Learning Intuitive Physics with Multimodal Generative Models

Sahand Rezaei-Shoshtari,[1,2] **Francois R. Hogan,**[1] **Michael Jenkin,**[1,3]
David Meger,[1,2] **Gregory Dudek**[1,2]

[1] Samsung AI Center Montreal
[2] McGill University
[3] York University

srezaei@cim.mcgill.ca, {f.hogan, greg.dudek}@samsung.com, {m.jenkin, david.meger}@partner.samsung.com

Abstract

Predicting the future interaction of objects when they come into contact with their environment is key for autonomous agents to take intelligent and anticipatory actions. This paper presents a perception framework that fuses visual and tactile feedback to make predictions about the expected motion of objects in dynamic scenes. Visual information captures object properties such as 3D shape and location, while tactile information provides critical cues about interaction forces and resulting object motion when it makes contact with the environment. Utilizing a novel See-Through-your-Skin (STS) sensor that provides high resolution multimodal sensing of contact surfaces, our system captures both the visual appearance and the tactile properties of objects. We interpret the dual stream signals from the sensor using a Multimodal Variational Autoencoder (MVAE), allowing us to capture both modalities of contacting objects and to develop a mapping from visual to tactile interaction and vice-versa. Additionally, the perceptual system can be used to infer the outcome of future physical interactions, which we validate through simulated and real-world experiments in which the resting state of an object is predicted from given initial conditions.

Introduction

Recently, several authors have pointed out the synergies between the senses of touch and vision: one enables direct measurement of 3D surface and inertial properties, while the other provides a holistic view of the projected appearance. Methods such as Li et al. (2019) have trained joint perceptual components, allowing better inference of physical properties from images, for example. This paper extends this reasoning into dynamic prediction: how can we predict the future motion of an object from visual and tactile measurements of its initial state? If a previously unseen object is dropped into a human's hand, we are able to infer the object's category and guess at some of its physical properties, but the most immediate inference is whether it will come to rest safely in our palm, or if we need to adjust our grasp on the object to maintain contact. Vision allows rapid indexing to capture overall object properties, while the tactile signal at the point of contact fills in a crucial gap, allowing direct physical reasoning about balance, contact forces
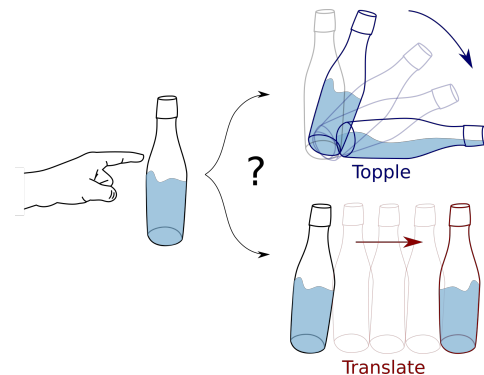
Figure 1: Predicting the outcome of physical interactions. Given an external perturbation on a bottle, how can we predict if the bottle will topple or translate? This paper reasons across visual and tactile modalities to infer the motion of objects in dynamic scenes.

and slippage. This paper shows that the combination of these signals is ideal to predict the object motion that will result in dynamic scenarios. Namely, we predict the final stable outcome of passive physical dynamics on objects based on sensing their initial state with touch and vision.

Previous research has shown that it is challenging to predict the trajectory of objects in motion, due to the unknown frictional and geometric properties and indeterminate pressure distributions at the interacting surface (Fazeli et al. 2020). To alleviate these difficulties, we focus on learning a predictor trained to capture the most informative and stable elements of a motion trajectory. This is in part inspired by recent findings in Time-Agnostic Prediction (Jayaraman et al. 2018), where the authors show that the prediction accuracy and reliability of predictive models can be vastly improved by focusing on outcomes at key events in the future. Furthermore, this approach mitigates the error drift from which visual prediction typically suffer; where uncertainties and errors propagate forward in time to produce blurry and imprecise predictions. For example, in Fig. 1, when predicting the outcome of an applied push on a bottle, an agent should reason about the most important consequence of this action: will the bottle topple over or will it translate forward? To study this problem, we present a novel artificial perception
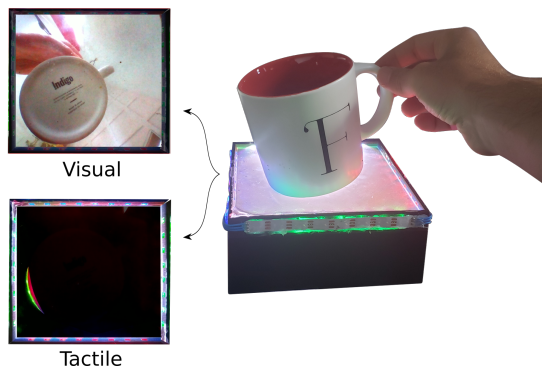
Figure 2: The visuotactile multimodal output of the See-Through-your-Skin (STS) sensor. Using controlled internal illumination, the surface of the sensor can be made to be transparent, as shown in the top left, allowing the camera to view the outside world. In the lower left figure, the sensor provides a tactile signature by maintaining the interior of the sensor bright relative to the outside.

system, composed of both hardware and software contributions, that allows measurement and prediction of the final resting configuration of objects falling on a surface. We prototype a novel sensor able to simultaneously capture visual images and provides tactile measurements. The data from this new *See-Through-your-Skin* (STS) sensor is interpreted by a multimodal perception system inspired by the multimodal variational autoencoder (MVAE) (Wu and Goodman 2018). The main contributions of this paper are:

Generative Multimodal Perception that integrates visual and tactile feedback within a unified framework. The system, built on Multimodal Variational Autoencoders (MVAE), exploits visual and/or tactile feedback when available, and learns a shared latent representation that encodes information about object pose, shape, and force to make predictions about object dynamics. To the best of our knowledge, this is the first study that learns multimodal dynamics models using collocated visual and tactile perception.

