PhysTacGen: Physics-Aware Visual-Tactile Sensor Image Generation

Guo Tang[1] **Yongtao Wang**[1,2]*
[1]Wangxuan Institute of Computer Technology, Peking University
[2] VGI Labs Co., Ltd.
confection_tang@stu.pku.edu.cn
wyt@pku.edu.cn

Abstract

Realistic physical interaction is a cornerstone of embodied intelligence, yet collecting paired visual–tactile data remains costly. Visual-to-tactile synthesis offers a promising approach to augmenting such data, but learning this mapping is complicated by the gap between visual appearance and contact-related material properties, as well as spatial misalignment in paired observations. To address these challenges, we present **PhysTacGen**, a visual-to-optical-tactile image generation framework that integrates material-aware descriptions with geometric conditioning. First, we introduce Group Tactile Policy Optimization (GTPO), a reinforcement learning strategy that refines a vision–language model to generate structured material descriptions using task-specific rewards. Second, we combine DINOv2-based pair curation with monocular relative-depth estimation to select training pairs and provide geometric priors. Finally, an SDXL ControlNet synthesizes optical tactile images conditioned on RGB, relative depth, and GTPO-generated text. Experiments on curated SSVTP data demonstrate improved structural similarity over the compared baselines, while a blinded user study shows a preference for GTPO-generated descriptions. Generated tactile inputs also improve performance on an attribute-derived force-coefficient prediction proxy. Together, these results demonstrate the effectiveness of PhysTacGen for optical tactile image synthesis and its utility in the evaluated downstream task.The code will be available at https://github.com/VDIGPKU/PhysTacGen.

1 Introduction

High-precision manipulation is the cornerstone of next-generation embodied intelligence. Tasks requiring sub-millimetre accuracy—such as threading a cable, inserting a USB connector [33], or handling deformable tissues [5]—often render vision insufficient due to occlusion or lack of contact information. Consequently, tactile sensing has emerged as an imperative modality for the last centimetre of interaction, providing the temporal feedback and subtle surface deformation cues necessary for complex reasoning [2]. The recent surge in Vision-Tactile-Language Models (VTLMs) further underscores the necessity of integrating touch with semantic understanding to achieve robust robot perception [33]. However, a critical bottleneck impedes this progress: the scarcity of large-scale, paired visual-tactile data. Unlike the web-scale abundance of visual data or the maturity of photorealistic visual rendering, collecting high-fidelity tactile datasets is prohibitively expensive, hardware-dependent, and labour-intensive. Furthermore, as highlighted in recent multi-embodied studies [11], the spatial relationship between external cameras and tactile sensors varies significantly across different robotic platforms. This inconsistency in camera-to-sensor layout introduces severe

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

spatial noise into existing datasets, making it difficult for conventional models to learn a robust mapping from 2D visual pixels to 3D tactile deformations. To address this scarcity, prior works have primarily turned to physical simulation. Simulators such as Tacto [26], Taxim [24], and TacSL [1] leverage rendering engines to generate synthetic tactile images. While scalable, these approaches suffer from a fundamental domain gap. The rendered visual counterparts in simulators often lack the photorealism and diverse textures of real-world objects. Furthermore, the simplified contact physics in simulators frequently fails to capture the complex, non-linear deformations of elastomers found in real optical tactile sensors.

Alternatively, data-driven Visual-to-Tactile (V2T) synthesis has gained traction. Approaches like SingleS2R [25] attempt to map visual features directly to tactile responses using pixel-level statistics. However, these methods expose two critical failure modes. First, they fail to distinguish between *visual texture* and *physical compliance*. A model might erroneously generate deep tactile deformations for a hard, marble-patterned surface simply because the visual texture is complex, or conversely, generate a flat response for a texture-less but soft sponge. Second, real-world visual-tactile pairs suffer from severe spatial misalignment due to mechanical tolerances. For deterministic pixel-wise MSE regression, uncertainty over spatial correspondence can average out detail; stochastic generators may instead produce sharp but misplaced features. Without explicit physical grounding and spatial alignment, these generative models hallucinate interactions that violate physical laws.

In this paper, we introduce **PhysTacGen**, a physics-aware generation framework that bridges the gap between visual appearance and tactile mechanics. PhysTacGen moves beyond simple modality mapping by explicitly modeling physical reasoning, filtering spatial noise, and injecting geometric constraints. On the evaluated curated SSVTP split, our method achieves the highest SSIM among the compared systems; the downstream experiment tests substitution of generated tactile images on a heuristic proxy task rather than robot transfer.

These three components form a cohesive pipeline: **GTPO** acts as the 'physical brain' to infer latent properties from semantic cues; the **Data Curation & Prior** stage functions as the 'geometric filter' to resolve spatial noise; and finally, the **SDXL-based Synthesis** serves as the 'integrator' that fuses these semantic and geometric insights into high-fidelity tactile images. To sum up, our system makes three core technical contributions:

- **Group Tactile Policy Optimization (GTPO) for Physical Grounding.** We propose GTPO, a novel alignment approach inspired by reasoning models. GTPO utilizes Reinforcement Learning with a specialized suite of reward functions to refine Large Multimodal Models. By rewarding reasoning chains that align with contact mechanics and tribology, we generate precise, physics-informed textual descriptions.
- **Semantic-Driven Data Curation and Spatial Visual Prior.** To resolve the blurriness caused by spatial misalignment, we introduce a rigorous data curation pipeline using Foundation Models (DINOv2) to filter out geometrically mismatched pairs. Concurrently, we utilize monocular depth estimation to provide a relative macro-geometric prior alongside visual texture.
- **Physics-Driven Multi-Modal Synthesis via SDXL.** We propose a generative architecture powered by Stable Diffusion XL (SDXL) that fuses semantic reasoning with geometric guidance. We utilize the physics-rich text generated by GTPO as the prompt for the dual text encoders, while employing the RGB image and its estimated depth map as 4-channel structural conditions for high-frequency detail generation.

