VIDEO GENERATORS CAN PREDICT TOUCH: TACTILE SYNTHESIS FROM EGOCENTRIC VIDEO

Anonymous authors
Paper under double-blind review

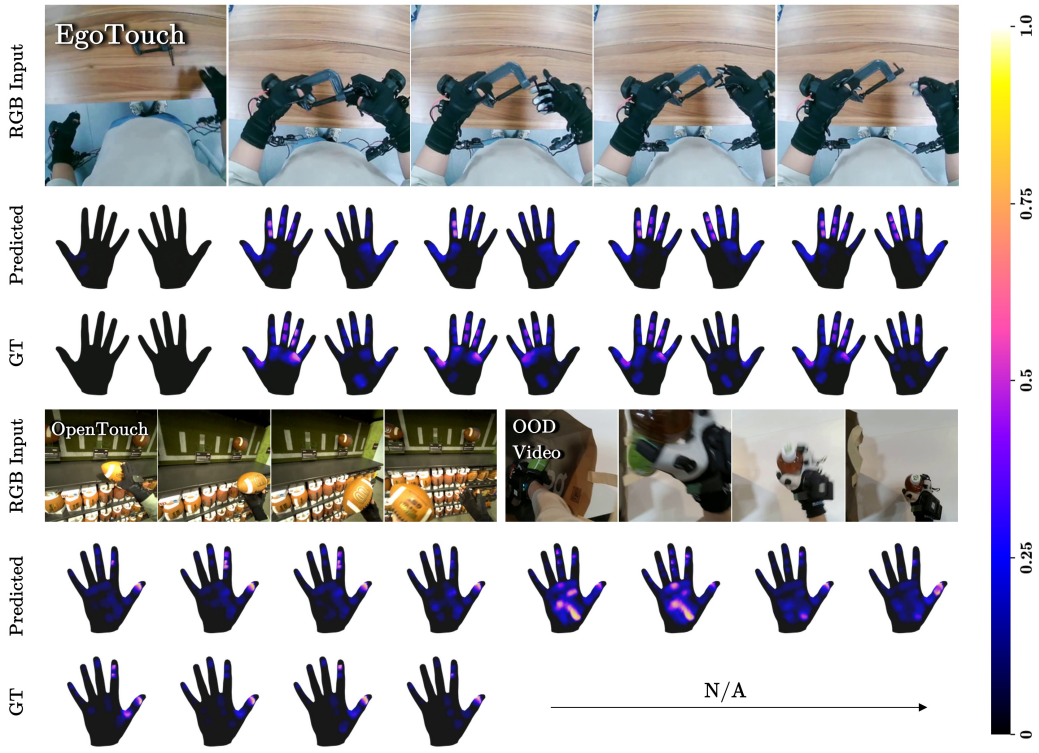


Figure 1: **Generating tactile video from egocentric video across diverse interaction settings.** Our model predicts temporally aligned tactile-pressure videos directly from egocentric RGB input. **Top:** bimanual EgoTouch interactions, where tactile pressure is generated independently for both hands. **Bottom left:** held-out OpenTouch examples showing single-hand tactile generation during object interaction. **Bottom right:** in-the-wild examples illustrating transfer to visually distinct scenes and interactions. Across these settings, the generated tactile signal captures the spatial distribution and temporal evolution of hand–object contact from RGB alone. Results can be viewed on our supplement website: https://iclr2027-15456.github.io/TouchPredict/

ABSTRACT

Tactile sensing matters for dexterous manipulation. However, tactile data is expensive and slow to collect at scale. In this work, we investigate how the interaction priors learned by video generators can be transferred to touch prediction from egocentric videos. We find that Wan2.1-Fun-Control can be easily repurposed for tactile pressure prediction via lightweight LoRA fine-tuning with its base weights frozen. On the OpenTouch test set, our model, using only video observations, achieves an MSE of 0.0104, hard IoU of 0.408, and soft IoU of 0.631 for tactile prediction. This setup also extends naturally to joint prediction of tactile pressure, hand configuration, and hand–object segmentation. We additionally evaluate tactile prediction on the bimanual EgoTouch dataset and qualitatively assess transfer to in-the-wild video clips without further fine-tuning. A randomly initialized generator fails to recover tactile structure in our control experiment, supporting the benefit of generative pretraining.

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

1 INTRODUCTION

"My hand is to me what your hearing and sight together are to you."
Helen Keller, *The World I Live In* (1908)

Touch matters for dexterous manipulation. While vision can reveal when and where contact occurs, touch provides direct feedback on the physical interaction between hands and objects. Indeed, for humans, touch is known to help regulate grip forces and respond to slip (Johansson & Westling, 1984; Johansson & Flanagan, 2009). However, unlike images or video, capturing representative, contact-rich touch data at scale remains a major bottleneck to applying machine-learning to touch. While videos on the internet capture diverse examples of physical interactions, they lack measurements of what the hands feel. To fill this gap, recent datasets such as OpenTouch (Song et al., 2025) make use of specialized sensors, such as egocentric video capture and haptic gloves, to record everyday interactions.

However, we also know that humans can learn to predict physical force from visual data. For example, surgeons using surgical robots routinely learn to estimate forces from visual signals (Chua et al., 2020). Can this be extended to more general manipulation? Specifically, can we estimate tactile signals from existing egocentric videos, potentially extending tactile annotations beyond recordings collected with specialized sensors? This is ill-posed: similar-looking grasps can produce different pressure distributions depending on grip force and object properties (Song et al., 2025), while contact regions are often occluded (Grady et al., 2022). Despite this ambiguity, we show the answer is yes: pretrained video generators can be adapted to predict tactile signals from egocentric video.

Our work builds on the observation that modern video generators are capable of producing highly realistic video clips involving physical interaction (DeepMind, 2025; OpenAI, 2024; Wan Team, 2025; NVIDIA, 2025). This suggests that pretrained video generators contain interaction priors that can help predict touch. Thus, we hypothesize that a pre-trained model that renders convincing hands, deforming objects and sliding contacts is a strong starting point to learn representations useful for inferring compliance, friction, and pressure. This is consistent with recent work showing that pretrained vision generators encode information about physical and geometrical scene properties, even without explicit supervision or with very limited supervision for these properties (Bhattad et al., 2023; Du et al., 2023; Bhattad et al., 2025; Hu et al., 2025; Zhang et al., 2026a; Bharadhwaj & Jangir, 2026).

Our recipe is straightforward: we fine-tune the Wan2.1-Fun-Control model with LoRA while freezing its base weights. For supervision, we use recorded tactile measurements paired with egocentric RGB videos. During inference, we supply just the RGB video and predict tactile pressure maps over time. Experiments on OpenTouch and EgoTouch demonstrate tactile generation in both single-hand and bimanual settings. On OpenTouch, a tactile-only model achieves an MSE of 0.0104 and soft IoU of 0.631. A control experiment comparing the pretrained model with the same architecture trained without generative pretraining performs substantially worse, demonstrating the importance of generative pretraining. We also explore generalization by qualitatively evaluating the fine-tuned model on unseen in-the-wild RGB videos. Finally, we extend the formulation to jointly predict hand configuration and hand–object segmentation by stacking the output latents. This extension produces all three outputs with a small reduction in tactile accuracy relative to tactile-only prediction.

In summary, our contributions are:

1. We show that a pretrained video generator can predict tactile pressure from egocentric RGB video through lightweight LoRA adaptation.
2. We evaluate tactile reconstruction, contact localization, and temporal agreement on OpenTouch and EgoTouch scenarios, quantify the benefit of generative pretraining, and qualitatively explore transfer to unseen in-the-wild videos.
3. We explore a simple extension to joint tactile, hand-configuration, and segmentation prediction through spatially stacked output latents, and evaluate its effect on tactile accuracy.

2 RELATED WORK

Tactile representation learning. Early work formulated tactile perception as recovering hardness and local surface geometry from tactile and kinesthetic measurements (Bajcsy & Hager, 1984).