Resting State Predictions that predicts the resting state of an object during physical interactions. We show that this approach is able to learn generalizable dynamics model capable of learning physical scenarios without explicit consideration of a full physical model.

Visuotactile Dataset of object motions in dynamic scenes. We consider three scenarios: objects freefalling on a flat surface, objects sliding down an inclined plane, and objects perturbed from their resting pose. Our code is made publicly available [1].

Related Work

Optical Tactile Sensors utilize visual information to encode touch-based interactions (Shimonomura 2019; Abad and Ranasinghe 2020). Using a combination of a camera and a light source to capture distortions in the contact surface introduced by tactile interactions, optical sensors such

as GelForce (Vlack et al. 2005), Gelsight (Johnson and Adelson 2009), Omnitact (Padmanabha et al. 2020), DIGIT (Lambeta et al. 2020), GelSlim (Donlon et al. 2018), and Soft-Bubble (Kuppuswamy et al. 2020) render a high resolution image of the contact geometry. These optical tactile sensors make use of an opaque membrane that obscures the view of the object at the critical moment prior to contact. This research builds on recent developments to optical sensors enabling them to simultaneously render both tactile and visual feedback from a same point of reference, such as the Semi-transparent Tactile Sensor (Hogan et al. 2021) and FingerVision (Yamaguchi and Atkeson 2017).

Multimodal Learning of Vision and Touch have been shown to produce rich perceptual systems that exploit the individual strengths of each modality. Multimodal approaches have been proposed for 3D shape reconstruction (Allen 1984; Bierbaum, Gubarev, and Dillmann 2008; Ottenhaus et al. 2016; Yi et al. 2016; Driess, Englert, and Toussaint 2017; Luo et al. 2016; Wang et al. 2018; Luo et al. 2018; Smith et al. 2020), pose estimation (Izatt et al. 2017), robotic manipulation (Li et al. 2014; Calandra et al. 2017, 2018; Lee et al. 2019; Watkins-Valls, Varley, and Allen 2019), and dynamics modeling (Tremblay et al. 2020). The connections between both modalities have been investigated in (Li et al. 2019; Lee, Bollegala, and Luo 2019) using GANs and concatenation of embedding vectors from different sensing modes (Yuan et al. 2017; Lee et al. 2019). The perceptual system proposed in this research maps vision and touch to a shared latent space in an end-to-end framework, resulting in a robust predictive model capable of handling missing modalities. We focus on learning dynamic models involving objects interacting with their environment rather than explicitly studying the connection between the two modes.

Temporal Abstraction in planning and prediction can mitigate the challenges in long-horizon tasks (Sutton, Precup, and Singh 1999) by temporally breaking down the trajectories into shorter segments. This can be achieved through the discovery of prediction bottlenecks, shown to be effective at generating intermediate sub-goals (McGovern and Barto 2001; Bacon, Harb, and Precup 2017) used by agents in a hierarchical reinforcement learning paradigm (Nair and Finn 2019; Nair, Savarese, and Finn 2020; Pertsch et al. 2020). Recent work on Time-Agnostic Prediction has shown that the notion of predictability can be exploited to identify bottlenecks (Jayaraman et al. 2018; Neitz et al. 2018), by skipping frames with high prediction uncertainty and focusing on frames that are more stable and easier to predict. This paper draws inspiration from this work by defining the state bottlenecks as stable object configurations during dynamic interactions and show that this assumption allows to learn more accurate and robust dynamics models.

Approach

This section outlines our approach for learning intuitive physical models that reason across visual and tactile sensing. Our long term objectives are twofold. First, we aim to understand how to develop reliable sensory perceptual models that integrate the senses of touch and sight to make infer-

---
[1] https://github.com/SAIC-MONTREAL/multimodal-dynamics

ences about the physical world. Second, we seek to exploit these models to enable autonomous agents to interact with the physical world. This paper focuses on the former, by investigating the core capability of a visuotactile sensor to make predictions about the evolution of dynamic scenarios.

While dynamic prediction is most often formulated as a high resolution temporal problem, we focus on predicting the final outcome of an object's motion in a dynamic scene, rather than predicting the fine-grained object trajectory through space. With an understanding that the main purpose of predictive models is to allow autonomous agents to take appropriate actions by reasoning through the consequences of those actions on the world, we believe that in many scenarios, reasoning about targeted future events is sufficient to make informed decisions. Importantly, we examine the relevant tactile information to make such predictions. Whereas motion prediction is most commonly approached as a purely visual product, we highlight the importance of reasoning through physical phenomena such as interaction forces, slippage, contact geometry, etc., to make informed decisions about the object state.

Visuotactile Sensing

This section describes a novel visuotactile sensor, named the See-Through-your-Skin (STS) sensor, that renders dual stream high resolution images of the contact geometry and the external world, as shown in Fig. 2. The key features are:

Multimodal Perception. By regulating the internal lighting conditions of the STS sensor, the transparency of the reflective paint coating of the sensor can be controlled, allowing the sensor to provide both visual and tactile feedback about the contacting object.

High-Resolution Sensing. Both visual and tactile signals are given as a high resolution image of 1640×1232. We use the Variable Focus Camera Module for Raspberry Pi by Odeseven, which provides a $160°$ field of view. This results in two sensing signals that have the same point of view, frame of reference, and resolution.

Sensor Design

Inspired by recent developments in the GelSight technology (Yuan, Dong, and Adelson 2017), the STS visuotactile sensor is composed of a compliant membrane, internal illumination sources, a reflective paint layer, and a camera. When an object is pressed against the sensor, a camera located within the sensor captures the view through the "skin" as well as the deformation of the compliant membrane, and produces an image that encodes tactile information, such as contact geometry, interactions forces, and stick/slip behavior.