2 Related Work

Optical Tactile Sensing and Simulation. High-resolution optical tactile sensors, such as GelSight [32] and Digit [9], have revolutionized robotic manipulation by providing dense geometric information. Recent advancements have further integrated multimodal fingertips to digitize touch [10]. To mitigate the high cost of real-world data collection, simulators like Tacto [26], Taxim [24] and TacSL [1] leverage rendering engines to synthesize tactile images. Other efforts utilize path tracing and IMPM to simulate complex dynamic contact like slip and rotation [23]. However, these analytical simulators often rely on simplified contact models that fail to capture the complex, non-linear deformations of elastomers found in real-world environments.

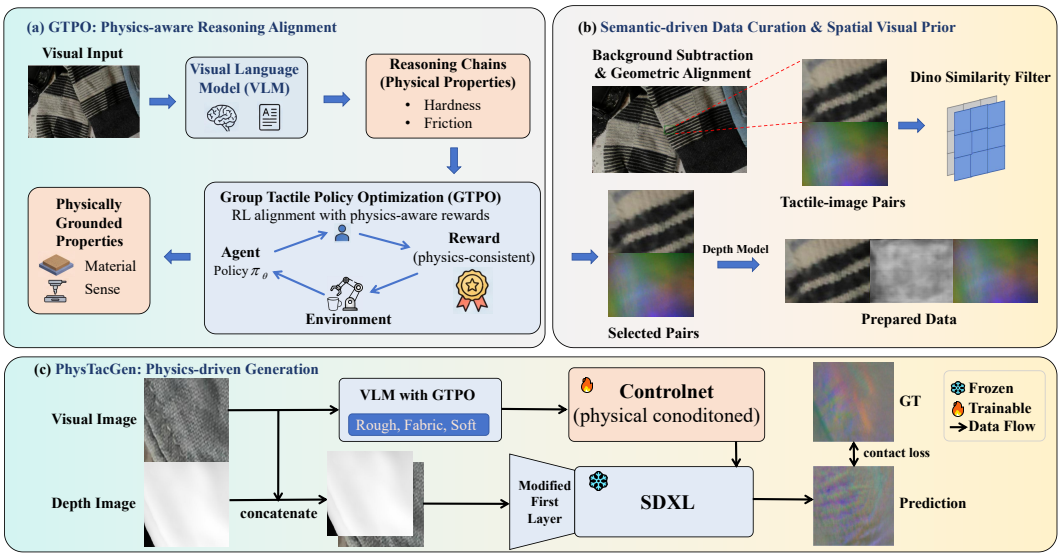

Figure 1: **The PhysTacGen Framework.** Given a visual observation, our system operates in three sequential stages: (1) **Physics-Aware Reasoning:** A VLM fine-tuned via *GTPO* infers latent physical properties. (2) **Data Curation & Spatial Prior:** The dataset is semantically purified using DINOv2, and input RGB is depth-estimated to form a 4-channel geometric condition. (3) **Multi-Modal Synthesis:** A physics-driven SDXL ControlNet synthesises the high-fidelity tactile sensor image I_t. Here I_v is the external visual RGB image, I_d the estimated relative-depth prior, and c_{phys} the GTPO-generated physical description; the optical tactile image is I_t.

Tactile Representation and Foundation Models. With the advent of self-supervised learning, models like Sparsh [6] and AnyTouch [2] have established unified representations across diverse sensors. These frameworks often leverage vision foundation models such as MAE [4] or multi-modal alignment strategies like CLIP [19] to bind touch to everything [29]. Specifically, CLTP [14] focuses on 3D contact geometry understanding, while Octopi [31] utilizes Large Multimodal Models for object property reasoning. Our work builds on this trend but shifts the focus from representation to generation, specifically addressing the semantic-physical gap that representation-only models might overlook.

Visual-to-Tactile (V2T) Synthesis and Force Perception. Mapping external RGB to optical tactile observations has been studied using image translation and reference-conditioned synthesis. Pix2Pix [7] provides a general paired translation baseline; Pix2Pix-Turbo [17] uses a one-step text-to-image backbone. VisGel [12] uses a query RGB image and a reference RGB–tactile pair, whereas Controllable Visual-Tactile Synthesis [3] generates visual and tactile outputs from editable sketches. The latter uses a different input and output protocol from our local optical-tactile prediction. Latent Diffusion Models [20], PolyTouch [34], SingleS2R [25], and contact-force estimation [21] provide further context, but the experiments below assess tactile image synthesis and a heuristic coefficient proxy, not measured-force control.

Tactile-Aware Manipulation and Real-world Challenges. Dynamic tactile perception is essential for contact-rich tasks, as explored in Reactive Diffusion Policy [28] and ViTacformer [5]. The integration of language has led to the emergence of Vision-Tactile-Language-Action (VTLA) models [33]. Despite these advances, real-world data collection remains plagued by spatial misalignment between cameras and sensors, especially in hardware-independent interfaces like Fast-UMI [27]. Large-scale datasets such as SSVTP [8] highlight the need for robust cross-modal alignment. Our framework contributes to this ecosystem by generating spatially aligned, physics-aware tactile data, motivating future validation in real manipulation tasks.

3 Method

The goal of PhysTacGen is to synthesise high-fidelity tactile sensor images $I_t \in \mathbb{R}^{H\times W\times 3}$ given a visual observation $I_v \in \mathbb{R}^{H\times W\times 3}$. As illustrated in Figure 1, the framework operates in three sequential stages to explicitly disentangle *semantic physical reasoning* from *geometric deformation generation*. Optical tactile RGB records a deformed, internally illuminated sensor membrane, rather than the contacting objectś visible color; its color-to-gradient mapping depends on the sensor calibration. The RGB input I_v yields both the physical description c_{phys} and estimated relative depth I_d, which condition the generator.