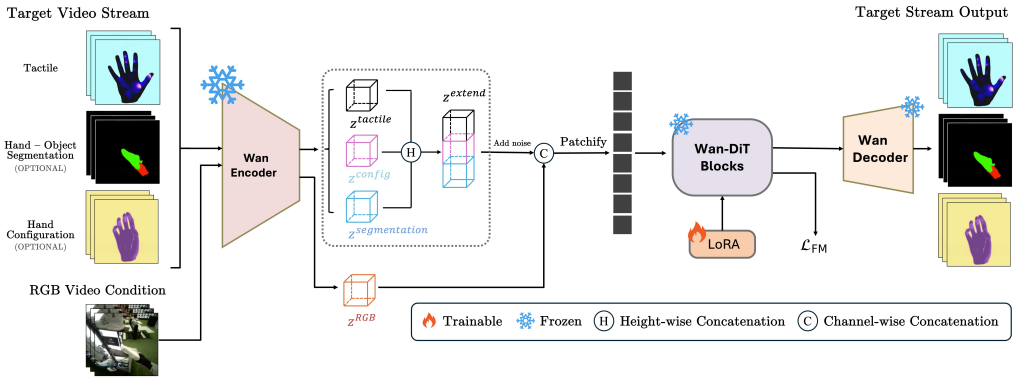


Figure 2: **Overview of the proposed multimodal generation pipeline.** Ground-truth tactile pressure, hand configuration, and hand/object segmentation videos are encoded with the frozen Wan VAE encoder and concatenated along the spatial height dimension to form a single target latent. Noise is added to the stacked latent, which is then channel-wise concatenated with the encoded egocentric RGB conditioning video and patchified before being passed through the pretrained Wan-DiT backbone. The backbone and VAE remain frozen, while a LoRA adapter is trained to jointly denoise the three target latents. The resulting latent is decoded and split back into tactile, configuration, and segmentation video streams. Snowflake and flame icons denote frozen and trainable components, respectively; H and C denote height-wise and channel-wise concatenation.

A large body of recent work learns general-purpose representations from tactile data for transfer across tasks and sensors. UniTouch (Yang et al., 2024) aligns tactile embeddings with a pretrained image embedding space, while Sparsh (Higuera et al., 2024) and AnyTouch (Feng et al., 2025) learn transferable representations from tactile observations. Other related methods learn to translate signals between tactile sensors (Rodriguez et al., 2025; Tu et al., 2026). These approaches focus on tactile representation and cross-sensor transfer, while we predict tactile pressure from egocentric RGB video from a diffusion model.

Vision-to-Tactile Prediction. Li et al. (2019) learned mappings between paired visual and tactile observations. Subsequent work applies vision-to-touch prediction to robotic manipulation (Wu et al., 2025). For human hands, PressureVision and PressureVision++ estimate pressure from RGB images (Grady et al., 2022; 2024), while HACO (Jung & Lee, 2026) estimates dense hand contact. EgoPressure (Zhao et al., 2025) explores joint hand-pose and pressure prediction, and TouchAnything (Zhou et al., 2026b) introduces EgoTouch for bimanual tactile estimation. Recent methods further study hand pressure and contact prediction from video (Jeon et al., 2026; Zhang et al., 2026b; Zhou et al., 2026a).

EgoTactile (Zeng et al., 2026) is particularly close to our approach. Its EgoPressureDiff model adapts pretrained Stable Video Diffusion to predict full-hand pressure, using spatial masks, physical-attribute text, and pressure prototypes as additional guidance. Our approach adapts Wan2.1-Fun-Control through LoRA while freezing its base weights, requires only RGB video at inference, and evaluates the benefit of generative pretraining for tactile prediction.

Repurposing video diffusion. Recent work adapts pretrained video generators to structured outputs beyond RGB, including depth estimation (Hu et al., 2025) and pose tracking (Zhang et al., 2026a). Other methods jointly generate video and depth (Zhai et al., 2024), or support multiple visual modalities such as depth and segmentation (Yang et al., 2025; Xi et al., 2026). Our work builds on these directions to predict tactile pressure from observed video. We also explore joint prediction of pressure, hand configuration, and segmentation by stacking their output latents.

3 Method

3.1 Problem Formulation

Given an egocentric RGB video $\mathbf{V}^{\text{rgb}}$, our goal is to jointly generate a set of spatially and temporally aligned hand–object interaction signals: tactile pressure $\hat{\mathbf{V}}^{\text{tac}}$, hand–object segmentation $\hat{\mathbf{V}}^{\text{seg}}$, and hand configuration video $\hat{\mathbf{V}}^{\text{config}}$. We formulate this task as conditional video generation and build on

the pretrained controllable video diffusion model Wan2.1-Fun-Control (Wan Team, 2025), trained with a flow-matching objective (Lipman et al., 2023). An overview is in Figure 2.

3.2 Multi-Stream Latent Construction

Let $\mathbf{V} = \{(\mathbf{V}^{\text{tac}}, \mathbf{V}^{\text{seg}}, \mathbf{V}^{\text{config}}, \mathbf{V}^{\text{rgb}})\}_{i=1}^{N}$denote the training set. We first normalize each target stream to $[0, 1]$ and render as a three-channel video so that it lies in the input domain of the pretrained VAE. We then encode each target stream, together with the RGB condition, using the frozen Wan VAE encoder $\mathcal{E}$:

$$\mathbf{z}^m = \mathcal{E}(\mathbf{V}^m) \in \mathbb{R}^{T' \times H' \times W' \times C}, \quad m \in \{\text{tac}, \text{seg}, \text{config}, \text{rgb}\}, \tag{1}$$

where $T' = 1 + F/4$, $H' = H/8$, and $W' = W/8$ follow the temporal and spatial compression ratios of the VAE. To model all target streams jointly with a single generative backbone, we stack their latents along the height axis to form one composite generation target:

$$\mathbf{z}^{\text{extend}} = \text{Concat}_H\big(\mathbf{z}^{\text{tac}}, \mathbf{z}^{\text{seg}}, \mathbf{z}^{\text{config}}\big) \in \mathbb{R}^{T' \times 3H' \times W' \times C}. \tag{2}$$

We fix the stream order during both training and inference. As shown in Chen et al. (2025), stacking inputs along the height dimension helps avoid modifying the model's underlying architecture, allowing it to seamlessly handle extended inputs without introducing any additional parameters. Also, spatial stacking enables full self-attention to facilitate information exchange both across modalities and between each individual modality and the RGB condition.

3.3 RGB-Conditioned Flow Matching

We sample a timestep $t \in [0, 1]$ and Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ to construct the noisy latent

$$\mathbf{z}_t^{\text{extend}} = (1 - t)\, \mathbf{z}^{\text{extend}} + t\, \boldsymbol{\epsilon}. \tag{3}$$

We follow Wan2.1-Fun-Control (Wan Team, 2025) by treating RGB video as the control signal by injecting through channel-wise concatenation. To match the new spatial extent of $\mathbf{z}_t^{\text{extend}}$, the RGB latent is replicated three times along the height axis, $\tilde{\mathbf{z}}^{\text{rgb}} = \text{Concat}_H(\mathbf{z}^{\text{rgb}}, \mathbf{z}^{\text{rgb}}, \mathbf{z}^{\text{rgb}})$, so that each target stream is paired with a pixel-aligned copy of the condition. The network input is

$$\mathbf{u}_t = \text{Concat}_C\big(\mathbf{z}_t^{\text{extend}}, \tilde{\mathbf{z}}^{\text{rgb}}\big) \in \mathbb{R}^{T' \times 3H' \times W' \times C'}. \tag{4}$$

This concatenated version is then patchified and processed by the Wan-DiT backbone. The model is then trained following the traditional latent flow matching loss:

$$\mathcal{L} = \mathbb{E}_{\mathbf{z}^{\text{extend}},\, t,\, \boldsymbol{\epsilon}} \left[\left\| \mathbf{v}_\theta(\mathbf{u}_t, t, \mathbf{c}) - (\boldsymbol{\epsilon} - \mathbf{z}^{\text{extend}}) \right\|_2^2 \right], \tag{5}$$

where $\mathbf{c}$ denotes the text embedding and we use an empty text during training this model. This single objective jointly denoises all three streams, without modality-specific output heads or additional losses. For efficient training, we keep the VAE and the 1.3B-parameter DiT backbone frozen and fine-tune only low-rank adapters (LoRA) (Hu et al., 2021) with rank 64. LoRA adapters are applied into the query, key, value, and output projections of every attention layer and into the feed-forward projections of each DiT block. At inference, we initialize the stacked latent with Gaussian noise $\hat{\mathbf{z}}_T^{\text{extend}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and iteratively denoise it. The resulting clean latent $\hat{\mathbf{z}}^{\text{extend}}$ is then split along the height axis into three stream latents $\{\hat{\mathbf{z}}^{\text{tac}}, \hat{\mathbf{z}}^{\text{seg}}, \hat{\mathbf{z}}^{\text{config}}\}$, and decode each with the VAE decoder to recover $\hat{\mathbf{V}}^{\text{tac}}$, $\hat{\mathbf{V}}^{\text{seg}}$, and $\hat{\mathbf{V}}^{\text{config}}$.

