}_\theta(\mathbf{x}_t|t,\,\mathbf{Z}_i^{\text{obj}},\mathbf{Z}_i^{\text{sen}}) &= v_\theta(\mathbf{x}_t|t,\;\mathbf{Z}_i^{\text{sen}}) \\ &\quad + s\big[v_\theta(\mathbf{x}_t|t,\;\mathbf{Z}_i^{\text{obj}},\mathbf{Z}_i^{\text{sen}}) - v_\theta(\mathbf{x}_t|t,\;\mathbf{Z}_i^{\text{sen}})\big].\end{aligned} \tag{7}$$

Here, $v_\theta(\mathbf{x}_t|t,\mathbf{Z}_i^{\text{sen}})$ serves as the sensor prior branch, corresponding to the non-contact process, while $v_\theta(\mathbf{x}_t|t,\mathbf{Z}_i^{\text{obj}},\mathbf{Z}_i^{\text{sen}})$ introduces the contact-specific guidance that reflects object-level semantics under the same sensor state. This design enforces a physically consistent transition, where the first item respects the sensor configuration and the second item incorporates contact semantics.

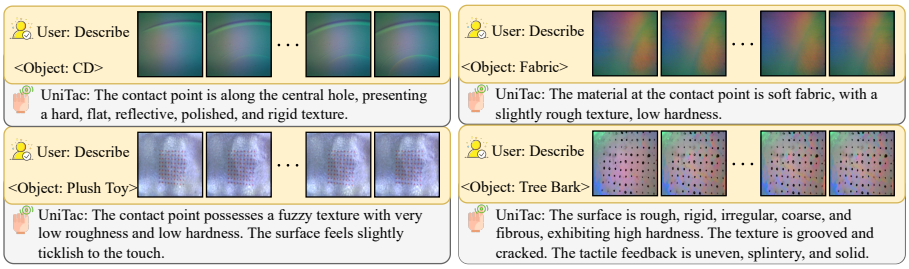


Fig. 4: Object property description of tactile videos across various tactile sensors. UniTac generates object-aware tactile descriptions that align with physical properties of the contacted materials.

4 Experiments

4.1 Implementation

We train UniTac on a large-scale visuo-tactile corpus integrated from five public datasets organized by AnyTouch [12]: Touch and Go [43], Tacquad [12], TVL [13], SSVTP [19], and PHYSICLEAR [45]. Except for SSVTP, all datasets include paired tactile videos and textual descriptions. We preprocess and unify them into a consistent format comprising approximately 400K video clips and 1.6M frames with aligned text annotations. We filter the contact frames with tactile deformations by calculating the difference between each tactile image and the corresponding background frame in these datasets. For tactile feature extraction, the pretrained touch encoder from AnyTouch is adopted to encode both tactile images and videos. The MLLM backbone is based on Qwen-VL 2.5 [37], and the touch decoder is implemented with SANA [39]. Additional implementation details are provided in Appendix A.

For tactile understanding, we evaluate simple perceptual recognition through hardness, roughness, and texture classification, and complex reasoning through Property Comparison (PC), Property-Object Matching (POM), and Property Superlative Selection (PSS) on the PHYSICLEAR-Test benchmark [45]. For tactile generation, following GVST [43] and TextToucher [35], we evaluate synthesis quality using SSIM and PSNR by randomly sampling 60K generated tactile images and comparing them with their corresponding ground-truth samples.