3.1 Group Tactile Policy Optimisation (GTPO)

Standard Supervised Fine-Tuning (SFT) often leads to superficial alignment where models mimic the style of training captions without grounding the underlying physics. Traditional RLHF requires training a critic model, which is extremely expensive and unstable in multimodal large models since the value model also needs to look at images, leading to memory explosion. To address this, we propose **Group Tactile Policy Optimisation (GTPO)**, a reinforcement learning framework built upon GRPO [22]. However, standard GRPO is primarily designed for general semantic alignment or mathematical reasoning and lacks the domain-specific physical common sense required for tactile tasks. It cannot inherently parse the subtle relationship between visual textures and latent material mechanics. To bridge this gap, we specialize the framework into GTPO, which adapts the group-based reward mechanism to the nuances of tactile physics. GTPO refines the Large Multimodal Model (LMM) to infer latent physical properties from visual cues by rewarding reasoning chains that strictly adhere to physical laws. Its primary objective is to serve as a semantic physical interpreter, which guides the subsequent image synthesis process.

Objective Function. Given a visual prompt x, we sample a group of G outputs $\{y_1, y_2, \ldots, y_G\}$ from the current policy π_θ. Unlike PPO, GRPO eliminates the need for a critic model by using group-based advantage estimation. The optimization objective is formulated as:

$$\rho_i(\theta) = \frac{\pi_\theta(y_i|x)}{\pi_{\theta_{\text{old}}}(y_i|x)}, \tag{1a}$$

$$\mathcal{L}_i(\theta) = \min\Big(\rho_i \hat{A}_i, \text{clip}(\rho_i, 1-\epsilon, 1+\epsilon)\hat{A}_i\Big), \tag{1b}$$

$$\mathcal{J}(\theta) = \mathbb{E}_{\mathcal{B}}\left[\frac{1}{G}\sum_{i=1}^{G}\Big(\mathcal{L}_i(\theta) - \beta\mathbb{D}_{\text{KL}}(\pi_\theta||\pi_{\text{ref}})\Big)\right], \tag{1c}$$

where ρ_i is the importance sampling ratio, ϵ is the clipping coefficient, and $\hat{A}_i$ is the group-normalized advantage. The policy π_{ref} is the frozen SFT reference; $\mathcal{B}$ is a training batch. The start/end tokens $t^{\text{r}}_{\text{start}}, t^{\text{r}}_{\text{end}}$ and $t^{\text{s}}_{\text{start}}, t^{\text{s}}_{\text{end}}$ delimit reasoning and solution sections, and $\mathcal{V}_{\text{phys}}$ is the predefined physical vocabulary. The total reward $R(y)$ is a weighted sum of three distinct physical alignment components: $R(y) = R_{\text{fmt}}(y) + R_{\text{sem}}(y, y^*) + R_{\text{vocab}}(y)$.

Reward Design. R_{fmt} enforces a structured Chain-of-Thought (CoT) process using reasoning and solution delimiters:

$$R_{\text{fmt}}(y) = 0.5 \cdot \mathbb{I}\left(t^{\text{r}}_{\text{start}} \in y \wedge t^{\text{r}}_{\text{end}} \in y\right) + 0.5 \cdot \mathbb{I}\left(t^{\text{s}}_{\text{start}} \in y \wedge t^{\text{s}}_{\text{end}} \in y\right), \tag{2}$$

R_{sem} aligns the generated physical description with ground truth physics in a semantic embedding space. We extract the text within the solution tags as y_{pred} and compare it with the ground truth y^*. Using a pre-trained sentence encoder $E(\cdot)$, we compute the cosine similarity $s_{\cos}$. To penalize hallucinations and reward high-fidelity descriptions, we apply a clamped linear transformation:

$$s_{\cos} = \frac{E(y_{\text{pred}}) \cdot E(y^*)}{\|E(y_{\text{pred}})\|\|E(y^*)\|}, \quad R_{\text{sem}}(y) = \text{clip}\left(1.5 \cdot (s_{\cos} + 0.1),\ 0.0,\ 2.0\right). \tag{3}$$

Finally, R_{vocab} measures the density of valid physical terms in the prediction to prevent the generation of useless descriptions:

$$R_{\text{vocab}}(y) = 0.5 \cdot \frac{\sum_{w \in W(y_{\text{pred}})} \mathbb{I}(w \in \mathcal{V}_{\text{phys}})}{|W(y_{\text{pred}})|}, \tag{4}$$

where $W(y_{\text{pred}})$ denotes the set of words in the generated solution.

Algorithm 1 Group Tactile Policy Optimization (GTPO)

Input: Dataset $\mathcal{D}$, Initial Policy π_θ, Reference Model π_{ref}, Semantic Encoder $E(\cdot)$, Group Size G, Learning Rate η, KL Coefficient β, Clip Ratio ϵ.
Initialize $\pi_{\text{ref}} \leftarrow \pi_\theta$
repeat
  Sample a batch of visual prompts and ground truths $\mathcal{B} \sim \mathcal{D}$
  Update old policy $\pi_{\theta_{\text{old}}} \leftarrow \pi_\theta$
  for each visual prompt $x \in \mathcal{B}$ **do**
    Sample G outputs $\{y_1, y_2, \ldots, y_G\} \sim \pi_{\theta_{\text{old}}}(\cdot|x)$
    for $i = 1$ **to** G **do**
      Rule-based Formatting Reward: $r_{\text{fmt}} \leftarrow R_{\text{fmt}}(y_i)$
      Rule-based Vocabulary Reward: $r_{\text{vocab}} \leftarrow R_{\text{vocab}}(y_i)$
      Model-based Semantic Reward: $r_{\text{sem}} \leftarrow \text{clip}(1.5 \cdot (\cos(E(y_i), E(y^*)) + 0.1), 0, 2)$
      $r_i \leftarrow r_{\text{fmt}} + r_{\text{sem}} + r_{\text{vocab}}$
    end for
    Compute Mean $\mu = \frac{1}{G}\sum_{j=1}^{G} r_j$ and Std $\sigma = \sqrt{\frac{1}{G}\sum_{j=1}^{G}(r_j - \mu)^2}$
    for $i = 1$ **to** G **do**
      $A_i \leftarrow (r_i - \mu)/(\sigma + \delta)$
    end for
  end for
  Optimize θ by maximizing the objective $\mathcal{J}_{\text{GTPO}}$:
  for each y_i in group **do**
    ratio $\rho_i = \pi_\theta(y_i|x)/\pi_{\theta_{\text{old}}}(y_i|x)$, **surrogate** $L_i^{\text{clip}} = \min(\rho_i A_i, \text{clip}(\rho_i, 1-\epsilon, 1+\epsilon)A_i)$
  end for
  Optimize θ by maximizing $\mathcal{J}$ in Eq. 1c
until convergence
Output: Optimized Physics-Aware Policy π_θ