Dataset-specific stream configuration. The formulation above corresponds to the single-hand setting of OpenTouch (Song et al., 2025) whose target streams only concerns with one hand (right hand). EgoTouch (Zhou et al., 2026b) additionally provides left-hand tactile and configuration signals, so we extend Eq. 2 to five streams (adding one more stream for each video modality) which yields a stacked latent of height $5H'$. Correspondingly, the RGB latent is replicated five times and no other component of the model or objective is modified.

3.4 Data Pipeline

OpenTouch (Song et al., 2025) provides 2,900 right-hand clips (960×960, 30 fps). Each clip includes egocentric RGB video, raw 16×16 tactile pressure, and per-frame hand configuration represented using a fitted MANO (Romero et al., 2017) hand model. We use up to 81 frames per clip and

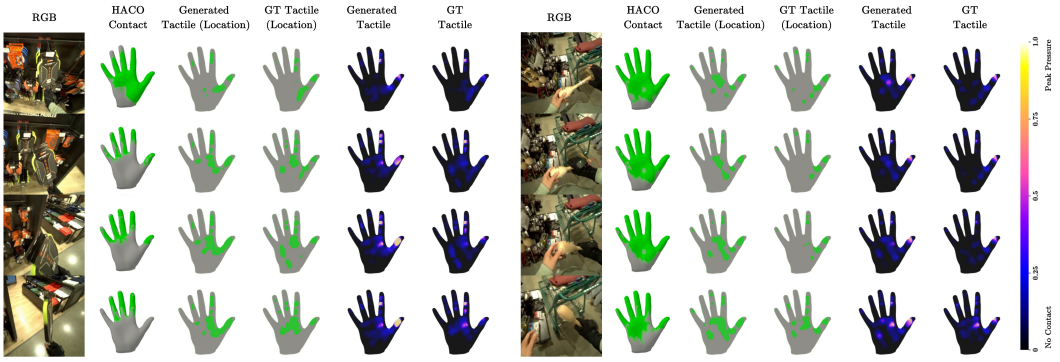


Figure 3: **Contact and tactile predictions on held-out OpenTouch data.** Representative examples compare the egocentric RGB input with contact estimates and tactile-pressure predictions. HACO (Jung & Lee, 2026) is used as an independent RGB-based contact baseline, while contact regions derived from the ground-truth and generated tactile signals provide corresponding tactile-based estimates. Green indicates predicted contact in the binary contact visualizations, while the final columns show the continuous ground-truth and generated tactile-pressure maps. HACO is used to assess spatial agreement with visually inferred hand–object contact.

retain 1,927 clips for training. Tactile pressure is rendered on a fixed and static hand mesh using the dataset's taxel-to-mesh mapping. We hold out 101 clips for testing.

EgoTouch (Zhou et al., 2026b) contains 1,891 two-hand clips. Similar to OpenTouch, each clip includes an egocentric RGB, raw 21×21 tactile pressure, and hand configuration provided by EgoTouch's Rokoko motion-capture glove. Longer videos allowed us to split sessions to 4,203 81-frame clips. 230 clips were used for held-out testing.

Hand–object segmentation. Neither dataset provides hand/object segmentation label, so we construct the segmentation masks for both datasets using the following pipeline: HACO (Jung & Lee, 2026) contact detection → HOI-DETR (Darkhalil et al., 2026) hand/object box proposals → SAM3 (Carion et al., 2026) segmentation. The resulting segmentation masks are used for the segmentation target stream for both OpenTouch and EgoTouch.

4 Experiments

4.1 Setup

We evaluate our framework on OpenTouch and EgoTouch using a held-out test set from each. The OpenTouch model jointly generates right-hand tactile pressure, hand–object segmentation, and hand configuration, while the EgoTouch model has two additional streams for left tactile and hand configuration. The models are both initialized from pretrained Wan2.1 weights and adapted using rank-64 LoRA, unless specified. Figure 3 shows a representative held-out OpenTouch example, showing the input RGB alongside the tactile prediction and contact estimates used throughout our evaluation.

4.2 Evaluation metrics

No single evaluation metric is sufficient to evaluate results. Reconstruction error can favor low-variance predictions whose means are close to the ground truth. Overlap metrics measure contact region accuracy but do not consider region intensity or temporal structure. We therefore evaluate pressure reconstruction, contact localization, and temporal consistency using complementary metrics.

Tactile reconstruction and contact. We evaluate tactile reconstruction using the masked mean-squared error (MSE) between generated and ground-truth tactile renderings. To distinguish between over- and under-prediction, we decompose the signed error into its positive and negative components. To assess alignment of spatial activity, we evaluate the hard and soft intersection-of-union (IoU). For hard IoU, each pixel in the generated tac-

Table 2: **Tactile generation performance on OpenTouch and EgoTouch.** We compare joint and tactile-only models for each dataset.

Configuration	MSE ↓	Over ↓	Under ↓	Hard IoU ↑	Soft IoU ↑	Optical Flow MSE ↓
Random initialization	0.0895	0.0982	0.0548	0.0670	0.1649	17.0767
OpenTouch (joint)	0.0146	0.0284	**0.0181**	0.3923	0.6123	0.4342
OpenTouch tactile-only	**0.0104**	**0.0215**	0.0190	**0.4081**	**0.6310**	**0.4294**
EgoTouch (joint)	0.0331	0.0486	**0.0270**	0.3275	0.4515	**0.1775**
EgoTouch tactile-only	**0.0111**	**0.0201**	0.0293	**0.3321**	**0.5419**	0.1868



Figure 4: **Bimanual tactile generation on held-out EgoTouch interactions.** Representative examples show the egocentric RGB observation together with predicted tactile maps for the left and right hands across several interactions. Unlike OpenTouch, EgoTouch contains bimanual contact, requiring the model to jointly predict tactile activity for both hands. Tactile magnitude is visualized using the shared color scale, ranging from no contact to peak pressure.

tile video is thresholded at 0.1 to obtain a binary contact location mask. Soft IoU considers pixel intensity and therefore assesses the spatial and tactile magnitude alignment.

Temporal consistency. We evaluate whether the generated tactile signal follows the ground-truth dynamics over time by computing optical-flow fields for the generated and ground-truth tactile videos and report the mean squared error.

Auxiliary streams. We evaluate auxiliary hand-configuration and segmentation streams using image-space MSE and optical-flow MSE, and report the mean semantic IoU for hand–object segmentation.

Table 1: **Auxiliary-output performance** for the OpenTouch multi-stream model.

Metric	Value
Hand configuration MSE ↓	0.0222
Hand configuration Flow MSE ↓	6.86
Segmentation MSE ↓	0.0436
Segmentation mean IoU ↑	0.474
Segmentation Optical Flow MSE ↓	6.11

4.3 Generation Performance

Table 2 reports the tactile reconstruction, spatial overlap and temporal consistency, while Table 1 separately evaluates the auxiliary hand-configuration and segmentation streams for the OpenTouch joint model. On OpenTouch, the tactile-only model improves MSE from 0.0146 to 0.0104, soft IoU from 0.6123 to 0.6310 and hard IoU from 0.3923 to 0.4081, while optical-flow error remains similar. Joint prediction reduces tactile prediction accuracy, while additionally predicting hand configuration and segmentation (Figure 5); auxiliary-stream performance is reported in Table 1.