4.2 Tactile Understanding

To evaluate the tactile understanding capability of UniTac, we compare it with advanced multimodal models like Qwen2.5-VL [1] and GPT-4o [16]. We further include representative UMMs that unify understanding and generation. The quantitative results are summarized in Tab. 1. UniTac-7B achieves the highest overall score of 66.51, surpassing all previous models. Notably, the overall improvement mainly comes from UniTac's stronger tactile reasoning capability rather than uniform gains across all tasks. On the three reasoning-oriented

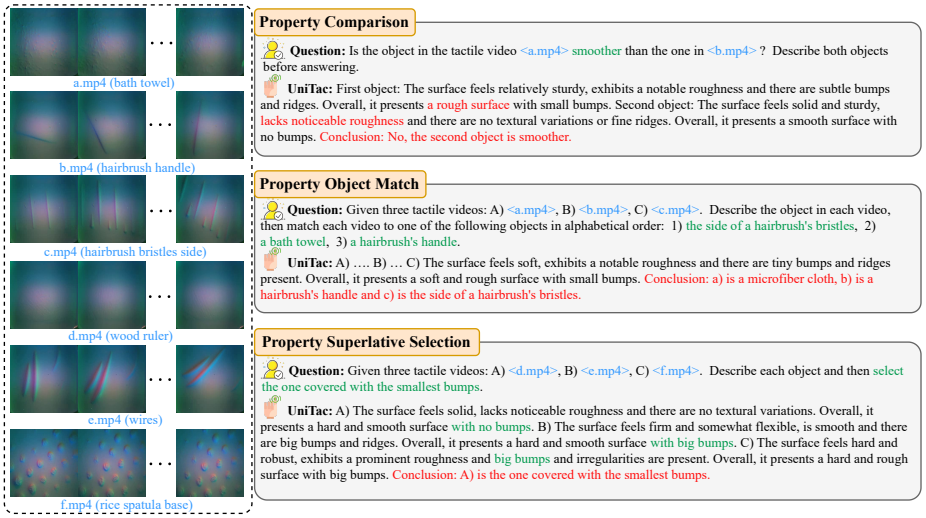


Fig. 5: Tactile video understanding across various understanding tasks. We evaluate UniTac on three representative understanding tasks: Property Comparison, Property–Object Matching, and Property Superlative Selection. UniTac first provides fine-grained descriptions of surface attributes (*e.g.*, roughness, bumps, hardness) and then performs reasoning to derive the final answer. The results show that UniTac can distinguish tactile differences through multi-step tactile reasoning capability.

tasks, i.e., PC, POM, and PSS, UniTac-7B improves over the strongest UMM baseline Octopi-7B by 11.80, 42.39, and 11.22 points, respectively. These results demonstrate that UniTac can better ground tactile observations into high-level physical and semantic concepts, leading to more reliable reasoning over material properties and tactile interactions.

In Fig. 4, we present qualitative results of the object property description (OPD) task, which is explicitly used during training to enhance tactile understanding. Across diverse tactile sensors, UniTac generates coherent and physically grounded descriptions that align with material properties and sensor configurations. For instance, it correctly characterizes the rigid, polished surface of a CD and distinguishes it from the soft, fuzzy texture of a plush toy, reflecting accurate reasoning over hardness and roughness cues. Building upon this capability, Fig. 5 further evaluates UniTac on complex tactile reasoning tasks. Rather than directly predicting answers, the model implicitly relies on fine-grained surface attributes, such as roughness, bump distribution, and stiffness, to support comparison, matching, and superlative selection.

4.3 Tactile Generation

Image Generation. We evaluate UniTac's tactile image generation on four representative sensors: Digit, GelSight, GelSight Mini, and Duragel. As shown in Tab. 2, UniTac achieves the highest average performance of 0.836 SSIM and

Table 1: Comparing the understanding capability of UniTac with other generative and unified multimodal models. UniTac-7B achieves the best average performance across all metrics, indicating stronger fine-grained tactile reasoning and understanding on PHYSICLEAR-Test benchmark. PC refers to Property Comparison task; POM refers to Property–Object Matching task; PSS refers to Property Superlative Selection task.