While optical tactile sensors typically make use of an opaque and reflective paint coating, we developed a membrane with a controllable transparency, allowing the sensor to provide tactile information about physical interactions and visual information about the world external to the sensor. This ability to capture a visual perspective of the region beyond the contact surface enables the sensor to visualize color and the appearance of the objects as they collide with the sensor. We control the duty cycle of tactile versus visual

measurements of the sensor by changing the internal lighting condition of the STS sensor, which sets the transparency of the reflective paint coating of the sensor. More details on the design can be found in Hogan et al. (2021).

Simulator

We developed a visuotactile simulator for the STS sensor within the PyBullet environment that reconstructs high resolution tactile signatures from the contact force and geometry. We exploit the simulator to quickly generate large visuotactile datasets of object interactions in dynamic scenes to validate the performance of perception models. The simulator maps the geometric information of the colliding objects via the shading equation (Yuan, Dong, and Adelson 2017):

$$\mathbf{I}(x,y) = \mathbf{R}(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}), \tag{1}$$

where $\mathbf{I}(x,y)$ is the image intensity, $z = f(x,y)$ is the height map of the sensor surface, and $\mathbf{R}$ is the reflectance function modeling the environment lighting and surface reflectance (Yuan, Dong, and Adelson 2017).

Following Gomes, Wilson, and Luo (2019), we implement the reflectance function $\mathbf{R}$ using Phong's reflection model, which breaks down the lighting into three main components of ambient, diffuse, and specular for each channel:

$$\mathbf{I}(x,y) = k_a i_a + \sum_{m \in lights} k_d (\hat{L}_m \cdot \hat{N}) i_{m,d} + k_s (\hat{R}_m \cdot \hat{V})^\alpha i_{m,s}, \tag{2}$$

where $\hat{L}_m$ is the direction vector from the surface point to the light source m, $\hat{N}$ is the surface normal, $\hat{R}_m$ is the reflection vector computed by $\hat{R}_m = 2(\hat{L}_m \cdot \hat{N})\hat{N} - \hat{L}_m$, and $\hat{V}$ is the direction vector pointing towards the camera. Additional information is provided in the supplemental material.

Multimodal Perception

We present a generative multimodal perceptual system that integrates visual, tactile and 3D pose (when available) feedback within a unified framework. We make use of Multimodal Variational Autoencoders (MVAE) (Wu and Goodman 2018) to learn a shared latent representation that encodes all modalities. We further show that this embedding space can encode key information about objects such as shape, color, and interaction forces, necessary to make inferences about intuitive physics.

The predicted outcome of a dynamic interaction can be formulated as a self-supervision problem, where the target visual and tactile images are generated given observed context frames. Our objective is to learn a generator that maps the current available observations to the predicted configuration of the resting state. We show that the MVAE architecture can be trained to predict the most stable and informative elements of a multimodal motion trajectory.

Variational Autoencoders

Generative latent variable models learn the joint distribution of the data and the unobservable representations in the form of $p_\theta(x,z) = p_\theta(z) p_\theta(x|z)$, where $p_\theta(z)$ and $p_\theta(x|z)$
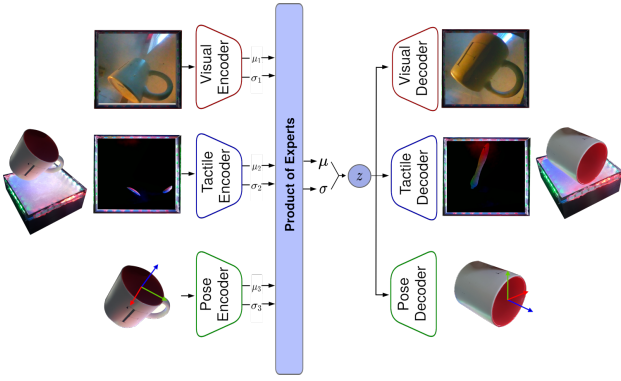
Figure 3: Multimodal dynamics modelling. A generative perceptual system that integrates visual, tactile and 3D pose feedback within a unified Multimodal Variational Autoencoder framework. The network gets the current object configuration and predicts its resting configuration.

denote the prior and the conditional distributions, respectively. The objective is to maximize the marginal likelihood given by $p_\theta(x) = \int p_\theta(z)p_\theta(x|z)dz$. Since the integration is in general intractable, variational autoencoders (VAE) (Kingma and Welling 2013) optimize a surrogate cost, the evidence lower bound (ELBO), by approximating the posterior $p_\theta(x|z)$ with an inference network $q_\phi(z|x)$. The ELBO loss is then given by:

$$\text{ELBO}(x) \triangleq \mathbb{E}_{q_\theta(z|x)}\big[\lambda \log p_\theta(x|z)\big] - \beta \text{KL}\big(q_\phi(z|x)||p_\theta(z)\big), \quad (3)$$

where the first term denotes the reconstruction loss measuring the expectation of the likelihood of the reconstructed data given the latent variables and the second term is the Kullback-Leibler divergence between the approximate and true posterior and acts as a regularization term. In order to regularize the terms in the ELBO loss, β (Higgins et al. 2017) and λ are used as weights.

Multimodal Variational Autoencoders

The VAE uses an inference network to map the observations to a latent space, followed by a decoder to map the latent variables back to the observation space. While this approach is practical with a constant observation space, it becomes challenging when using multiple modalities, where the dimensions of the observation space vary with the availability of the modalities. For example, tactile information only becomes available when contact is made with the sensor. For such multimodal problems that present variability in the availability of data, we would require training an inference network $q(z|X)$ for each subset of modalities $X \subseteq \{x_1, x_2, \ldots, x_N\}$, resulting in a total of 2^N combinations. To deal with this combinatorial explosion of modalities, Wu and Goodman (2018) propose the notion of Product of Experts (PoE) to efficiently learn the approximate joint posterior of different modalities as the product of individual posteriors of each modality. This method has the advantage of training only N inference networks, one for each modality, allowing for better scaling.