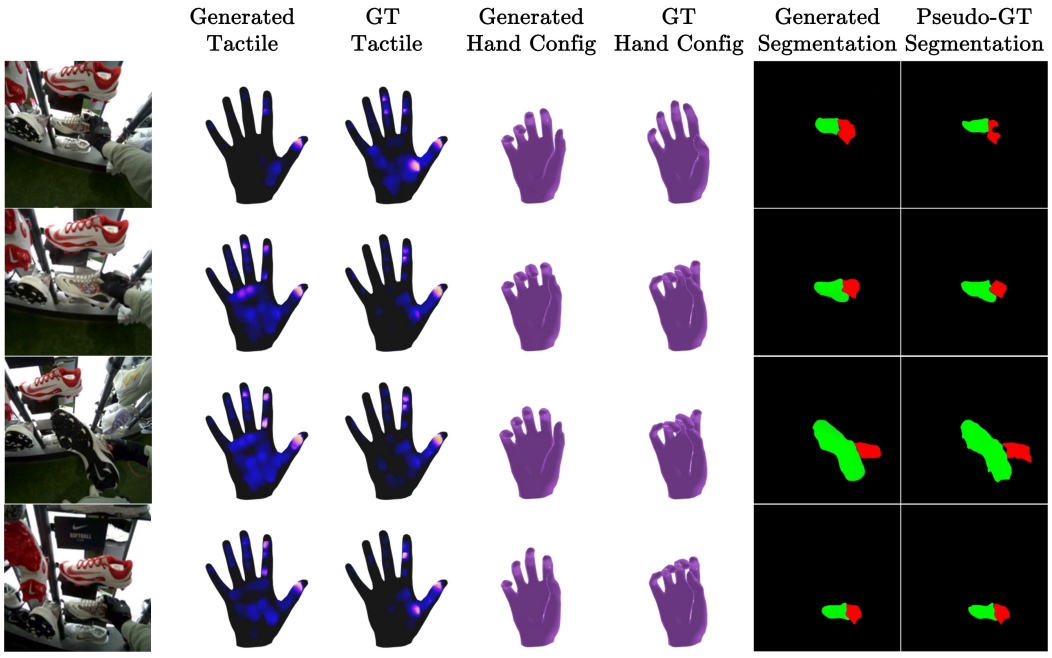


Figure 5: **Joint prediction of tactile pressure, hand configuration, and hand–object segmentation from egocentric RGB video.** Representative held-out OpenTouch frames are shown in the left column. For each timestep, the model jointly predicts three target streams: tactile pressure, hand configuration, and hand–object segmentation. The left panel shows the ground-truth targets and the right panel shows the corresponding generated outputs. The generated predictions largely preserve the spatial structure of contact, hand pose, and interaction layout across time, illustrating that the model can synthesize multiple correlated video outputs from RGB input alone.

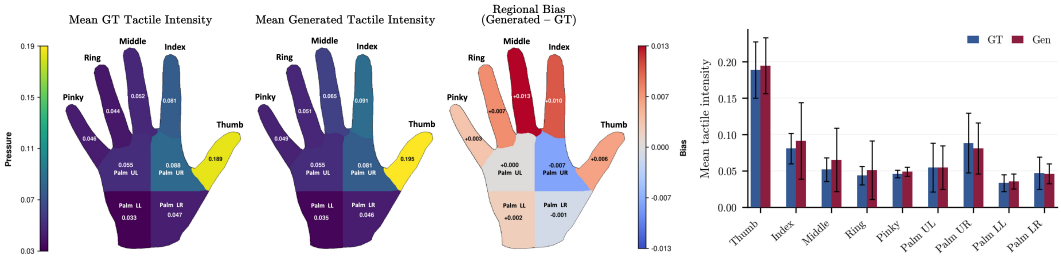


Figure 6: **Regional tactile activity on held-out OpenTouch sequences.** Tactile pressure is aggregated over nine predefined anatomical regions of the hand: the five fingers and four palm quadrants. From left to right, the hand maps show mean ground-truth tactile intensity, mean generated tactile intensity, and the regional difference between generated and ground-truth activity. The rightmost panel compares the mean tactile intensity of the generated and ground-truth signals across regions, with error bars showing the corresponding variability across the held-out set. The generated signal follows the overall anatomical activity profile of the ground truth, with a mild tendency toward overprediction across several regions.

We further examine how the generated tactile signal is distributed across regions of the hand. Figure 6 shows the mean tactile signal across nine labeled hand regions. The generated signal follows the overall ground-truth activity profile across the fingers and palm, and indicates a mild tendency toward higher activity across all regions. This is consistent with the higher overprediction bias reported in Table 2, where the mean overprediction error (0.0215) exceeds its underprediction error (0.0190). Figure 7 further shows that this error is not uniformly distributed across the hand, rather, it is spatially structured.

We additionally evaluate bimanual tactile-only generation on EgoTouch, and also observe stronger tactile predictions in the tactile-only model over the joint model. Figure 4 shows the model produc-

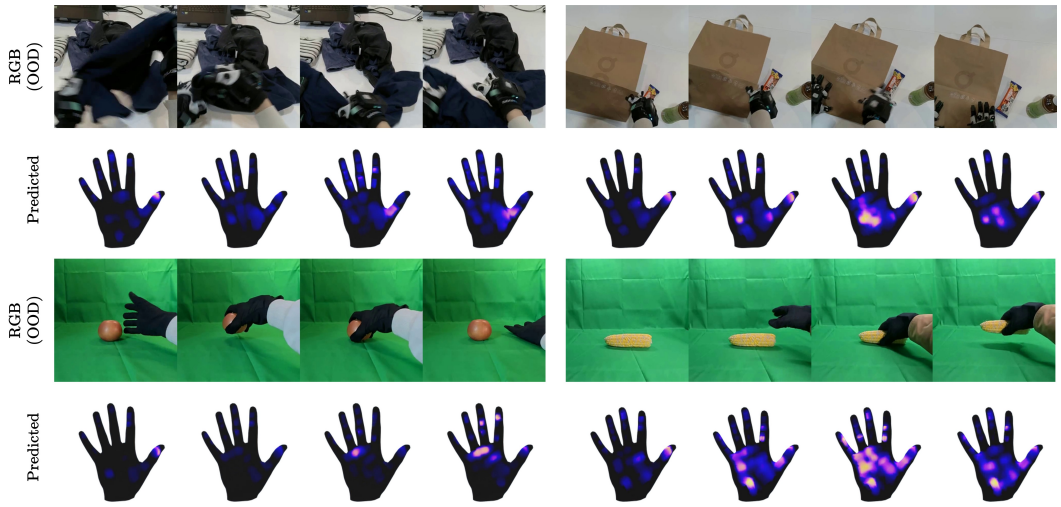


Figure 8: **Zero-shot transfer to out-of-distribution.** Given an out-of-distribution egocentric RGB input, our model generates tactile pressure without additional fine-tuning. Top row evaluates on the Manus Egocentric Hand Tracking (Labs, 2025), and bottom row on the EgoTactile dataset (Zeng et al., 2026).

ing left- and right-hand tactile predictions corresponding to the observed RGB. This demonstrates that the latent-stacking formulation can extend from single-hand to bimanual tactile generation.

4.4 Contact-Localization Benchmark

HACO (Jung & Lee, 2026) providesindependently trained RGB-based contact estimate for comparison. Figure 3 compares the contact regions obtained from HACO and contact extracted from the generated tactile signals. Quantitatively, generated tactile achieves a pixel-space IoU of 0.1715 with recorded contact, compared with 0.0433 for HACO (Table 3). We note that this comparison is not fully symmetric as our model has been trained on OpenTouch data while HACO is an independently trained model.