3.2 Semantic-Driven Data Curation and Visual Prior

Real-world visual-tactile datasets often suffer from severe spatial misalignment due to mechanical tolerances. Spatial uncertainty can blur the outputs of deterministic pixel-wise regression and displace details in stochastic synthesis.

DINOv2 Semantic Filtering. To resolve this, we employ DINOv2 [15], a foundation model highly sensitive to material textures, to compute the semantic correlation between visual crops and tactile readings. Let $f_{\text{DINO}}(\cdot)$ denote the CLS token feature extraction of DINOv2. For each data pair (I_v, I_t), we compute the semantic similarity score:

$$S_{\text{DINO}} = \frac{f_{\text{DINO}}(I_v) \cdot f_{\text{DINO}}(I_t)}{\|f_{\text{DINO}}(I_v)\|_2 \|f_{\text{DINO}}(I_t)\|_2}. \tag{5}$$

We use CLS features from `facebook/dinov2-base`, rank calibrated RGB–tactile pairs by this cosine score, and discard the bottom 30%. This feature-based heuristic is applied before the split to all pairs; it does not use target labels or evaluation feedback, but it shapes the retained test distribution. Cross-modal DINO similarity is not a calibrated geometric alignment measure.

Depth Isolation and Calibration. Concurrently, we explicitly model the transformation $\mathcal{T} : \mathcal{I}_v \to \mathcal{I}_t$ to map the visual observation to the tactile sensor space using center-crop operations and rotations. We use the pretrained relative-depth model `Depth-Anything-V2-Large-hf` [30]. Its output is min–max normalized per image to $[0, 255]$, stored as uint8, and scaled to $[0, 1]$ for ControlNet. This is a macro-geometric prior, not calibrated metric depth or the contact-induced elastomer deformation field.

3.3 Physics-Driven Multi-Modal Tactile Image Generation

To synthesize high-fidelity tactile responses that respect both macroscopic geometry and microscopic material properties, we adopt Stable Diffusion XL (SDXL) [18] as our generative backbone. Unlike

earlier models, SDXL features dual text encoders (CLIP ViT-L and OpenCLIP ViT-bigG), which significantly enhances its comprehension of complex GTPO-generated physical attributes.

Four-Channel Geometric-Aware Input Representation. We augment the aligned RGB prior I_{rgb} with its corresponding depth map I_{depth} to form a 4-channel control tensor $\mathbf{C} \in \mathbb{R}^{H\times W\times 4}$. We modify the first convolutional layer of the SDXL ControlNet to accommodate this input. To preserve the pretrained visual feature extraction capabilities, we employ a channel-wise weight inheritance strategy. Let $\mathbf{W}_{\text{old}} \in \mathbb{R}^{K\times 3\times S\times S}$ be the original weights. The new weights $\mathbf{W}_{\text{new}} \in \mathbb{R}^{K\times 4\times S\times S}$ are initialized as:

$$\mathbf{W}_{\text{new}}^{(c)} = \begin{cases} \mathbf{W}_{\text{old}}^{(c)} & \text{if } c \in \{0,1,2\}, \\ \frac{1}{3}\sum_{j=0}^{2} \mathbf{W}_{\text{old}}^{(j)} & \text{if } c = 3, \end{cases} \tag{6}$$

This initialization ensures that the depth channel starts with a neutral but feature-aware contribution. The SDXL VAE, U-Net, and text encoders remain frozen; only ControlNet is trained.

Physically-Grounded Training Strategy. To prevent the model from overfitting to the non-contact background, we propose a **Contact-Aware Loss Reweighting** strategy. We generate a spatial weight mask $\mathbf{M}$ based on the depth prior $I_{\text{depth}} \in [0,1]$:

$$\mathbf{M}_{u,v} = \begin{cases} \lambda_{\text{boost}} & \text{if } I_{\text{depth}}(u,v) > \tau, \\ 1.0 & \text{otherwise}, \end{cases} \tag{7}$$

where $\tau = 0.5$ and $\lambda_{\text{boost}} = 5.0$. The depth mask is resized to latent resolution. The diffusion training objective incorporates this mask to focus gradient descent on reconstructing fine-grained deformation details:

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

where $\odot$ denotes element-wise multiplication, $\mathbf{z}_t$ is the noisy latent, and $\boldsymbol{\epsilon}_\theta$ is the predicted noise. This forces the model to focus gradient descent steps on reconstructing deformation details.

Furthermore, we implement a **Classifier-Free Guidance with Text Dropout**, randomly dropping the physics-aware text condition $\mathbf{c}_{\text{txt}}$ with a probability $p = 0.15$ during training. Finally, a **Physics-Prompt Normalization** protocol is applied to parse the LMM reasoning output into a structured set of concise physical tags.

