Generating Visual Scenes from Touch

Fengyu Yang Jiacheng Zhang Andrew Owens

University of Michigan

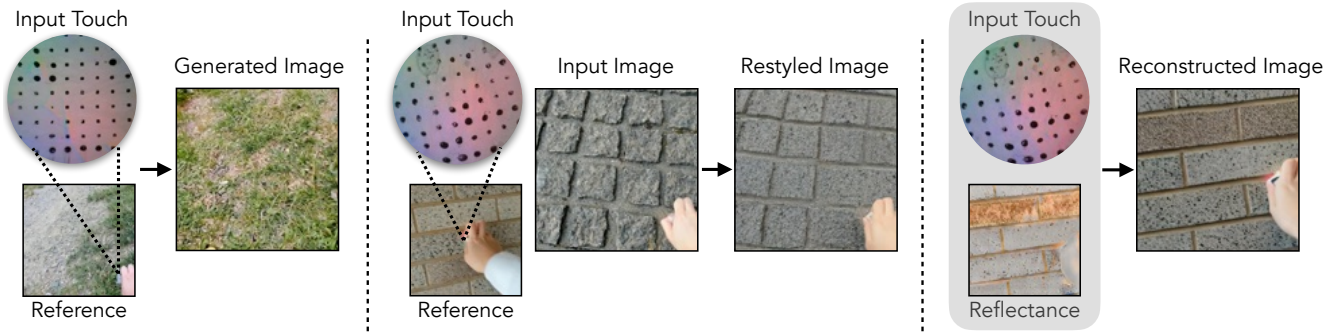


Figure 1: **Generating and manipulating images via touch**. We propose a model based on latent diffusion that translates between touch and images (and vice versa), unifying many previous visuo-tactile image synthesis tasks and enabling new ones. (a) We generate an image of a scene given a tactile signal. (b) We perform tactile-driven image stylization, *e.g.* restyling a rough rock to match the smoother texture of a brick. (c) We propose the novel task of *tactile-driven shading estimation*: predicting an image from its reflectance and tactile signal. To aid visualization, we show reference images next to the touch signal. We present a circular crop from the touch signal to emphasize the part of the signal that is in contact with the object.

Abstract

An emerging line of work has sought to generate plausible imagery from touch. Existing approaches, however, tackle only narrow aspects of the visuo-tactile synthesis problem, and lag significantly behind the quality of cross-modal synthesis methods in other domains. We draw on recent advances in latent diffusion to create a model for synthesizing images from tactile signals (and vice versa) and apply it to a number of visuo-tactile synthesis tasks. Using this model, we significantly outperform prior work on the tactile-driven stylization problem, i.e., manipulating an image to match a touch signal, and we are the first to successfully generate images from touch without additional sources of information about the scene. We also successfully use our model to address two novel synthesis problems: generating images that do not contain the touch sensor or the hand holding it, and estimating an image's shading from its reflectance and touch. Project Page: https://fredfyyang.github.io/vision-from-touch/

1. Introduction

Humans rely crucially on cross-modal associations between sight and touch to physically interact with the world [57]. For example, our sense of sight tells us how the ground in front of us will feel when we place our feet on it, while our sense of touch conveys the likely visual appearance of an unseen object in our hand. Translating between these modalities requires an understanding of physical and material properties. Models trained to solve this problem must, for instance, learn to associate rapid changes in visual shading with rough microgeometry, and smooth visual textures with soft surfaces.

Touch is arguably the most important sensory modality for humans [47, 42, 39], due to its role in basic survival [39, 9, 22] and physical interaction. Yet touch sensing has received comparably little attention in multimodal learning. An emerging line of work has addressed the problem of translating touch to sight, such as by learning joint embeddings [63, 38], manipulating visual styles to match a tactile signal [63], or adding a plausible imagery of a robotic arm to an existing photo of a scene [37]. While these tasks each capture important parts of the cross-modal prediction problem, each currently requires a separate, special-purpose method. Existing methods also lag significantly behind those of other areas of multimodal perception, which provide general-purpose methods for cross-modal synthe-

sis, and can translate between modalities without the aid of extra conditional information.

In this paper, we generate plausible images of natural scenes from touch (and vice versa), drawing on recent advances in diffusion models [50, 12, 20, 21, 44]. We adapt latent diffusion models to a variety of visuo-tactile synthesis problems. Our proposed framework obtains strong results on several novel synthesis problems, and unifies many previously studied visuo-tactile synthesis tasks.