Table 3: **Pixel-space contact agreement on OpenTouch.**

Comparison	IoU ↑
HACO	0.0433
Gen	**0.1715**

4.5 Zero-Shot Cross-Dataset Transfer

We evaluate how well our model has learned to predict tactile signals beyond what it has been trained on by evaluating the tactile-only OpenTouch-trained model on the EgoTactile (Zeng et al., 2026) and the Manus Egocentric Hand Tracking (Labs, 2025) datasets, without additional finetuning. Figure 8 shows that the model continues to predict tactile signals across the two different datasets despite different scene settings. These datasets do not provide tactile ground-truth signals that are directly comparable to OpenTouch and EgoTouch, hence we treat these experiments as qualitative evidence of cross-domain transfer.

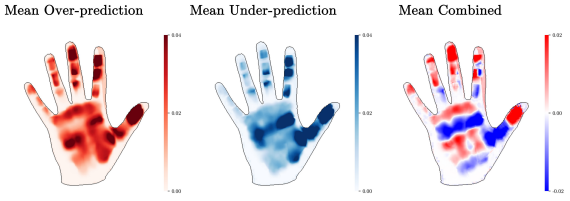
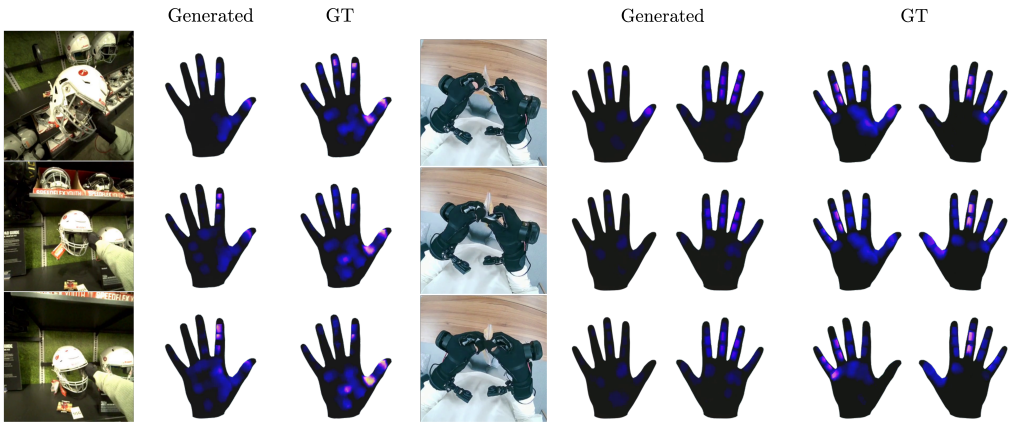


Figure 7: **Spatial tactile error on OpenTouch.** Mean overprediction, underprediction, and signed bias.

4.6 Qualitative Results

Qualitative examples in Figures 1,3–5 show that the generated pressure signals are generally coherent with hand–object contact, in both single-hand and bimanual interactions. Because tactile prediction from RGB is an ill-posed problem, disagreement with the recorded tactile signal does not necessarily imply an implausible prediction.



(a) OpenTouch failure case. (b) EgoTouch failure case.

Figure 9: **Representative failure cases on OpenTouch and EgoTouch.** Examples from **(a)** OpenTouch and **(b)** EgoTouch where the generated tactile activity diverges substantially from the recorded ground-truth tactile signal. These examples illustrate failure modes in both the single-hand and bimanual settings.

Figure 9 demonstrates two examples of failed predictions where the predicted pressure diverged both spatially and temporally from the ground-truth, and was not coherent with the observed hand–object interaction in the scene.

4.7 Ablation Studies

We evaluate the effect of generative pretraining and LoRA rank on OpenTouch, (quantitative results in Appendix C. We observe moderate gains as LoRA capacity increases. Randomly initialized backbone performs substantially worse than all pretrained configurations, failing to recover distinguishable tactile, segmentation, and hand-configuration structure despite reducing its training loss, emphasizing the contribution of pretrained prior in successfully predicting tactile signals.

5 Discussion

We presented a framework for predicting tactile signals from egocentric RGB video by adapting a frozen Wan2.1-Fun-Control backbone with LoRA. Our results on OpenTouch show that pretrained video generators can be repurposed to predict spatially and temporally structured tactile patterns. The same formulation also supports joint prediction of tactile pressure, hand configuration, and hand–object segmentation through spatially stacked output latents, although tactile-only prediction achieves slightly better tactile accuracy.

Predicting touch from video remains ill-posed: similar-looking interactions can involve different pressure distributions depending on grip force, object properties, and hidden contact geometry. Our predictions therefore estimate tactile signals from visible cues without uniquely determining the underlying contact forces. Evaluation against a single sensor recording measures agreement with the observed outcome, but does not capture other potentially valid solutions. Determining whether alternative predictions are physically plausible remains an open question.

Our study is also constrained by the scale and diversity of available egocentric tactile datasets, limiting the strength of cross-domain generalization claims. Representing tactile measurements as rendered videos makes them compatible with the pretrained backbone, but may lose sensor-specific structure in the raw data.

A longer-term goal is to extend tactile prediction beyond human hands to different robots and end effectors. This will require representations that account for differences in contact geometry, sensor layout, and the forces each embodiment can apply, together with paired data spanning these settings. Such a model could help annotate robot interaction videos with tactile estimates and support learning across embodiments.

AI Disclosure

In this work, we used generative AI tools to assist with method implementation, clean and reformat datasets and create visuals for the manuscript. We have not used generative AI tools for generating synthetic datasets and developing theoretical models. Formulating or proving mathematical claims, or assisting with translation are not applicable to this work. Additionally, we used generative AI tools to create and edit software code, summarize and analyze existing literature and identify related research topics and gaps. We have reviewed all AI-assisted code, analyses, and experimental suggestions. All generated code was tested before being used in reported experiments, and literature suggestions were checked against their original sources. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

Ethics statement

This work presents no ethical concerns.

Reproducibility statement

We provide the implementation and experimental details necessary to reproduce our training and evaluation pipeline.

- **Datasets and preprocessing.** We use publicly available OpenTouch and EgoTouch datasets. All rendering details are described in Appendix B and available through the datasets. We describe additional preprocessing steps in the Appendix.
- **Architecture and training.** Our framework builds on the publicly available Wan2.1-Fun-Control 1.3B video diffusion model. The pretrained VAE and DiT backbone are kept frozen, while LoRA adapters of rank 64 are optimized using a weighted rectified flow-matching objective. Training hyperparameters, including the optimizer, learning rate, precision, and other implementation settings, are provided in Appendix A.
- **Inference and evaluation.** We specify the inference procedure, including 50 denoising steps, input/output resolution, and sampling configuration. Evaluation protocols are described in Sections 3 and 4.

References

Ruzena Bajcsy and Gregory Hager. Tactile information processing-the bottom up approach. In *ICPR*, 1984. 2

Homanga Bharadhwaj and Yash Jangir. Motionforesight: Re-purposing video models for future 3d scene-flow prediction. 2026. 2

Anand Bhattad, Daniel McKee, Derek Hoiem, and D. A. Forsyth. StyleGAN knows normal, depth, albedo, and more. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2

Anand Bhattad, Konpat Preechakul, and Alexei A Efros. Visual jenga: Discovering object dependencies via counterfactual inpainting. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=yMXn86pzWx. 2

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *International conference on learning representations*, volume 2026, pp. 138846–138923, 2026. 5, 14

Zhaoxi Chen, Tianqi Liu, Long Zhuo, Jiawei Ren, Zeng Tao, He Zhu, Fangzhou Hong, Liang Pan, and Ziwei Liu. 4dnex: Feed-forward 4d generative modeling made easy. *arXiv preprint arXiv:2508.13154*, 2025. 4

Zonghe Chua, Anthony M Jarc, Sherry M Wren, Ilana Nisky, and Allison M Okamura. Task dynamics of prior training influence visual force estimation ability during teleoperation. *IEEE Transactions on Medical Robotics and Bionics*, 2(4):586–597, 2020. 2