4 Experiments

4.1 Implementation Details

We train PhysTacGen on SSVTP, which contains 4,587 original visual–tactile pairs with human material descriptions. DINOv2 curation retains 70% of calibrated pairs before the train/test split. Paired crop/rotation augmentation yields approximately 13k training instances, and SSIM/LPIPS use approximately 1.2k held-out processed test instances; the separate reasoning benchmark has 452 pairs. All compared image-generation methods use the same retained split. The GTPO model is Qwen3-VL-8B-Instruct with group size 4, learning rate 2×10^{-6}, KL coefficient 0.05, clip ratio 0.2, and 1,000 training steps. The format, semantic, and vocabulary reward scales are 1.0, 1.5, and 0.5. Its vision encoder, vision–language adapter, and output head are frozen, while LoRA trains language-model attention and MLP modules; the reference policy is the frozen SFT model. For the SDXL generator, RGB and relative depth form the four-channel ControlNet condition. We train only ControlNet on 512×512 images with AdamW at 10^{-5}, fp16, batch size 1 with gradient accumulation 8, 10,000 steps, and 15% text dropout. Final GTPO and generator runs used NVIDIA A100 GPUs; RTX 3090/4090 systems were used only for pilot work. SSIM uses aligned RGB images with `data_range=255`; LPIPS uses VGG with inputs in $[-1,1]$. We report legacy nFD, computed by dividing the Fréchet feature distance by the evaluation-set size,which is not standard FID.

4.2 Visual-to-Tactile Synthesis Quality

We compare PhysTacGen with three external baselines and the original generator variants on the same curated SSVTP split. Pix2Pix [7] uses U-Net-256/PatchGAN at 512×512 for 200 epochs;

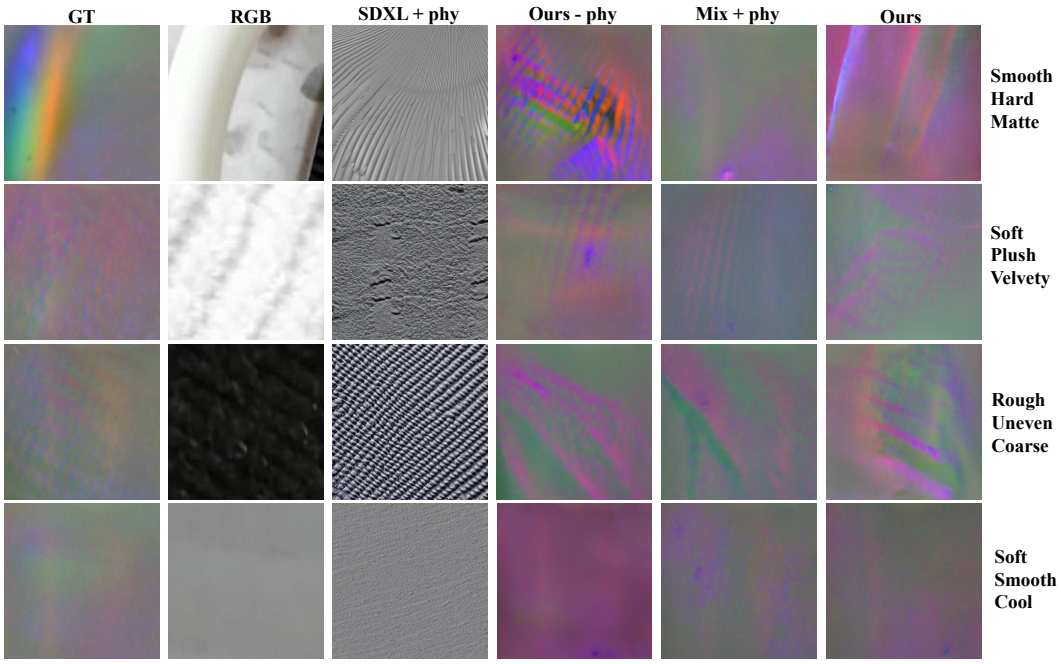

Figure 2: **Qualitative Synthesis Results.** Visualisation of generated tactile images across different settings. SDXL (Baseline) fails to ground the contact location. The Mix-Fusion and four-channel variants change both fusion and loss and therefore do not isolate a single cause of their visual differences. The examples illustrate generated contact appearance under the original conditions.

Table 1: External baselines and original generator variants. All rows use the same retained SSVTP split. The last column is the legacy Fréchet statistic divided by evaluation-set size (legacy nFD), *not* FID.

Method and condition	SSIM ↑	LPIPS ↓	legacy nFD ↓
Pix2Pix / RGB	0.106	0.819	0.829
Pix2Pix-Turbo / RGB + generic prompt	0.417	0.692	0.628
VisGel (adapted) / RGB + train reference	0.715	0.629	0.433
SDXL baseline / physical text only	0.190	0.859	0.838
Mix-Fusion / RGB + depth, physical text	0.938	0.617	**0.241**
Four-channel / RGB + depth, generic text	0.908	0.637	0.349
PhysTacGen / RGB + depth, GTPO text	**0.964**	**0.559**	0.278

Pix2Pix-Turbo [17] uses one-step SD-Turbo at 512×512 for 6,000 steps with a fixed generic prompt. VisGel [12] uses the query RGB plus a reference RGB–tactile pair retrieved using RGB features from the training split only; query tactile data are never accessed. Because SSVTP lacks VisGel's original temporal/no-contact reference, this baseline is an adaptation.

Table 1 shows the highest SSIM and lowest LPIPS for PhysTacGen among the original variants and the evaluated external baselines; Mix-Fusion obtains the lowest legacy nFD. Since Mix-Fusion changes both fusion and loss, that row cannot isolate either effect. The generic-text variant tests the impact of semantic text with the same four-channel architecture and contact-aware training.