Type	Model	Method	**PC↑**	**POM↑**	**PSS↑**	**Hardness↑**	**Roughness↑**	**Texture↑**	**Average↑**
Und. Only	GPT-4o [16]	AR	30.87	22.62	38.05	31.37	30.48	36.51	31.65
	Qwen2.5-VL-7B [1]	AR	21.47	19.92	23.05	25.73	33.33	26.62	25.01
	Gemini-2.5-Pro-Exp [9]	AR	26.71	23.47	25.06	29.19	28.44	32.18	27.50
	LLaVA-OneVision-7B [21]	AR	35.42	29.11	28.44	33.33	32.18	36.51	32.49
	Octopi-7B [45]	AR	45.50	22.22	48.00	**64.10**	73.92	87.17	57.31
Und. and Gen.	TokenFlow-7B [29]	AR	38.05	29.48	32.56	47.68	38.70	48.31	39.13
	JanusPro-7B [5]	AR + Diff	30.18	26.71	38.05	41.63	35.82	45.37	36.29
	BLIP3o-3B [3]	AR+ Diff	26.71	21.47	32.18	36.51	32.18	48.00	32.84
	UniTac-3B(Ours)	AR + Diff	54.97	42.13	58.90	51.28	**76.92**	79.48	60.61
	UniTac-7B(Ours)	AR + Diff	**57.30**	**64.61**	**59.22**	61.53	74.35	**82.05**	**66.51**

Table 2: Comparing the generation capability of UniTac with other generative and unified multimodal models. Results are reported across four tactile sensors (Digit, GelSight, GelSight Mini, and Duragel). UniTac achieves the highest average SSIM and PSNR among unified multimodal models and remains competitive with generation-only methods, demonstrating strong cross-sensor tactile synthesis capability.

Type	Model	**Digit**		**Gelsight**		**Gelsight Mini**		**Duragel**		**Average**	
		SSIM↑	**PSNR↑**	**SSIM↑**	**PSNR↑**	**SSIM↑**	**PSNR↑**	**SSIM↑**	**PSNR↑**	**SSIM↑**	**PSNR↑**
Gen. Only	GVST [44]	0.841	18.72	0.655	14.58	0.883	18.90	0.429	14.98	0.762	18.43
	UniTouch [42]	0.859	18.17	0.636	14.18	0.851	18.63	0.440	14.31	0.769	17.71
	PixArt-α [4]	0.877	18.10	0.641	14.09	0.869	18.77	0.446	14.36	0.785	17.74
	TextToucher [35]	0.901	20.90	0.662	15.42	0.896	20.38	**0.473**	16.36	0.816	18.65
Und. and Gen.	TokenFlow-7B [29]	0.834	17.47	0.628	14.04	0.758	15.67	0.431	12.72	0.726	15.35
	JanusPro-7B [5]	0.854	20.18	0.653	15.34	0.894	19.67	0.464	17.73	0.753	18.46
	BLIP3o-7B [3]	0.821	16.30	0.618	13.46	0.746	14.74	0.405	10.92	0.713	13.66
	UniTac	**0.915**	**21.26**	**0.683**	**16.28**	**0.946**	**24.56**	0.472	**18.44**	**0.836**	**19.93**

19.93 PSNR, surpassing all state-of-the-art generative and unified models. We also observe that all models perform relatively lower on Duragel due to unstable acquisition conditions and significant variation in sensor calibration and gel state. The collected tactile images are often blurrier and less consistent in lighting and contact geometry, increasing intra-domain variance and making it harder for models to learn stable patterns between tactile appearance and physical properties, leading to overall lower metrics across all methods.

In Fig. 6, UniTac consistently generates realistic and physically coherent tactile images across different sensor types and configurations. It can also accurately reflect distinct sensor configuration even the same sensor type, such as the results in Digit-1 and Digit-2. Unlike other UMMs that misrepresent sensor states or lose spatial fidelity, UniTac preserves fine structural and illumination cues, demonstrating robust cross-sensor consistency enabled by its joint modeling of sensor- and object-level information.