Ahmad Darkhalil, Dima Damen, and David Fouhey. Improving and evaluating hand-object interaction detection. *arXiv preprint arXiv:2606.17384*, 2026. 5, 14

Google DeepMind. Veo 3 technical report. Technical report, Google, 2025. URL https://storage.googleapis.com/deepmind-media/veo/Veo-3-Tech-Report.pdf. 2

Xiaodan Du, Nicholas Kolkin, Greg Shakhnarovich, and Anand Bhattad. Generative models: What do they know? do they know things? let's find out! *arXiv preprint arXiv:2311.17137*, 2023. 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. In *International Conference on Learning Representations*, volume 2025, pp. 31265–31285, 2025. 3

Patrick Grady, Chengcheng Tang, Samarth Brahmbhatt, Christopher D Twigg, Chengde Wan, James Hays, and Charles C Kemp. Pressurevision: estimating hand pressure from a single rgb image. In *ECCV*, 2022. 2, 3

Patrick Grady, Jeremy A Collins, Chengcheng Tang, Christopher D Twigg, Kunal Aneja, James Hays, and Charles C Kemp. Pressurevision++: Estimating fingertip pressure from diverse rgb images. In *WACV*, 2024. 3

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. 3

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 4

Wenbo Hu, Xiangjun Gao, Xiaoyu Li, Sijie Zhao, Xiaodong Cun, Yong Zhang, Long Quan, and Ying Shan. Depthcrafter: Generating consistent long depth sequences for open-world videos. In *CVPR*, 2025. 2, 3

Subin Jeon, Byungjun Kim, and Hanbyul Joo. Hope: Hand-object pressure estimation from monocular videos. *arXiv preprint arXiv:2608.06192*, 2026. 3

Roland S. Johansson and J. Randall Flanagan. Coding and use of tactile signals from the fingertips in object manipulation tasks. *Nature Reviews Neuroscience*, 10(5):345–359, 2009. 2

Roland S. Johansson and Göran Westling. Roles of glabrous skin receptors and sensorimotor memory in automatic control of precision grip when lifting rougher or more slippery objects. *Experimental Brain Research*, 56(3):550–564, 1984. 2

Daniel Jung and Kyoung Mu Lee. Learning dense hand contact estimation from imbalanced data. *Advances in Neural Information Processing Systems*, 38:120351–120384, 2026. 3, 5, 8, 14

OpenGraph Labs. Manus egocentric hand tracking dataset, 2025. URL https://huggingface.co/datasets/opengraph-labs/manus-egocentric-sample. 8

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)*, pp. 10601–10610. IEEE, 2019. 3

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2023. 4

NVIDIA. Cosmos world foundation model platform for physical AI. *arXiv preprint arXiv:2501.03575*, 2025. 2

OpenAI. Video generation models as world simulators. https://openai.com/research/video-generation-models-as-world-simulators, 2024. 2

Samanta Rodriguez, Yiming Dou, Miquel Oller, Andrew Owens, and Nima Fazeli. Cross-sensor touch generation. *arXiv preprint arXiv:2510.09817*, 2025. 3

Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics*, 36(6):245:1–245:17, 2017. doi: 10.1145/3130800.3130883. 4

Yuxin Ray Song, Jinzhou Li, Rao Fu, Devin Murphy, Kaichen Zhou, Rishi Shiv, Yaqi Li, Haoyu Xiong, Crystal Elaine Owens, Yilun Du, et al. Opentouch: Bringing full-hand touch to real-world interaction. *arXiv preprint arXiv:2512.16842*, 2025. 2, 4, 14

Jiahang Tu, Fengyu Yang, Chenyang Ma, Xihang Yu, Ziyao Zeng, Shaokai Wu, Hanbin Zhao, Zhi Tao, Chao Zhang, Hui Qian, et al. Unitac: A unified multimodal model for cross-sensor tactile understanding and generation. *arXiv preprint arXiv:2606.31451*, 2026. 3

Wan Team. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 2, 4

Zhiyuan Wu, Yijiong Lin, Yongqiang Zhao, Xuyang Zhang, Zhuo Chen, Nathan Lepora, and Shan Luo. Vitacgen: Robotic pushing with vision-to-touch generation. *IEEE Robotics and Automation Letters*, 2025. 3

Dianbing Xi, Jiepeng Wang, Yuanzhi Liang, Xi Qiu, Yuchi Huo, Rui Wang, Chi Zhang, and Xuelong Li. Omnivdiff: Omni controllable video diffusion for generation and understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 10915–10923, 2026. 3

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26330–26343. IEEE, 2024. 3

Lehan Yang, Lu Qi, Xiangtai Li, Sheng Li, Varun Jampani, and Ming-Hsuan Yang. Unified dense prediction of video diffusion. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28963–28973. IEEE, 2025. 3

Yuan Zeng, Yujia Shi, Tiao Tan, Xingting Li, Yaqi Qin, Zongqing Lu, Wenming Yang, Jing-Hao Xue, and Qingmin Liao. Egotactile: Learning grasp pressure for everyday objects from egocentric video. *arXiv preprint arXiv:2606.09243*, 2026. 3, 8

Yuanhao Zhai, Kevin Lin, Linjie Li, Chung-Ching Lin, Jianfeng Wang, Zhengyuan Yang, David Doermann, Junsong Yuan, Zicheng Liu, and Lijuan Wang. IDOL: Unified dual-modal latent diffusion for human-centric joint video-depth generation. In *European Conference on Computer Vision (ECCV)*, 2024. 3

Ruihang Zhang, Felix Taubner, Pooja Ravi, Kiriakos N Kutulakos, and David B Lindell. Proxypose: 6-dof pose tracking via video-to-video translation. *arXiv preprint arXiv:2607.06555*, 2026a. 2, 3

Wenkang Zhang, Chengbo Yuan, Zicheng Zhang, Zhengxue Cheng, and Yang Gao. Egotac: In-the-wild tactile prediction from egocentric vision. *arXiv preprint arXiv:2608.15060*, 2026b. 3

Yiming Zhao, Taein Kwon, Paul Streli, Marc Pollefeys, and Christian Holz. Egopressure: A dataset for hand pressure and pose estimation in egocentric vision. In *CVPR*, 2025. 3

Danyan Zhou, Jinxuan Lu, Jiawei Lin, Tianxing Chen, Chuqiao Lyu, and Wenbo Ding. Touchsight: Bare-handed tactile prediction from egocentric video via generative visual augmentation. *arXiv preprint arXiv:2609.20414*, 2026a. 3

Jianyi Zhou, Ziteng Gao, Feiyang Hong, Zirui Liu, Guannan Zhang, Weisheng Dai, Ruichen Zhen, Chuqiao Lyu, Haotian Wu, Yinian Mao, et al. Touchanything: A dataset and framework for bimanual tactile estimation from egocentric video. *arXiv preprint arXiv:2605.13083*, 2026b. 3, 4, 5, 14

Table 4: **Implementation details for the primary models.** Settings are shared across OpenTouch and EgoTouch unless otherwise noted.

Setting	Value
Backbone	Wan2.1-Fun-Control 1.3B
Video autoencoder	Wan VAE, frozen
Diffusion backbone	Wan-DiT, base weights frozen
Adaptation	LoRA
LoRA rank	64
LoRA target modules	q, k, v, o, `ffn.0`, `ffn.2`
Optimizer	AdamW
Learning rate	1×10^{-4}
LR warmup	10 steps
Training precision	bfloat16
Clip length	Up to 81 frames
Training objective	Weighted rectified-flow matching
Prediction target	Velocity, $v = \epsilon - x_0$

A Implementation Details

A.1 Model and Optimization

All experiments ran in this study use the Wan2.1-Fun-Control 1.3B video diffusion model as the generative backbone. During training, the pretrained video VAE and base WAN-DiT parameters remain frozen, and low-rank adaptation (LoRA) is used to adapt the model. LoRA adapters are applied to the attention adn feed-forward projections of each DiT block, and we use rank-64 adapters for experiments, unless otherwise specified. The aim is to take advantage of the pretrained priors and adapt them to tactile and auxiliary video generation without the need to update the full backbone.