Multimodal generative modeling learns the joint distribution of all modalities as:

$$p_\theta(x_1, \ldots, x_N, z) = p(z)p_\theta(x_1|z) \ldots p_\theta(x_N|z), \quad (4)$$

where x_i denotes the observation associated with mode i, N is the total number of available modes, and z is the shared latent space. Assuming conditional independence between modalities, we can rewrite the joint posterior as:

$$\begin{aligned} p(z|x_1, \ldots, x_N) &= \frac{p(x_1, \ldots, x_N|z)p(z)}{p(x_1, \ldots, x_N)} \\ &= \frac{p(z)}{p(x_1, \ldots, x_N)} \prod_{i=1}^{N} p(x_i|z) \\ &= \frac{p(z)}{p(x_1, \ldots, x_N)} \prod_{i=1}^{N} \frac{p(z|x_i)p(x_i)}{p(z)} \\ &\propto \frac{\prod_{i=1}^{N} p(z|x_i)}{\prod_{i=1}^{N-1} p(z)}. \end{aligned} \quad (5)$$

By approximating $p(z|x_i)$ in Equation (5) with $q(z|x_i) \equiv \widetilde{q}(z|x_i)p(z)$, where $\widetilde{q}(z|x_i)$ is the inference network of modality i, we obtain:

$$p(z|x_1, \ldots, x_N) \propto p(z) \prod_{i=1}^{N} \widetilde{q}(z|x_i), \quad (6)$$

that is recognized as the Product of Expert (PoE). In the case of variational autoencoders where $p(z)$ and $\widetilde{q}(z|x_i)$ are multivariate Gaussians, the PoE can be computed analytically as the product of two Gaussians (Cao and Fleet 2014).

An important advantage of the MVAE is that unlike other multimodal generative models, it can be efficiently scaled up to several modalities, as it requires training only N inference models rather than the 2^N multimodal inference networks. Additionally, the notion of PoE allows for continuous inference in the case of discontinuous and unavailable modalities.

Learning Intuitive Physics with MVAEs

We use the MVAE architecture to learn a shared representation that exploits multiple sensing modalities for learning the underlying dynamics of intuitive physics. A key advantage of this formulation is that it enables combining sensing modalities, while naturally dealing with intermittent contacts, during which tactile measurements are discontinuous.

While variational autoencoders are often trained by reconstructing the encoder inputs, we introduce a time-lag element into the network architecture (Hernández et al. 2018; Kipf et al. 2018), where the outputs of the decoder are set to predict future frames. We adapt the ELBO loss in Eq. (3) to:

$$\text{ELBO}(x_t, x_T) \triangleq \mathbb{E}_{q_\theta(z|x_t)}\big[\lambda \log p_\theta(x_T|z)\big] - \beta \text{KL}\big(q_\phi(z|x_t)||p_\theta(z)\big), \quad (7)$$

where t and T denote the input and output time instances.

Figure 3 describes our dynamics model learning framework, where visual, tactile, and 3D pose are fused together to learn a shared embedding space via three unimodal encoder-decoders connected through the Product of Experts. To train
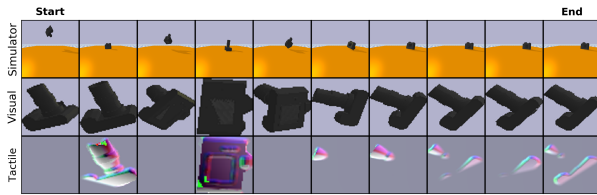
(a) Freefalling objects on a flat surface.

(b) Objects sliding down an inclined plane.
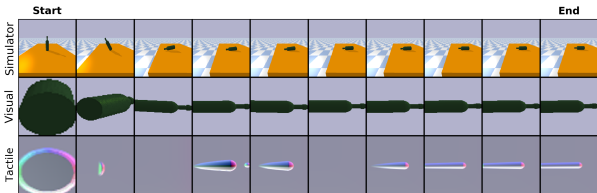
(c) Objects perturbed from a stable resting pose.

Figure 4: Simulated example episodes for three dynamic simulated scenarios. The top row shows the 3D object view while the middle and bottom rows show the visual and tactile measurement captured by the STS sensor, respectively. Some frames have been removed for clarity.

the model loss, we follow the sampling methodology proposed in (Wu and Goodman 2018), where we compute the ELBO loss by enumerating the subsets of the modalities $\mathcal{M} = \{\text{visual}, \text{tactile}, \text{pose}\}$:

$$\mathcal{L}(x_t) = \sum_{X \in \mathbb{P}(\mathcal{M})} \text{ELBO}(X_t, X_T), \qquad (8)$$

where $\mathbb{P}(\mathcal{M})$ is the powerset of the modalities set $\mathcal{M}$.

In the cases where there is an input to the dynamics model (e.g., force perturbation in the third simulated scenario), we include the conditional dependence of the input condition c on the ELBO loss as:

$$\begin{aligned} \text{ELBO}(x_t, x_T | c) \triangleq \ &\mathbb{E}_{q_\theta(z|x_t,c)} \big[\lambda \log p_\theta(x_T|z,c) \big] \qquad (9) \\ &- \beta \text{KL} \big(q_\phi(z|x_t,c) || p_\theta(z|c) \big), \end{aligned}$$

Data Collection

The simulated dataset was collected using the PyBullet simulator described earlier, and the real-world dataset was collected using a prototype of the STS sensor. While more details are provided in the supplemental material [2], we give an overview of the experimental setup.
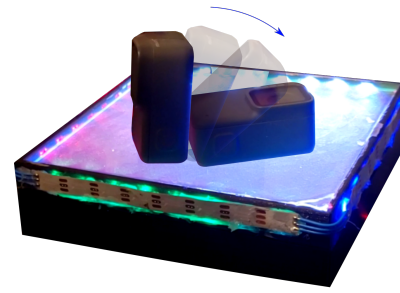
Figure 5: Real-world data collection. An electronic device (GoPro camera) is released from an unstable initial configuration. The task is to predict the resting configuration of the object from its initial measurements.