Matched prompt-source comparison. For a matched comparison of prompt sources, we train three separate ControlNet generators from the same pretrained initialization.Classifier, SFT, and GTPO prompts are each used during both training and test time for their own model. The DINO-retained pairs, split, RGB-depth inputs, architecture, contact-aware loss, optimizer, training steps, random seed, and inference settings are held fixed. The seven-label classifier is trained from original SSVTP

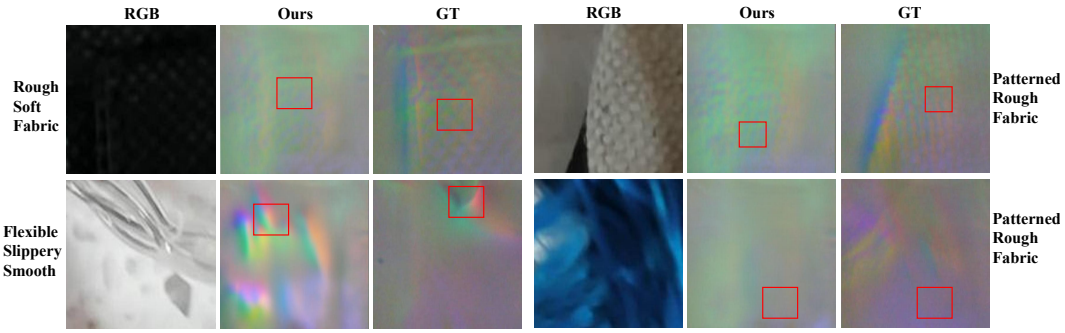

Figure 3: **Material and Textural Consistency under Spatial Misalignment.** Red boxes highlight areas for qualitative texture comparison.The details in the areas outlined in red in each set of images are similar.

Table 2: Final matched end-to-end prompt-source comparison. Each prompt source trains and tests a separate generator under the same conditions; legacy nFD is not standard FID.

Prompt source	SSIM ↑	LPIPS ↓	legacy nFD ↓
7-label classifier	0.919	**0.557**	0.305
SFT	0.948	0.570	**0.255**
GTPO	**0.964**	0.559	0.278

human descriptions, not distilled from GTPO; its vocabulary is textured, hard, smooth, rough, soft, flexible, and fabric. Since captions may omit valid properties, absent labels need not be true negatives.

Relative to SFT, GTPO improves SSIM from 0.948 to 0.964 and LPIPS from 0.570 to 0.559; SFT has lower legacy nFD. Relative to the classifier, GTPO improves SSIM and legacy nFD, while the classifier has LPIPS lower by 0.002. A simplified lexical consistency check counts descriptions containing both hard and soft or both smooth and rough. Contradiction rates are 7.6% for the classifier, 3.2% for SFT, and 1.3% for GTPO. Figure 3 offers qualitative examples of residual spatial offsets.

4.3 Downstream Application: Synthetic-Tactile Substitution for Proxy Force-Coefficient Prediction

The force-coefficient target $y \in [0.1, 0.9]$ is assigned heuristically from normalized SSVTP material descriptions, rather than measured with a force sensor. In the fixed attribute mapping, "slippery" adds 0.25 and "soft" subtracts 0.15. A late-fusion ResNet-18 $\pi(I_v, I_t)$ is trained on real RGB–real tactile pairs with a 90:10 split. At test time we compare RGB-only, RGB plus generated tactile, and RGB plus real tactile (oracle). This tests substitution of a synthetic observation on the same proxy task; it does not test simulator-to-robot transfer or closed-loop manipulation.

Generated-tactile input improves on RGB-only in both runs, and its MSE is 0.002 above the real-tactile oracle in each. As the target is built from material attributes also used in the semantic pipeline, these results indicate consistency on a proxy task, not measured-force utility. Two final generator models trained on independently selected 9:1 folds differed by less than 5% in SSIM/LPIPS.

4.4 Comprehensive Evaluation of Physical Reasoning (GTPO)

We assess GTPO-generated physical descriptions against SFT and a seven-label classifier on the standardized 452-pair benchmark, using lexical metrics, an auxiliary automatic judge, and blinded human preference.

Data Curation and Vocabulary Standardization. The raw SSVTP dataset contains unstructured human descriptions often polluted by purely visual adjectives or vague terms. To construct a high-quality ground truth for physical reasoning, we implemented a rigorous text normalization pipeline: (1) **Visual Noise Filtering:** We constructed a blacklist to discard non-tactile descriptors. (2) **Physical**

Table 3: Test MSE for the heuristic force-coefficient prediction proxy. Sample SD describes variation across from two random seeds; the last column reports their mean and sample standard deviation.

Input modality	Run 1	Run 2	Mean $\pm$ sample SD
RGB only	0.144	0.147	0.1455 ± 0.0021
RGB + generated tactile	0.132	0.134	0.1330 ± 0.0014
RGB + real tactile (oracle)	0.130	0.132	0.1310 ± 0.0014

Table 4: Description evaluation on 452 pairs. The human column is a blinded forced choice among classifier, SFT, and GTPO descriptions by 22 participants, with randomized presentation and one choice per image.

Method	SES ↑	BLEU ↑	ROUGE ↑	PPL ↓	Judge ↑	Human ↑
7-label classifier	**0.732**	**29.3**	**52.7**	12.9	2.94	16%
SFT Qwen	0.461	18.4	32.7	14.3	3.42	27%
GTPO Qwen	0.586	26.8	44.5	**10.8**	**4.65**	**57%**

Attribute Normalization: We mapped synonymous adjectives to a standardized physics vocabulary. For instance, terms like “rigid", “stiff", and “ceramic" are mapped to the canonical label *“hard"*. This process yielded a refined dataset of 452 unique object-material pairs for testing.

Evaluation Metrics. We employ BLEU [16] and ROUGE [13] to measure linguistic overlap with expert annotations, while Perplexity (PPL) is used to measure fluency. As an auxiliary automatic measure, we use a **VLM Physics Judge**. We prompt GPT-4o to act as an expert which reviews the generated descriptions for apparent physical plausibility and outputs a *Physics Plausibility Score* between 1-5. Furthermore, we extract the physical keywords from the tags and compute the Semantic Embedding Similarity (SES) using `all-MiniLM-L6-v2` to capture objective vector-space alignment.