We encode each target modality independently using the frozen Wan VAE, and the resulting latent videos are spatially concatenated along their height, giving a single target latent. The RGB video remains noise-free and is used as a conditioning signal during training. Flow-matching noise is added to the target latent and concatenated along the channel with the clean RGB video, and passed through the Wan DiT. The number of target videos differs between the OpenTouch and EgoTouch Datasets, but the same formulation is used for both.

Table 4 summarized the implementation details and settings throughout our experiment.

A.2 Training Configurations

The different model configurations used throughout the experiments is summarized in Table 5. The joint Opentouch model generates three streams: right-hand tactile, hand–object segmentation, left-hand configuration. We use this model as the primary model throughout this paper and all ablations are done using it. Joint EgoTouch extends this same formulation to five streams, adding left-hand tactile and configuration. We also train both datasets on tactile-only to compare generation results and analyze whether the additional streams improve tactile predictions.

A.3 Inference

the model receives the egocentric RGB video at inference, encodes it using the Wan VAE and channel-concatenates it with noisy target latent. Starting from Gaussian noise, the finetuned Wan-DiT iteratively denoises the target latent. For the joint models, the latent is split at the predefined spatial locations and each component is then decoded using the Wan VAE, giving us the output videos. Inference uses 50 denoising steps under the flow-matching schedule (σ shift 5.0), with a denoising strength of 1.0 and no classifier-free guidance (CFG scale 1.0). All videos are generated at 480×480 resolution, 81 frames, in bfloat16 precision.

Table 5: **Primary training configurations.** Joint models jointly generate tactile and auxiliary videos, while single-modal models generate tactile only.

Dataset	Configuration	Generated Videos	Initialization	LoRA rank
OpenTouch	Joint	tactile_R, segmentation, config_R	Pretrained	64
	Tactile-only	tactile_R	Pretrained	64
EgoTouch	Joint	tactile_R, tactile_L, segmentation, config_R, config_L	Pretrained	64
	Tactile-only	tactile_R, tactile_L	Pretrained	64

B Dataset and Target Construction

We train our model on OpenTouch (2025) and EgoTouch (Zhou et al., 2026b), which provide synchronized egocentric RGB video, tactile measurements, and hand configuration data. However, the datasets differ in the layout of the tactile-sensor and the hand configuration representation. We convert the tactile and hand-configuration data into video representation. All videos are then truncated to the maximum length of 81 frames.

B.1 OpenTouch Target Construction

The raw tactile data in the OpenTouch dataset is represented as a 16×16 taxel grid. We convert these measurements to a tactile video by mapping the taxels onto a fixed hand mesh, using the provided taxel-to-mesh correspondence code provided by the dataset. OpenTouch additionally provides fitted MANO hand configuration for each frame, and we render it accordingly to get a hand-configuration video. We retain 1,927 clips for training

B.2 EgoTouch Target Construction

EgoTouch raw tactile data differs from OpenTouch in its dimension (21×21) and the tactile video rendering follows the dataset's taxel-to-mesh correspondence. Hand configuration is provided by the dataset through the Rokoku motion-capture representation. EgoTouch recordings are substantially longer than the fixed-length clips used for training, allowing us to extract multiple 81-frame length videos from each recording session. We retain clips where the Rokoko hand-configuration data is present throughout the recording. 4,203 training clips are retained the processed dataset.

B.3 Hand-Object Segmentation Target

As hand–object segmentation is not provided by either datasets, we generate these targets from the corresponding RGB videos. Figure 10 demonstrates our segmentation pipeline combines HACO(Jung & Lee, 2026), HOI-DETR (Darkhalil et al., 2026), and SAM3 (Carion et al., 2026). HACO's desnse-per-vertiex hand-contact estimator runs across each clip and identifies all frames whos estimates contact-vertex count crosses a fixed threshold. Clips where none of the frames pass the threshold are discarded from the dataset. At the second stage, HOI-DETR runs across all proposed candidate frames and selects the one which yields the highest-confidence hand–object interaction detection, and producing bounding boxes for both the hand and object. In the final stage, each box is used as a prompt for SAM3's box-conditioned video segmentation. SAM3 propagates forward and backward from the chosen frame. This is done once for the object and again for the hand segmentation. The resulting masks are then combined to a single two-color mask where the object segmentation is green and the hand is red.

C Additional Ablations

Figure 11 show results of LoRA rank ablation and joint prediction versus tactile only ablations. Table 6 quantifies these results.

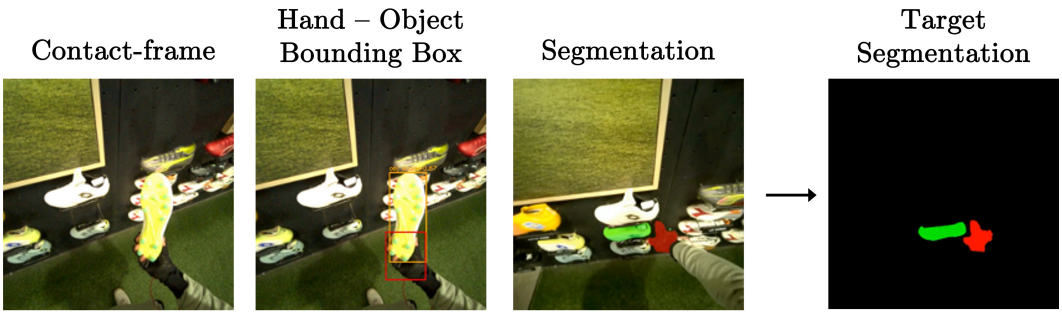


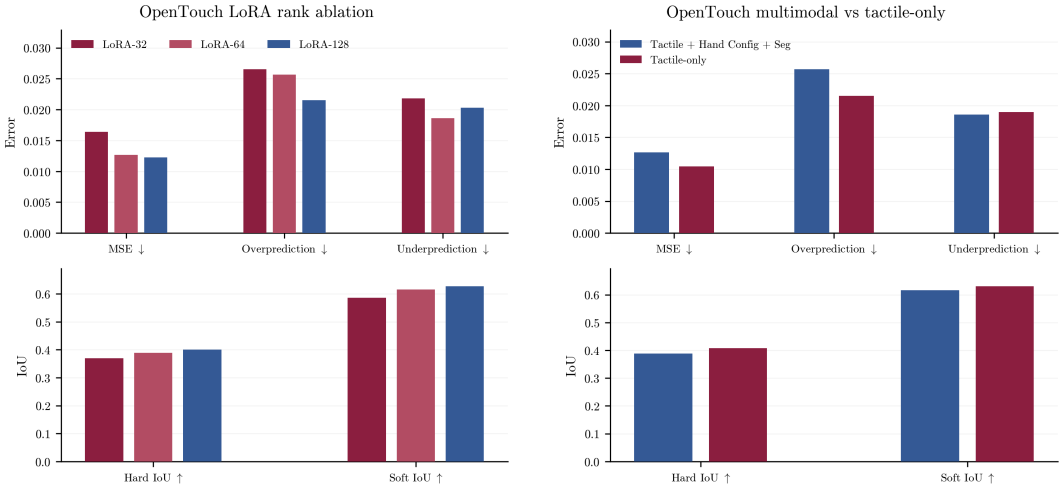
Figure 10: **Hand–object segmentation target construction.** We generate segmentation supervision offline from the egocentric RGB video using a three-stage pipeline. HACO first identifies contact-relevant hand regions, HOI-DETR predicts the interacting hand–object bounding boxes, and SAM3 produces pixel-level hand and object masks. The resulting masks are converted into the fixed semantic segmentation target used for training, with hand and object represented as separate classes.

Table 6: **Ablation of generative pretraining and LoRA rank on OpenTouch.** All models are evaluated on the same 101 held-out OpenTouch clips.