Simulated Dataset We consider three simulated physical scenarios, as shown in Fig. 4, involving eight household object classes [3] drawn from the 3D ShapeNet dataset (Chang et al. 2015). The tasks ordered in increasing difficulty are:
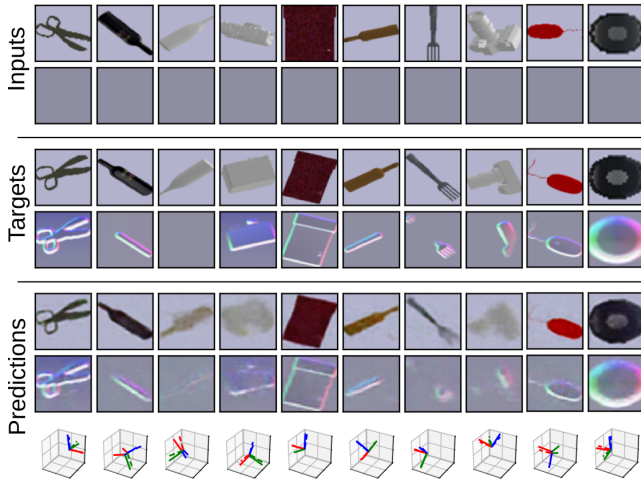
Freefalling Objects on a Flat Surface. This experiment releases objects with random initial poses over the STS sensor, where they collide multiple times with the sensor prior to coming to rest. We collect a total of 1700 trajectories comprising 100k images.

Objects Sliding Down an Inclined Plane. This experiment, inspired by Wu et al. (2015), places objects with random initial poses atop an inclined surface, where they can either stick due to friction or slide down. While sliding down, the objects can roll, causing the final configuration to be significantly different than the initial. We collect a total of 2400 trajectories comprising 145k images.

Objects Perturbed from a Stable Resting Pose. In this scenario, we consider an object initially stably resting on the sensor, that is perturbed from its equilibrium by a randomly sampled quick lateral acceleration of the sensor. This experiment only considers bottles due to their elongated and unstable shape allowing for different outcomes (e.g., toppling, sliding or standing) based on the direction and magnitude of the applied force. Due to such diverse outcomes, this task is considerably more complicated than the other two. We collect a total of 2500 trajectories comprising 150k images.

Real-World Dataset We validate the predictive accuracy of our proposed framework on a small real-world dataset collected manually using the STS sensor. We collect 2000 images from 500 trajectories using a small electronic device (GoPro). This object has been selected due to its small form factor (small enough to fit on the 15cm×15cm sensor prototype) and its mass (heavy enough to leave a meaningfully tactile signature on the sensor). Each trajectory includes the initial and final visual and tactile images, obtained by rapidly turning on/off the internal lights of the sensor. The object is released from an unstable initial position, while being in contact with the sensor, as illustrated in Fig. 5, and the end of the episode is determined once the object is immobile.

(a) Freefalling objects on a flat surface. The motion starts from a non-contacting initial position, accounting for the initial unavailability of tactile measurements.
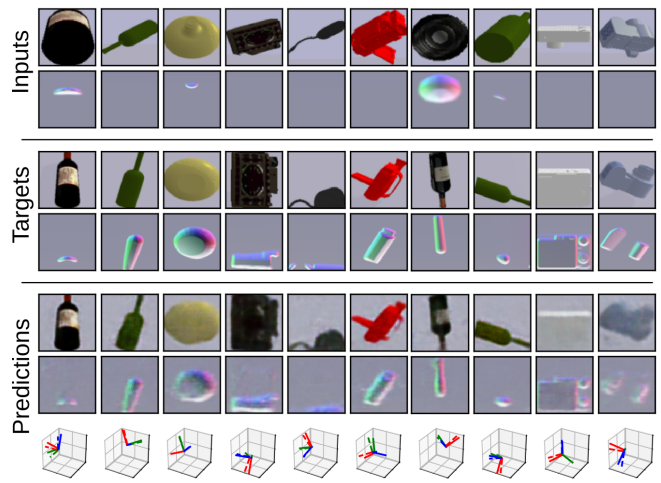
(b) Objects sliding down an inclined plane. The motion starts in proximity to the surface, accounting for the the initially availability of tactile measurements that may be either active or inactive.

Figure 6: Multimodal predictions for three simulated scenarios evaluated on the validation set. The model predicts the final resting pose in addition to the visual and tactile measurements of the STS sensor. The bottom row compares the predicted pose (solid coordinates) with the ground truth (dashed coordinates).

Experimental Results

We validate the ability of our approach to predict the evolution of physical configurations on simulated and real scenes. We downsample the sensor images to a resolution of 64×64 images and an identical network architecture and training parameters consistent across all evaluations. More details are provided in the supplemental material.

Simulation   We compare the performance of our proposed framework against two baselines. First, we highlight the value of multimodal sensing for learning dynamics models. Second, we compare our model against dynamic models parameterized using higher temporal resolutions.

In Fig. 6 and 7, we present the multimodal predictions for three simulated scenarios evaluated on the validation set. We show the MVAE's ability to predict the raw visual and tactile measurements of the resting configuration of an object with high accuracy, with the predictions closely matching the ground truth labels. Interestingly, the model learns a mapping between the visual, tactile and 3D pose modes allowing it to sample correlating outputs from the learned shared embedding. Fig. 6(a) highlights the ability of the MVAE model to handle missing modalities, such as when tactile information is unavailable in the input. Finally, the model learns to accurately predict instances where the object has fallen from the surface of the sensor, resulting in empty output images.