Video Generation. Beyond static image synthesis, we extend UniTac to tactile video generation by replacing the touch decoder with Wan v2.2 [36], while

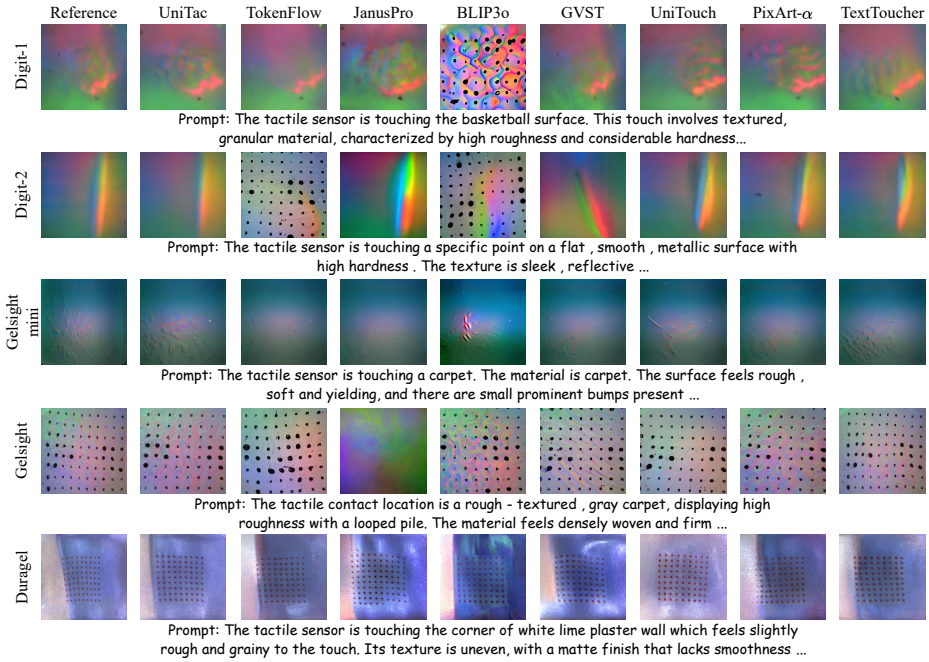


Fig. 6: Qualitative comparison of tactile image generation across various tactile sensors. UniTac consistently generates realistic and physically coherent tactile images across diverse sensors and configurations.

keeping the same sensor-aware conditioning mechanism. This design enables temporally coherent synthesis without altering the unified multimodal backbone or the sensor-specific conditioning strategy. As shown in Fig. 7, UniTac generates continuous tactile sequences across different sensors, capturing both the onset of contact and the gradual pressure propagation over time. The generated videos preserve fine-grained texture cues (*e.g.*, the granular surface of an orange) as well as sensor-specific deformation patterns and illumination dynamics. Besides, UniTac maintains temporal consistency in dot-grid displacement and contact region evolution. These results demonstrate that the proposed framework generalizes naturally from static tactile image synthesis to dynamic tactile video modeling, enabling realistic simulation of contact processes under diverse sensor configurations. More results are provided in Appendix C.

4.4 Ablation Studies

To examine the contribution of each component in UniTac, we conduct a series of ablation experiments, as summarized in Tab. 3. More ablation studies are provided in Appendix B. Removing the Sensor Identification objective causes a clear drop on PHYSICLEAR (from 60.61 to 57.38), indicating that explicitly modeling sensor-level configurations substantially enhances the UMM's ability to interpret tactile data across various tactile sensors. Without the Dual-Level

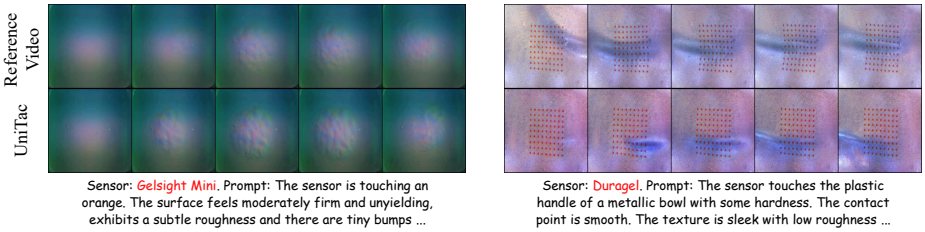


Fig. 7: Cross-sensor tactile video generation results. Left (GelSight Mini, orange): UniTac reproduces the fine, bumpy texture and progressive deformation consistent with the orange peel surface. Right (Duragel, bowl handle): UniTac generates smooth contact patterns with localized indentation and realistic dynamic changes throughout the contact process.