Configuration	MSE ↓	Over ↓	Under ↓	Hard IoU ↑	Soft IoU ↑	Flow MSE ↓
LoRA rank 32	0.0164	0.0265	0.0218	0.3699	0.5855	0.4224
LoRA rank 64	0.0126	0.0257	**0.0186**	0.3882	0.6160	0.4232
LoRA rank 128	**0.0122**	**0.0215**	0.0203	**0.4009**	**0.6275**	**0.4190**
Random initialization	0.0895	0.0982	0.0548	0.0670	0.1649	17.0767

D Additional Qualitative Results

We provide additional qualitative examples in Figures 12 and 13.



(a) **LoRA rank ablation.** Comparison of LoRA ranks 32, 64, and 128 on OpenTouch. Rank 128 achieves the lowest MSE and overprediction error and the highest hard and soft IoU, while rank 64 yields the lowest underprediction error.

(b) **Joint multimodal versus tactile-only training.** We compare the *Tactile + Hand Config + Seg* model with the tactile-only model on OpenTouch. Tactile-only achieves lower MSE and overprediction error and higher hard and soft IoU, while joint multimodal training yields slightly lower underprediction error.

Figure 11: **OpenTouch ablations.** Reconstruction errors are shown in the top panels (↓), while contact overlap metrics are shown in the bottom panels (↑).

Table 7: **Regional tactile pressure statistics on OpenTouch.** Ground-truth (GT) and generated (Gen.) values are reported as mean ± standard deviation across clips. Bias is computed as generated minus ground-truth pressure, such that positive values indicate overprediction and negative values indicate underprediction.

Region	GT	Gen.	Bias
Thumb	0.189 ± 0.039	0.195 ± 0.038	+0.006
Index	0.081 ± 0.021	0.091 ± 0.053	+0.010
Middle	0.052 ± 0.016	0.065 ± 0.044	+0.013
Ring	0.044 ± 0.013	0.051 ± 0.040	+0.007
Pinky	0.046 ± 0.005	0.049 ± 0.006	+0.003
Palm UL	0.055 ± 0.034	0.055 ± 0.030	+0.000
Palm UR	0.088 ± 0.041	0.081 ± 0.035	-0.007
Palm LL	0.033 ± 0.012	0.035 ± 0.010	+0.002
Palm LR	0.047 ± 0.022	0.046 ± 0.014	-0.001

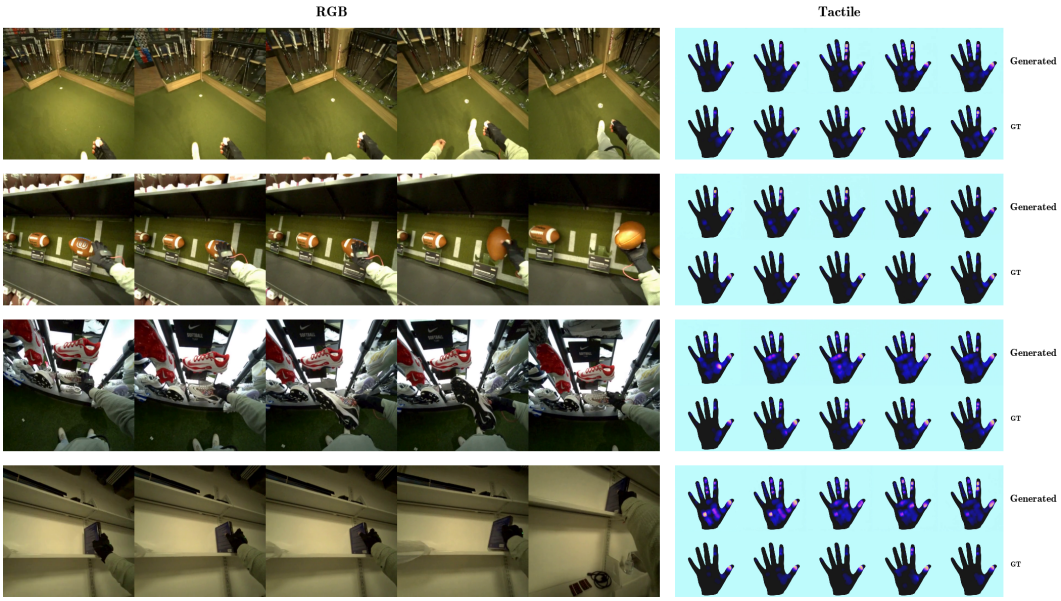


Figure 12: **Additional qualitative tactile generation for OpenTouch.**

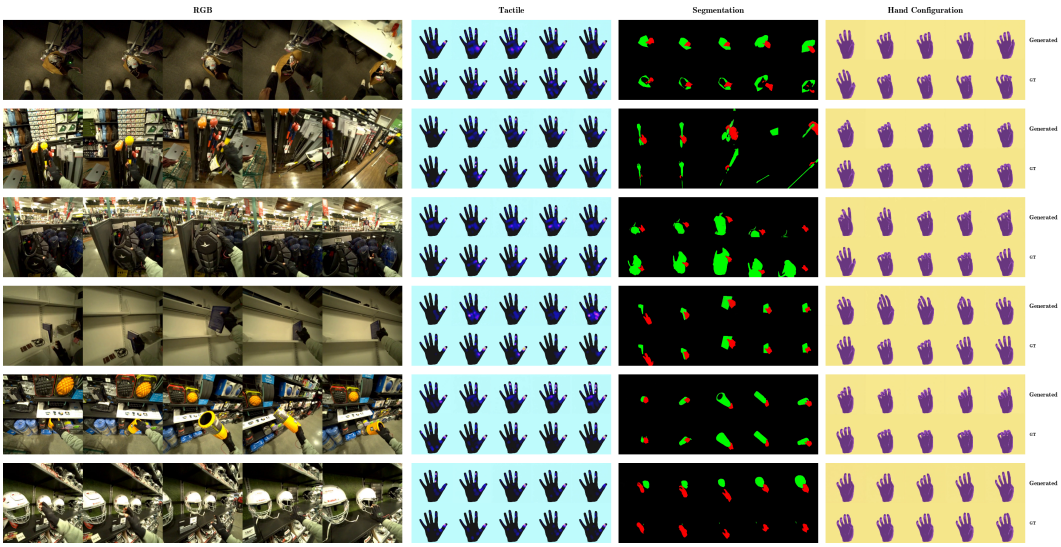
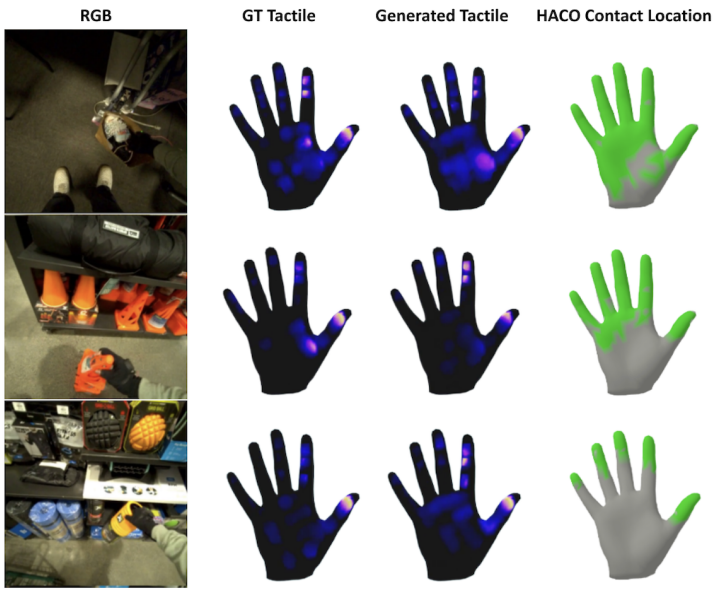


Figure 13: **Additional qualitative results for predicted videos streams on OpenTouch.**



(a) Example Contact Predictions

Figure 14: **Contact-localization agreement with HACO. (a)** Representative held-out OpenTouch examples comparing RGB input, ground-truth tactile, generated tactile, and HACO contact predictions.