In Figure 7, we present the prediction results for the scenario where objects are perturbed from an initial stable pose. Unlike the first two experiments, this scenario includes a random lateral force applied on the system that plays a significant role in determining the resting state of the objects, thus making it considerably more complicated. To account

for this, we condition the MVAE with information about the magnitude and direction of the lateral force using Eq. (9). The results in Fig. 7 indicate that the model successfully integrates information about applied forces by correctly predicting the outcome about the object motion (i.e., toppling or falling). Further comparison with the unimodal models shows that the tactile mode has played an important role in
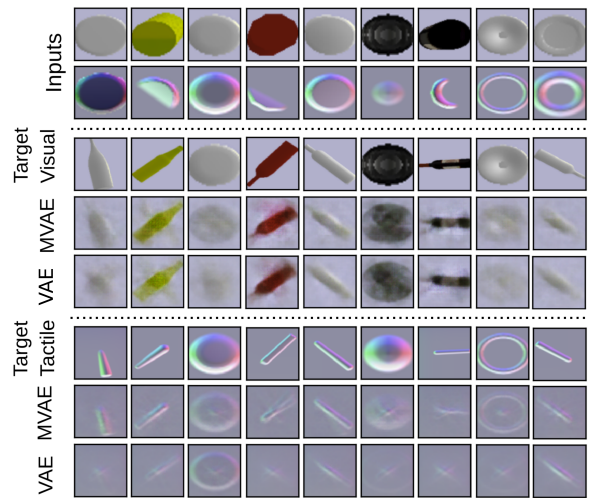

Figure 7: Qualitative comparison of visual and tactile predictions of MVAE with unimodal VAE for the simulated force perturbation scenario obtained on the validation set. MVAE leverages the tactile mode to provide clearer predictions of resting configuration.
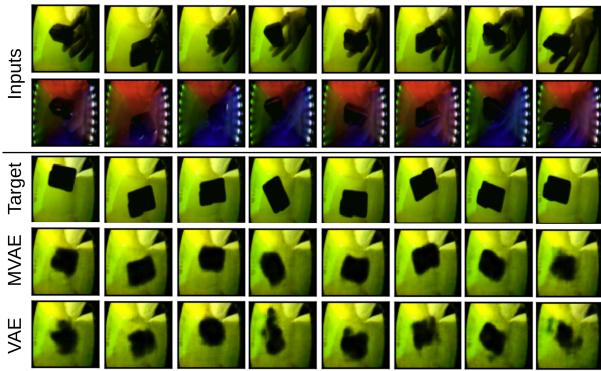
Figure 8: Qualitative comparison of visual predictions of MVAE with unimodal VAE for the real dataset obtained on the validation set. MVAE leverages the tactile mode to provide more accurate and clearer visual predictions of the resting configuration.

↓ Task	Setting → Model ↓	Visual Perf. (×1e-4)		Tactile Perf. (×1e-4)	
		Multi Step	Final Step	Multi Step	Final Step
Freefall	VAE-visual only	6522	5750	NA	NA
	VAE-tactile only	NA	NA	6770	**6703**
	MVAE w/ pose	6548	**5741**	6726	**6703**
	MVAE w/o pose	6635	5752	6735	**6703**
Inclined	VAE-visual only	6751	5895	NA	NA
	VAE-tactile only	NA	NA	6714	6711
	MVAE w/ pose	6625	**5891**	6719	**6709**
	MVAE w/o pose	6549	**5890**	6713	**6710**
Perturb.	VAE-visual only	7369	6158	NA	NA
	VAE-tactile only	NA	NA	6879	6705
	MVAE w/ pose	7552	**6054**	6868	**6701**
	MVAE w/o pose	6967	6095	6896	**6702**

Table 1: Prediction performance of fixed-step and final-step predictions for unimodal and multimodal VAE models for the three simulated scenarios. Performance is reported as the average of the binary cross-entropy error on the validation set. The bold cells indicate preferred values.

Setting → Model ↓	Visual (×1e-4)	Tactile (×1e-4)
VAE-visual only	3532	NA
VAE-tactile only	NA	2872
MVAE w/o pose	**3509**	**2819**

Table 2: Prediction performance of final-step predictions for unimodal and multimodal VAE models on the real dataset. The performance is reported as the average of the binary cross-entropy error (BCE) on the validation set. The bold cells indicate the best values.

reducing the uncertainty and blurriness of the predictions.

In Table 1, we present the quantitative results comparing the multimodal and unimodal models for both one-step (high temporal resolution) and resting state predictions (resting object configuration). We draw two conclusions from this analysis. First, models that exploit multimodal sensing outperform those relying on visual/tactile modalities alone. Importantly, we find that tactile information improves the prediction accuracy (for both visual and tactile predictions) by reasoning about interaction forces and the geometry of contact. Second, we show that predicting the resting state of the object outperforms dynamic models using high temporal resolution. This is due to the fact that uncertainties and errors propagate forward in time, leading to blurry and imprecise predictions. The results show that in dynamic scenarios where the intermediate states are not of interest, we can learn to predict the final outcome with higher accuracy without explicitly reasoning about intermediate steps.

Real Sensor This section validates the predictive abilities of the perceptual system on a real-world experiment setup using the STS sensor. In Figure 8, we show the network's ability to predict the resting object configuration, reasoning through both the visual and tactile images. Qualitative comparison of visual predictions of MVAE with unimodal VAE shows that the MVAE model leverages the tactile mode to reason more accurately about the resting configuration. In Table 2, we quantitatively compare the predictive performances of multimodal and unimodal models, highlighting the important role played by both visual and tactile measurements to determine the outcome of physical interactions.