Table 3: Performance impact of different modules in UniTac through ablation study.

Component	PHYSICLEAR	SSIM	PSNR
UniTac	**60.61**	**0.836**	**19.93**
w/o Sensor Identification	57.38	0.822	19.91
w/o Dual-Level Comprehension	26.52	0.758	18.14
w/o DiT Projector	60.61	0.794	19.25
w/o Sensor-Prior Sampling	60.61	0.817	19.49

Comprehension, the performance decreases sharply across all metrics, as the MLLM degenerates into a standard Qwen-VL model lacking tactile understanding. This confirms that tactile comprehension not only improves reasoning performance but also provides critical semantic priors for generation. In contrast, removing the DiT Projector or Sensor-Prior Sampling, which are trained independently from the MLLM, has limited impact on understanding but reduces generative fidelity. This reflects that sensor-level representation remains essential for maintaining structural realism and physical consistency in generation.

4.5 Real-World Deployment and Validation

To evaluate the practicality of UniTac, we deploy UniTac in a real robotic platform and validate both its understanding capability and the utility of generated tactile data.

Understanding in Real Interaction. We apply UniTac in a real-world grasping scenario involving material selection for delicate skin contact, aiming to identify the fabric most suitable for wiping an infant's skin. This task requires subtle tactile discrimination beyond visual similarity. As shown in Fig. 8, given two visually similar fabrics with distinct tactile properties, UniTac analyzes the tactile feedback and performs fine-grained comparison of surface attributes. The model identifies the purple towel as exhibiting lower roughness and a smoother texture. Based on this tactile comparison outcome, the robotic system selects and grasps the smoother fabric as the preferred option for baby skin contact.

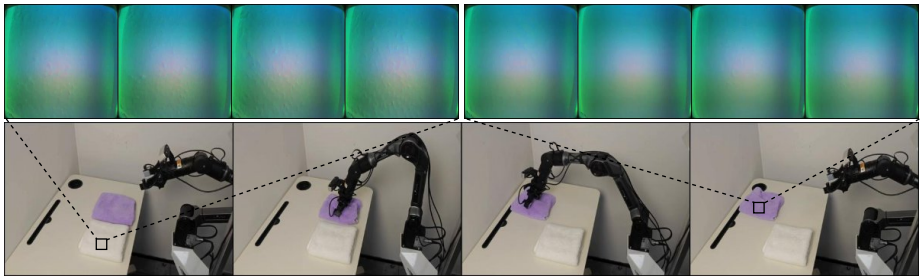


Fig. 8: The robot compares two visually similar fabrics through the Property Comparison task, identifies the smoother one as more suitable for baby skin contact. Tactile differences between the two materials are magnified for clarity.

Table 4: Rollout results for language-guided tactile towel selection and grasping. VLA is an RGB-only baseline that maps visual observations, robot state, and language instructions directly to actions. VTLA-real additionally uses real GelSight tactile observations during rollout. VTLA-pred predicts tactile representations from RGB/state inputs at inference, testing whether tactile-aware control can be achieved without real-time tactile sensing. Implementation details are provided in Appendix B.

Method	Selection Success ↑	Grasping Success ↑	Overall Success ↑
VLA	11/20 (55%)	6/20 (30%)	4/20 (20%)
VTLA-real	20/20 (100%)	19/20 (95%)	19/20 (95%)
VTLA-pred	18/20 (90%)	16/20 (80%)	16/20 (80%)

We further quantify this setting as a language-guided tactile fabric selection and grasping task. A rollout is considered fully successful only when the robot selects the baby-suitable fabric, grasps a single-layer edge or corner, lifts it stably, and avoids visible slipping. As shown in Tab. 4, the RGB-only VLA baseline can often localize the fabric regions visually, but it struggles with tactile-semantic selection and single-layer contact-state estimation. In contrast, VTLA-real achieves the best performance by directly observing tactile cues such as softness, contact area, and slip tendency. VTLA-predict, while not using real-time tactile input during inference, still substantially outperforms VLA, suggesting that predicted tactile representations preserve useful physical cues for contact-aware manipulation. Additional deformation-aware grasping results are provided in Appendix B.1.