Results Analysis. As Table 4 shows, the closed-vocabulary classifier leads SES, BLEU, and ROUGE; this favors frequent canonical words. GTPO has the lowest reported PPL, highest auxiliary judge score (4.65), and highest human preference (57%). The study used 22 participants; every evaluated image was presented to all 22, with descriptions in randomized order and no ties. These measures support the quality of GTPO descriptions but do not directly validate contact mechanics. The matched generator table additionally shows a modest SSIM and LPIPS advantage over SFT, with trade-offs against the classifier.

5 Limitations and Broader Impacts

Limitations and Future Work. The generation results are limited to one curated SSVTP dataset. DINO filtering occurred before the split, shaping the retained test distribution, and direct DINO feature similarity across camera and optical tactile images is a heuristic. Homogeneous surfaces and residual spatial offsets can still cause errors. The two-seed and two-fold checks show consistency but not formal significance or cross-sensor generalization. The force-coefficient labels are heuristic proxies, so the downstream result cannot demonstrate measured force or real robot transfer. The distributional comparison uses legacy nFD rather than standard FID/KID. The heuristic coefficient mapping is specified only in part, and the current experiments do not isolate every component or assess cross-sensor generalization.

Broader Impacts. This study investigates synthetic optical tactile observations as a potential supplement to costly paired data collection. Generated outputs can be inaccurate, especially outside the curated distribution; applications involving force control or safety-critical contact require validation with calibrated sensors and real robots.

6 Conclusion

We presented **PhysTacGen**, which combines GTPO-derived material descriptions, DINOv2-based pair curation, relative-depth conditioning, and an SDXL ControlNet for visual-to-optical-tactile

synthesis. On the curated SSVTP test split, PhysTacGen achieves an SSIM of 0.964 and an LPIPS of 0.559. In a matched comparison of separately trained generators, GTPO prompts yield the highest SSIM, while seven-label-classifier prompts achieve a marginally lower LPIPS. In a blinded study with 22 participants, GTPO descriptions receive the highest preference rate of 57%. Generated tactile inputs also reduce test MSE relative to RGB-only inputs in both runs of the attribute-derived force-coefficient proxy task. Together, these results demonstrate PhysTacGen's optical tactile image synthesis performance and its utility for the evaluated proxy task on curated SSVTP data.

Broader Impacts Statement. This work investigates physics-aware visual-to-optical-tactile image generation to support research on multimodal perception and embodied intelligence. Synthetic tactile observations may supplement costly paired visual–tactile data collection and facilitate the development of tactile learning methods. However, generated images can contain inaccurate material or contact cues, particularly outside the evaluated data distribution. Using these outputs in force control or contact-rich robotic applications without further validation could lead to unreliable decisions or unsafe interactions. Practical deployment therefore requires validation with calibrated tactile sensors and real robots.

Acknowledgments. This work was supported in part by the Beijing Major Science and Technology Project under Contract No. Z261100000226036.

References

[1] Iretiayo Akinola, Jie Xu, Jan Carius, Dieter Fox, and Yashraj Narang. Tacsl: A library for visuotactile sensor simulation and learning. *IEEE Transactions on Robotics*, 2025.

[2] Ruoxuan Feng, Jiangyu Hu, Wenke Xia, Tianci Gao, Ao Shen, Yuhao Sun, Bin Fang, and Di Hu. Anytouch: Learning unified static-dynamic representation across multiple visuo-tactile sensors. *arXiv preprint arXiv:2502.12191*, 2025.

[3] Ruihan Gao, Wenzhen Yuan, and Jun-Yan Zhu. Controllable visual-tactile synthesis. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7017–7029. IEEE, 2023.

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

[5] Liang Heng, Haoran Geng, Kaifeng Zhang, Pieter Abbeel, and Jitendra Malik. Vitacformer: Learning cross-modal representation for visuo-tactile dexterous manipulation. *arXiv preprint arXiv:2506.15953*, 2025.

[6] Carolina Higuera, Akash Sharma, Chaithanya Krishna Bodduluri, Taosha Fan, Patrick Lancaster, Mrinal Kalakrishnan, Michael Kaess, Byron Boots, Mike Lambeta, Tingfan Wu, et al. Sparsh: Self-supervised touch representations for vision-based tactile sensing. *arXiv preprint arXiv:2410.24090*, 2024.

[7] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pages 5967–5976. Ieee, 2017.

[8] Justin Kerr, Huang Huang, Albert Wilcox, Ryan Hoque, Jeffrey Ichnowski, Roberto Calandra, and Ken Goldberg. Self-supervised visuo-tactile pretraining to locate and follow garment features. *arXiv preprint arXiv:2209.13042*, 2022.

[9] Mike Lambeta, Po-Wei Chou, Stephen Tian, Brian Yang, Benjamin Maloon, Victoria Rose Most, Dave Stroud, Raymond Santos, Ahmad Byagowi, Gregg Kammerer, 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.

[10] Mike Lambeta, Tingfan Wu, Ali Sengul, Victoria Rose Most, Nolan Black, Kevin Sawyer, Romeo Mercado, Haozhi Qi, Alexander Sohn, Byron Taylor, et al. Digitizing touch with an artificial multimodal fingertip. *arXiv preprint arXiv:2411.02479*, 2024.

[11] Hongyu Li, Mingxi Jia, Tuluhan Akbulut, Yu Xiang, George Konidaris, and Srinath Sridhar. V-hop: Visuo-haptic 6d object pose tracking. *arXiv preprint arXiv:2502.17434*, 2025.

[12] Yunzhu Li, Jun-Yan Zhu, Russ Tedrake, and Antonio Torralba. Connecting touch and vision via cross-modal prediction. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10601–10610. IEEE, 2019.