Conclusion

We have designed and implemented a system that exploits visual and tactile feedback to make physical predictions about the motion of objects in dynamic scenes. We exploit a novel visuotactile sensor, the See-Through-your-Skin (STS) sensor, that represents both modalities as high resolution images. The perceptual system uses a multimodal variational

autoencoder (MVAE) neural network architecture that maps the sensing modalities to a shared embedding, used to infer the stable resting configuration of objects during physical interactions. By focusing on predicting the most stable elements of a trajectory, we validate the predictive abilities of our dynamics models in three simulated dynamic scenarios and a real-world experiment using the STS sensor. Results show that the MVAE framework can accurately predict the future state of objects during physical interactions with a surface. Importantly, we find that predicting object motions benefits from exploiting both modalities: visual information captures object properties such as 3D shape and location, while tactile information provides critical cues about interaction forces and resulting object motion and contacts.

Ethics Statement

The research in this work is related to the development of smarter devices, possibly including assistive devices, but we do not foresee any other substantive ethical implications.

References

Abad, A.; and Ranasinghe, A. 2020. Visuotactile Sensors with Emphasis on GelSight Sensor: A Review. *IEEE Sensors Journal* PP: 1–1. doi:10.1109/JSEN.2020.2979662.

Allen, P. 1984. Surface descriptions from vision and touch. In *IEEE International Conference on Robotics and Automation (ICRA)*, volume 1, 394–397. IEEE.

Bacon, P.-L.; Harb, J.; and Precup, D. 2017. The option-critic architecture. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Bierbaum, A.; Gubarev, I.; and Dillmann, R. 2008. Robust Shape Recovery for Sparse Contact Location and Normal Data from Haptic Exploration. In *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 3200–3205. Nice, France.

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

Calandra, R.; Owens, A.; Upadhyaya, M.; Yuan, W.; Lin, J.; Adelson, E. H.; and Levine, S. 2017. The feeling of success: Does touch sensing help predict grasp outcomes? In *1st Conf. on Robot Learning (CoRL)*. Mountain View, CA.

Cao, Y.; and Fleet, D. J. 2014. Generalized product of experts for automatic and principled fusion of Gaussian process predictions. *arXiv preprint arXiv:1410.7827* .

Chang, A. X.; Funkhouser, T.; Guibas, L.; Hanrahan, P.; Huang, Q.; Li, Z.; Savarese, S.; Savva, M.; Song, S.; Su, H.; et al. 2015. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012* .

Donlon, E.; Dong, S.; Liu, M.; Li, J.; Adelson, E.; and Rodriguez, A. 2018. GelSlim: A high-resolution, compact, robust, and calibrated tactile-sensing finger. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 1927–1934. Madrid, Spain: IEEE.

Driess, D.; Englert, P.; and Toussaint, M. 2017. Active Learning with Query Paths for Tactile Object Shape Exploration. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. Vancouver, Canada.

Fazeli, N.; Zapolsky, S.; Drumwright, E.; and Rodriguez, A. 2020. Fundamental limitations in performance and interpretability of common planar rigid-body contact models. In Amato, N.; Hager, G.; Thomas, S.; and Torres-Torriti, M., eds., *Robotics Research*, 555–571. Springer.

Gomes, D. F.; Wilson, A.; and Luo, S. 2019. GelSight Simulation for Sim2Real Learning. In *CRA ViTac Workshop*. Montreal, Canada. Held in conjunction with IEEE ICRA.

Hernández, C. X.; Wayment-Steele, H. K.; Sultan, M. M.; Husic, B. E.; and Pande, V. S. 2018. Variational encoding of complex dynamics. *Physical Review E* 97(6): 062412.

Higgins, I.; Matthey, L.; Pal, A.; Burgess, C.; Glorot, X.; Botvinick, M.; Mohamed, S.; and Lerchner, A. 2017. beta-vae: Learning basic visual concepts with a constrained vari-

ational framework. In *International Conference on Learning Representations (ICLR)*. Toulon, France.

Hogan, F. R.; Jenkin, M.; Rezaei-Shoshtari, S.; Girdhar, Y.; Meger, D.; and Dudek, G. 2021. Seeing Through your Skin: Recognizing Objects with a Novel Visuotactile Sensor. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1218–1227.

Izatt, G.; Mirano, G.; Adelson, E.; and Tedrake, R. 2017. Tracking objects with point clouds from vision and touch. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 4000–4007. IEEE.

Jayaraman, D.; Ebert, F.; Efros, A.; and Levine, S. 2018. Time-Agnostic Prediction: Predicting Predictable Video Frames. In *International Conference on Learning Representations (ICLR)*. Vancouver, Canada.

Johnson, M. K.; and Adelson, E. 2009. Retrographic sensing for the measurement of surface texture and shape. In *IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 1070–1077. Miami, FL.

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

Kipf, T.; Fetaya, E.; Wang, K.-C.; Welling, M.; and Zemel, R. 2018. Neural Relational Inference for Interacting Systems. In *International Conference on Machine Learning*, 2688–2697.

Kuppuswamy, N.; Alspach, A.; Uttamchandani, A.; Creasey, S.; Ikeda, T.; and Tedrake, R. 2020. Soft-Bubble grippers for robust and perceptive manipulation. *arXiv preprint arXiv:2004.03691* .

Lambeta, M.; Chou, P.-W.; Tian, S.; Yang, B.; Maloon, B.; Most, V. R.; Stroud, D.; Santos, R.; Byagowi, A.; Kammerer, G.; et al. 2020. DIGIT: A Novel Design for a Low-Cost Compact High-Resolution Tactile Sensor With Application to In-Hand Manipulation. *IEEE Robotics and Automation Letters* 5(3): 3838–3845.

Lee, J.-T.; Bollegala, D.; and Luo, S. 2019. "Touching to See" and "Seeing to Feel": Robotic Cross-modal Sensory Data Generation for Visual-Tactile Perception. In *2019 International Conference on Robotics and Automation (ICRA)*, 4276–4282. IEEE.