Effectiveness of Generated Tactile Data. Tactile sensing hardware evolves rapidly, making large-scale data recollection for each newly introduced device both costly and time-consuming. This challenge is particularly pronounced in cross-sensor scenarios, where models trained on one sensor often fail to generalize to another due to differences in illumination patterns, marker layouts, and deformation characteristics. UniTac alleviates this issue by generating sensor-specific tactile data conditioned on object and contact information, enabling efficient cross-domain adaptation without additional real data collection. We evaluate this capability in a cross-sensor grasp classification task. In Tab. 5,

Table 5: Effectiveness of UniTac-generated GelSight data in improving cross-sensor grasp classification performance. By augmenting with UniTac-generated GelSight data, cross-sensor grasp accuracy improves while maintaining high Digit performance.

Data	Digit Grasp(%)	Gelsight Grasp(%)
Digit	98.89	50.00
Digit+UniTac-Gelsight	99.07	99.37

when trained only on Digit data, the classifier achieves strong in-domain performance (98.89%) but suffers a severe drop when tested on GelSight (50.00%), highlighting the substantial sensor gap. By augmenting the training set with UniTac-generated GelSight samples, the GelSight accuracy improves to 99.37%, while Digit performance remains stable (99.07%). This suggests that UniTac can serve as a scalable data generation framework to support rapid adaptation for emerging tactile devices.

5 Conclusion

This paper introduces UniTac, the first unified multimodal model (UMM) designed for cross-sensor tactile understanding and generation. UniTac bridges the gap between tactile reasoning and synthesis by jointly modeling sensor-level configurations and object-level semantics through a dual-level mixture comprehension framework. A two-stage aligned generation paradigm with a sensor-prior sampling strategy further ensures physically consistent tactile synthesis. Extensive experiments demonstrate that UniTac outperforms existing UMMs and specialized tactile models in both understanding and generation, highlighting its potential to advance unified multimodal modeling.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China under Grant No. 62402430. Hanbin Zhao was also partially supported by the Zhejiang Provincial Natural Science Foundation of China under Grant No. LQN25F020008. Alex Wong was supported by the NSF Athena AI Institute under Grant No. 2112562 and by the Global Industrial Technology Cooperation Center (GITCC) through a grant agreement with the Korea Institute for Advancement of Technology (KIAT), Project No. P0028922.

References

1. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
2. Baishya, S.S., Bäuml, B.: Robust material classification with a tactile skin using deep learning. In: 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 8–15. IEEE (2016)
3. Chen, J., Xu, Z., Pan, X., Hu, Y., Qin, C., Goldstein, T., Huang, L., Zhou, T., Xie, S., Savarese, S., et al.: Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. arXiv preprint arXiv:2505.09568 (2025)

4. Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wu, Y., Wang, Z., Kwok, J., Luo, P., Lu, H., et al.: Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. arXiv preprint arXiv:2310.00426 (2023)
5. Chen, X., Wu, Z., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C.: Janus-pro: Unified multimodal understanding and generation with data and model scaling. arXiv preprint arXiv:2501.17811 (2025)
6. Cheng, N., Xu, J., Chen, J., Han, W.: Stola: Self-adaptive touch-language framework with tactile commonsense reasoning in open-ended scenarios. arXiv preprint arXiv:2505.04201 (2025)
7. Cheng, N., Xu, J., Guan, C., Gao, J., Wang, W., Li, Y., Meng, F., Zhou, J., Fang, B., Han, W.: Touch100k: A large-scale touch-language-vision dataset for touch-centric multimodal representation. Information Fusion p. 103305 (2025)
8. Cheng, Z., Zhang, Y., Zhang, W., Li, H., Wang, K., Song, L., Zhang, H.: Omnivtla: Vision-tactile-language-action model with semantic-aligned tactile sensing. arXiv preprint arXiv:2508.08706 (2025)
9. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
10. Dou, Y., Yang, F., Liu, Y., Loquercio, A., Owens, A.: Tactile-augmented radiance fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26529–26539 (2024)
11. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first international conference on machine learning (2024)
12. Feng, R., Hu, J., Xia, W., Gao, T., Shen, A., Sun, Y., Fang, B., Hu, D.: Anytouch: Learning unified static-dynamic representation across multiple visuo-tactile sensors. arXiv preprint arXiv:2502.12191 (2025)
13. Fu, L., Datta, G., Huang, H., Panitch, W.C.H., Drake, J., Ortiz, J., Mukadam, M., Lambeta, M., Calandra, R., Goldberg, K.: A touch, vision, and language dataset for multimodal alignment. arXiv preprint arXiv:2402.13232 (2024)
14. Gao, R., Deng, K., Yang, G., Yuan, W., Zhu, J.Y.: Tactile dreamfusion: Exploiting tactile sensing for 3d generation. Advances in Neural Information Processing Systems **37**, 29839–29863 (2024)
15. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
16. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
17. Jamali, N., Sammut, C.: Material classification by tactile sensing using surface textures. In: 2010 IEEE International Conference on Robotics and Automation. pp. 2336–2341. IEEE (2010)
18. Johnson, M.K., Adelson, E.H.: Retrographic sensing for the measurement of surface texture and shape. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1070–1077. IEEE (2009)
19. Kerr, J., Huang, H., Wilcox, A., Hoque, R., Ichnowski, J., Calandra, R., Goldberg, K.: Self-supervised visuo-tactile pretraining to locate and follow garment features. arXiv preprint arXiv:2209.13042 (2022)

20. Lambeta, M., Chou, P.W., Tian, S., Yang, B., Maloon, B., Most, V.R., Stroud, D., Santos, R., Byagowi, A., Kammerer, G., et al.: 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 (2020)
21. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al.: Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024)
22. Li, S., Kallidromitis, K., Gokul, A., Liao, Z., Kato, Y., Kozuka, K., Grover, A.: Omniflow: Any-to-any generation with multi-modal rectified flows. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 13178–13188 (2025)
23. Li, Y., Zhu, J.Y., Tedrake, R., Torralba, A.: Connecting touch and vision via cross-modal prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10609–10618 (2019)
24. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
25. Luo, D., Yu, K., Shahidzadeh, A.H., Fermüller, C., Aloimonos, Y., Gao, R.: Controltac: Force-and position-controlled tactile data augmentation with a single reference image. arXiv preprint arXiv:2505.20498 (2025)
26. Luu, Q.K., Nguyen, N.H., et al.: Simulation, learning, and application of vision-based tactile sensing at large scale. IEEE Transactions on Robotics **39**(3), 2003–2019 (2023)
27. Ma, W., Cao, X., Zhang, Y., Zhang, C., Yang, S., Hao, P., Fang, B., Cai, Y., Cui, S., Wang, S.: Cltp: Contrastive language-tactile pre-training for 3d contact geometry understanding. arXiv preprint arXiv:2505.08194 (2025)
28. Ma, Y., Liu, X., Chen, X., Liu, W., Wu, C., Wu, Z., Pan, Z., Xie, Z., Zhang, H., Yu, X., et al.: Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 7739–7751 (2025)
29. Qu, L., Zhang, H., Liu, Y., Wang, X., Jiang, Y., Gao, Y., Ye, H., Du, D.K., Yuan, Z., Wu, X.: Tokenflow: Unified image tokenizer for multimodal understanding and generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2545–2555 (2025)
30. Rodriguez, S., Dou, Y., Oller, M., Owens, A., Fazeli, N.: Touch2touch: Cross-modal tactile generation for object manipulation. arXiv preprint arXiv:2409.08269 (2024)
31. Shi, W., Han, X., Zhou, C., Liang, W., Lin, X.V., Zettlemoyer, L., Yu, L.: Lmfusion: Adapting pretrained language models for multimodal generation. arXiv preprint arXiv:2412.15188 (2024)
32. Stefani, A.L., Bisagno, N., Conci, N., De Natale, F.: Splattouch: Explicit 3d representation binding vision and touch. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 118–127 (2025)
33. Sun, F., Liu, C., Huang, W., Zhang, J.: Object classification and grasp planning using visual and tactile sensing. IEEE Transactions on Systems, Man, and Cybernetics: Systems **46**(7), 969–979 (2016)
34. Team, C.: Chameleon: Mixed-modal early-fusion foundation models. arXiv preprint arXiv:2405.09818 (2024)
35. Tu, J., Fu, H., Yang, F., Zhao, H., Zhang, C., Qian, H.: Texttoucher: Fine-grained text-to-touch generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 7455–7463 (2025)

36. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
37. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
38. Wang, X., Zhang, X., Luo, Z., Sun, Q., Cui, Y., Wang, J., Zhang, F., Wang, Y., Li, Z., Yu, Q., et al.: Emu3: Next-token prediction is all you need. arXiv preprint arXiv:2409.18869 (2024)
39. Xie, E., Chen, J., Chen, J., Cai, H., Tang, H., Lin, Y., Zhang, Z., Li, M., Zhu, L., Lu, Y., et al.: Sana: Efficient high-resolution image synthesis with linear diffusion transformers. arXiv preprint arXiv:2410.10629 (2024)
40. Xie, J., Mao, W., Bai, Z., Zhang, D.J., Wang, W., Lin, K.Q., Gu, Y., Chen, Z., Yang, Z., Shou, M.Z.: Show-o: One single transformer to unify multimodal understanding and generation. arXiv preprint arXiv:2408.12528 (2024)
41. Xie, Y., Li, M., Li, S., Li, X., Chen, G., Ma, F., Yu, F.R., Ding, W.: Universal visuo-tactile video understanding for embodied interaction. arXiv preprint arXiv:2505.22566 (2025)
42. Yang, F., Feng, C., Chen, Z., Park, H., Wang, D., Dou, Y., Zeng, Z., Chen, X., Gangopadhyay, R., Owens, A., et al.: Binding touch to everything: Learning unified multimodal tactile representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26340–26353 (2024)
43. Yang, F., Ma, C., Zhang, J., Zhu, J., Yuan, W., Owens, A.: Touch and go: Learning from human-collected vision and touch. arXiv preprint arXiv:2211.12498 (2022)
44. Yang, F., Zhang, J., Owens, A.: Generating visual scenes from touch. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22070–22080 (2023)
45. Yu, S., Lin, K., Xiao, A., Duan, J., Soh, H.: Octopi: Object property reasoning with large tactile-language models. arXiv preprint arXiv:2405.02794 (2024)
46. Yuan, W., Dong, S., Adelson, E.H.: Gelsight: High-resolution robot tactile sensors for estimating geometry and force. Sensors **17**(12), 2762 (2017)
47. Zhang, C., Hao, P., Cao, X., Hao, X., Cui, S., Wang, S.: Vtla: Vision-tactile-language-action model with preference learning for insertion manipulation. arXiv preprint arXiv:2505.09577 (2025)
48. Zhang, S., Yang, Y., Sun, F., Bao, L., Shan, J., Gao, Y., Fang, B.: A compact visuo-tactile robotic skin for micron-level tactile perception. IEEE Sensors Journal **24**(9), 15273–15282 (2024)
49. Zhou, C., Yu, L., Babu, A., Tirumala, K., Yasunaga, M., Shamis, L., Kahn, J., Ma, X., Zettlemoyer, L., Levy, O.: Transfusion: Predict the next token and diffuse images with one multi-modal model. arXiv preprint arXiv:2408.11039 (2024)