[13] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[14] Wenxuan Ma, Xiaoge Cao, Yixiang Zhang, Chaofan Zhang, Shaobo Yang, Peng Hao, Bin Fang, Yinghao Cai, Shaowei Cui, and Shuo Wang. Cltp: Contrastive language-tactile pre-training for 3d contact geometry understanding. *Biomimetic Intelligence and Robotics*, page 100324, 2026.

[15] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[16] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[17] Gaurav Parmar, Taesung Park, Srinivasa Narasimhan, and Jun-Yan Zhu. One-step image translation with text-to-image models. *arXiv preprint arXiv:2403.12036*, 2024.

[18] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

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

[20] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[21] Amir-Hossein Shahidzadeh, Gabriele M Caddeo, Koushik Alapati, Lorenzo Natale, Cornelia Fermüler, and Yiannis Aloimonos. Feelanyforce: Estimating contact force feedback from tactile sensation for vision-based tactile sensors. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 251–257. IEEE, 2025.

[22] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[23] Zirong Shen, Yuhao Sun, Shixin Zhang, Zixi Chen, Heyi Sun, Fuchun Sun, and Bin Fang. Simulation of optical tactile sensors supporting slip and rotation using path tracing and impm. *IEEE Robotics and Automation Letters*, 9(12):11218–11225, 2024.

[24] Zilin Si and Wenzhen Yuan. Taxim: An example-based simulation model for gelsight tactile sensors. *IEEE Robotics and Automation Letters*, 7(2):2361–2368, 2022.

[25] Jing Tang, Zeyu Gong, Bo Tao, and Zhouping Yin. Singles2r: Single sample driven sim-to-real transfer for multi-source visual-tactile information understanding using multi-scale vision transformers. *Information Fusion*, 108:102390, 2024.

[26] Shaoxiong Wang, Mike Lambeta, Po-Wei Chou, and Roberto Calandra. Tacto: A fast, flexible, and open-source simulator for high-resolution vision-based tactile sensors. *IEEE Robotics and Automation Letters*, 7(2):3930–3937, 2022.

[27] Ziniu Wu, Tianyu Wang, Chuyue Guan, Zhongjie Jia, Shuai Liang, Haoming Song, Delin Liu, Dong Wang, Zhigang Wang, Nieqing Cao, et al. Fast-umi: A scalable and hardware-independent universal manipulation interface. 2024.

[28] Han Xue, Jieji Ren, Wendi Chen, Gu Zhang, Yuan Fang, Guoying Gu, Huazhe Xu, and Cewu Lu. Reactive diffusion policy: Slow-fast visual-tactile policy learning for contact-rich manipulation. *arXiv preprint arXiv:2503.02881*, 2025.

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

[30] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in neural information processing systems*, 37:21875–21911, 2024.

[31] Samson Yu, Kelvin Lin, Anxing Xiao, Jiafei Duan, and Harold Soh. Octopi: Object property reasoning with large tactile-language models. *arXiv preprint arXiv:2405.02794*, 2024.

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

[33] Chaofan Zhang, Peng Hao, Xiaoge Cao, Xiaoshuai Hao, Shaowei Cui, and Shuo Wang. Vtla: Vision-tactile-language-action model with preference learning for insertion manipulation. *Biomimetic Intelligence and Robotics*, page 100333, 2026.

[34] Jialiang Zhao, Naveen Kuppuswamy, Siyuan Feng, Benjamin Burchfiel, and Edward Adelson. Polytouch: A robust multi-modal tactile sensor for contact-rich manipulation using tactile-diffusion policies. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 104–110. IEEE, 2025.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction describe optical tactile image synthesis, material-description evaluation, and an attribute-derived force-coefficient prediction proxy. These claims are supported by Section 4, with their scope and limitations discussed in Section 5.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 5 discusses the curated single-dataset evaluation, residual spatial discrepancies, heuristic proxy targets, and the scope of the generalization and downstream claims.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper focuses on empirical deep learning and generative modeling. It does not introduce new theoretical mathematical proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Section 3 describes the proposed pipeline in detail, including data curation, material-description optimization, and tactile image generation. Section 4 provides the training settings and evaluation protocols.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We use the existing SSVTP dataset and plan to release part of the implementation at `https://github.com/VDIGPKU/PhysTacGen`. A complete code-and-data reproduction package is not currently provided.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4.1 describes the data retention and splits, optimizer, learning rates, batch size, training steps, and evaluation settings.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Table 3 reports results from two random seeds, including the mean and sample standard deviation. The main image-generation comparisons are reported without error bars or statistical significance tests.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Section 4.1 reports the GPU hardware used for our experiments and the corresponding execution times.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research conforms to the NeurIPS Code of Ethics. We do not use proprietary datasets without permission or conduct harmful research.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: The Broader Impacts Statement in Section 5 discusses the potential of synthetic tactile observations to support multimodal perception research and supplement costly paired data collection. It also discusses risks from inaccurate generated cues in robotic applications and the need for validation with calibrated tactile sensors and real robots.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The proposed generator targets optical tactile sensor images and is not intended for general-purpose content generation. Potential application risks from inaccurate tactile predictions are discussed in Section 5.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The original creators of all models, code, and datasets used in this work are properly cited. We comply with the licenses of all models, code, and datasets used in this work.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper uses the existing SSVTP dataset. No new dataset or model-checkpoint release accompanies the current version, and an implementation release is planned.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [No]

    Justification: Section 4.4 describes a blinded forced-choice study with 22 participants, randomized presentation, and one choice per image. The full participant instructions and compensation details are not included.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [No]

Justification: The paper reports a 22-participant preference study, but does not state its institutional review or exemption status or describe the information provided to participants about potential risks.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: Section 3.1 describes the use of Qwen3-VL in GTPO, and Section 4.4 describes GPT-4o as an auxiliary automatic judge for generated material descriptions.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.