Lee, M. A.; Zhu, Y.; Srinivasan, K.; Shah, P.; Savarese, S.; Fei-Fei, L.; Garg, A.; and Bohg, J. 2019. Making Sense of Vision and Touch: Self-Supervised Learning of Multimodal Representations for Contact-Rich Tasks. In *2019 International Conference on Robotics and Automation (ICRA)*, 8943–8950.

Lee, M. A.; Zhu, Y.; Srinivasan, K.; Shah, P.; Savarese, S.; Fei-Fei, L.; Garg, A.; and Bohg, J. 2019. Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks. In *2019 International Conference on Robotics and Automation (ICRA)*, 8943–8950. IEEE.

Li, R.; Platt, R.; Yuan, W.; ten Pas, A.; Roscup, N.; Srinivasan, M. A.; and Adelson, E. 2014. Localization and manipulation of small parts using gelsight tactile sensing.

In *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, 3988–3993. Hong Kong, China: IEEE.

Li, Y.; Zhu, J.-Y.; Tedrake, R.; and Torralba, A. 2019. Connecting touch and vision via cross-modal prediction. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 10609–10618. Long Beach, CA.

Luo, S.; W. Mou, W.; Althoefer, K.; and Liu, H. 2016. Iterative Closest Labeled Point for Tactile Object Shape Recognition. In *IEEE/RSJ IROS*, 3137–3142. Daejon, Korea. doi:10.1109/IROS.2016.7759485.

Luo, S.; Yuan, W.; Adelson, E.; Cohn, A. G.; and Fuentes, R. 2018. Vitac: Feature sharing between vision and tactile sensing for cloth texture recognition. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2722–2727. IEEE.

McGovern, A.; and Barto, A. G. 2001. Automatic Discovery of Subgoals in Reinforcement Learning using Diverse Density. In *Proceedings of the Eighteenth International Conference on Machine Learning*, 361–368.

Nair, S.; and Finn, C. 2019. Hierarchical Foresight: Self-Supervised Learning of Long-Horizon Tasks via Visual Subgoal Generation. In *International Conference on Learning Representations*.

Nair, S.; Savarese, S.; and Finn, C. 2020. Goal-Aware Prediction: Learning to Model What Matters. *arXiv preprint arXiv:2007.07170* .

Neitz, A.; Parascandolo, G.; Bauer, S.; and Schölkopf, B. 2018. Adaptive skip intervals: Temporal abstraction for recurrent dynamical models. In *Advances in Neural Information Processing Systems*, 9816–9826.

Ottenhaus, S.; Miller, M.; Schiebener, D.; Vahrenkamp, N.; and Asfour, T. 2016. Local implicit surface estimation for haptic exploration. In *IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, 850–856. doi:10.1109/HUMANOIDS.2016.7803372.

Padmanabha, A.; Ebert, F.; Tian, S.; Calandra, R.; Finn, C.; and Levine, S. 2020. OmniTact: A Multi-Directional High-Resolution Touch Sensor. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 618–624. IEEE.

Pertsch, K.; Rybkin, O.; Yang, J.; Zhou, S.; Derpanis, K.; Daniilidis, K.; Lim, J.; and Jaegle, A. 2020. Keyframing the Future: Keyframe Discovery for Visual Prediction and Planning. In *Learning for Dynamics and Control*, 969–979. PMLR.

Shimonomura, K. 2019. Tactile image sensors employing camera: A review. *Sensors* 19(18): 3933.

Smith, E.; Calandra, R.; Romero, A.; Gkioxari, G.; Meger, D.; Malik, J.; and Drozdzal, M. 2020. 3d shape reconstruction from vision and touch. *Advances in Neural Information Processing Systems* 33.

Sutton, R. S.; Precup, D.; and Singh, S. 1999. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence* 112(1-2): 181–211.

Tremblay, J.-F.; Manderson, T.; Noca, A.; Dudek, G.; and Meger, D. 2020. Multimodal dynamics modeling for off-road autonomous vehicles. *arXiv preprint arXiv:2011.11751* .

Vlack, K.; Mizota, T.; Kawakami, N.; Kamiyama, K.; Kajimoto, H.; and Tachi, S. 2005. GelForce: a vision-based traction field computer interface. In *CHI'05 extended abstracts on Human factors in computing systems*, 1154–1155.

Wang, S.; Wu, J.; Sun, X.; Yuan, W.; Freeman, W. T.; Tenenbaum, J. B.; and Adelson, E. H. 2018. 3d shape perception from monocular vision, touch, and shape priors. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 1606–1613. IEEE.

Watkins-Valls, D.; Varley, J.; and Allen, P. 2019. Multimodal geometric learning for grasping and manipulation. In *2019 International conference on robotics and automation (ICRA)*, 7339–7345. IEEE.

Wu, J.; Yildirim, I.; Lim, J. J.; Freeman, B.; and Tenenbaum, J. 2015. Galileo: Perceiving physical object properties by integrating a physics engine with deep learning. In *Advances in neural information processing systems*, 127–135.

Wu, M.; and Goodman, N. 2018. Multimodal generative models for scalable weakly-supervised learning. In *32nd Conf. on Neural Information Processing Systems (NeurIPS)*, 5575–5585. Montreal, Canada.

Yamaguchi, A.; and Atkeson, C. G. 2017. Implementing tactile behaviors using fingervision. In *2017 IEEE-RAS 17th International Conference on Humanoid Robotics (Humanoids)*, 241–248. IEEE.

Yi, Z.; Calandra, R.; Filipe, F. V.; van Hoof, H.; Tucker, H.; Yilei, Z.; and Peters, J. 2016. Active Tactile Object Exploration with Gaussian Processes. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 4925–4930. Stockholm, Sweden. doi:10.1109/IROS.2016.7759723.

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

Yuan, W.; Wang, S.; Dong, S.; and Adelson, E. 2017. Connecting look and feel: Associating the visual and tactile properties of physical materials. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5580–5588.