First, we study the problem of generating images from touch (and vice versa). We address the task of generating images from touch without any image-based conditioning, where we are the first method to successfully generate images for natural scenes (Fig. 1a). We also address the task of adding an arm to a photo of an existing scene, where we significantly outperform prior work [37].

Second, we address the recently proposed *tactile-driven image stylization* task, *i.e.*, the problem of manipulating an image to match a given touch signal [63] (Fig. 1b), using an approach based on guided image synthesis [43]. Our approach obtains results that are higher fidelity and that match the tactile signal significantly more closely than those of prior work. It also provides the ability to control the amount of image content preserved from the input image.

Finally, we show that we can augment our model with additional conditional information. Taking inspiration from the classic problem of intrinsic image decomposition [40, 3], we perform *tactile-driven shading estimation*, predicting an image after conditioning on reflectance and touch (Fig. 1c). Since changes in tactile microgeometry often manifest as changes in shading (*i.e.*, the information missing from reflectance), this tests the model's ability to link the two signals. We also use segmentation masks to create "hand-less" images that contain the object being pressed but not the tactile sensor or arm that pressed it.

We demonstrate our framework's effectiveness using natural scenes from the *Touch and Go* dataset [63], a collection of egocentric videos that capture a wide variety of materials and objects using GelSight [27], and using robot-collected data from *VisGel* [37].

2. Related Work

Cross-modal synthesis with diffusion models. Diffusion models have recently become a favored generative model family due to their ability to produce high-quality samples. However, one major concern for diffusion models is their slow inference speed due to the iterative generation process on high dimensional data. Recently, latent diffusion [50] addressed this drawback by working on a compressed latent space of lower dimensionality, which allows diffusion models to work on more extensive tasks with accelerating the speed. These models have demonstrated remarkable success in tasks such as image synthesis [12, 20, 21, 44], super-resolution [53], and image editing [56, 43, 8]. Additionally, the advancements in multimodal learning [24, 26, 16] have enabled diffusion models to be utilized for cross-modal synthesis tasks. For vision-language generation, diffusion models have been studied for text-to-image synthesis [1, 28, 45, 49, 52], text-to-speech generation [7, 30, 34], text-to-3D generation [41, 54]. In addition, diffusion models also show promising results in audio synthesis including text-to-audio generation [55], waveform generation [31, 18, 6]. In this work, we are the first to employ diffusion model on real-world visual-tactile data, exploring the possibility of utilizing tactile data as a prompt for image synthesis.

Tactile sensing. Early touch sensors recorded simple, low-dimensional sensory signals, such as measures of force, vibration, and temperature [32, 33, 10]. Beginning with GelSight [64, 27], researchers proposed a variety of vision-based tactile sensors, which convert the deformation of an illuminated membrane using a camera, thereby providing detailed information about shape and material properties [58, 35]. We focus on these sensors, particularly using GelSight, since it is widely used applications [37, 4], and available in visuo-tactile datasets [15, 17, 63]. Crucially, these sensors produce images as output, allowing us to use the same network architectures for both images and touch [65]. Other work proposes collocated vision and touch sensors [61, 5].

Cross-modal models for vision and touch. Li *et al*. [37] used a GAN [23] to translate between tactile signals and images, using a dataset acquired by a robot. In contrast, they require conditioning their touch-to-image model on another photo from the same scene. This is a task that amounts to adding an arm grasping the correct object (given several possible choices), rather than generating an object that could have plausibly led to a touch signal according to its physical properties. It is not straightforward to adapt their method to the other touch-to-image synthesis problems we address without major modifications. Yang *et al*. [63] proposed a visuo-tactile dataset and used a GAN to restyle images to match a touch signal. Their approach only learns a limited number of visual styles, and cannot be straightforwardly adopt extra conditional information (such as reflectance) or be applied to unconditional cross-modal translation tasks. Other work has learned multimodal visuo-tactile embeddings [63, 38]. Other work learns to associate touch and sight for servoing and manipulation [5].

3. Method

Our goal is to translate touch to vision (and vision to touch) using a generative model. We will do this using a model based on latent diffusion [50]. We will use this model to solve a number of tasks, including: 1) cross-modal visual-tactile synthesis, 2) tactile-driven image stylization, and 3) tactile-driven shading estimation.

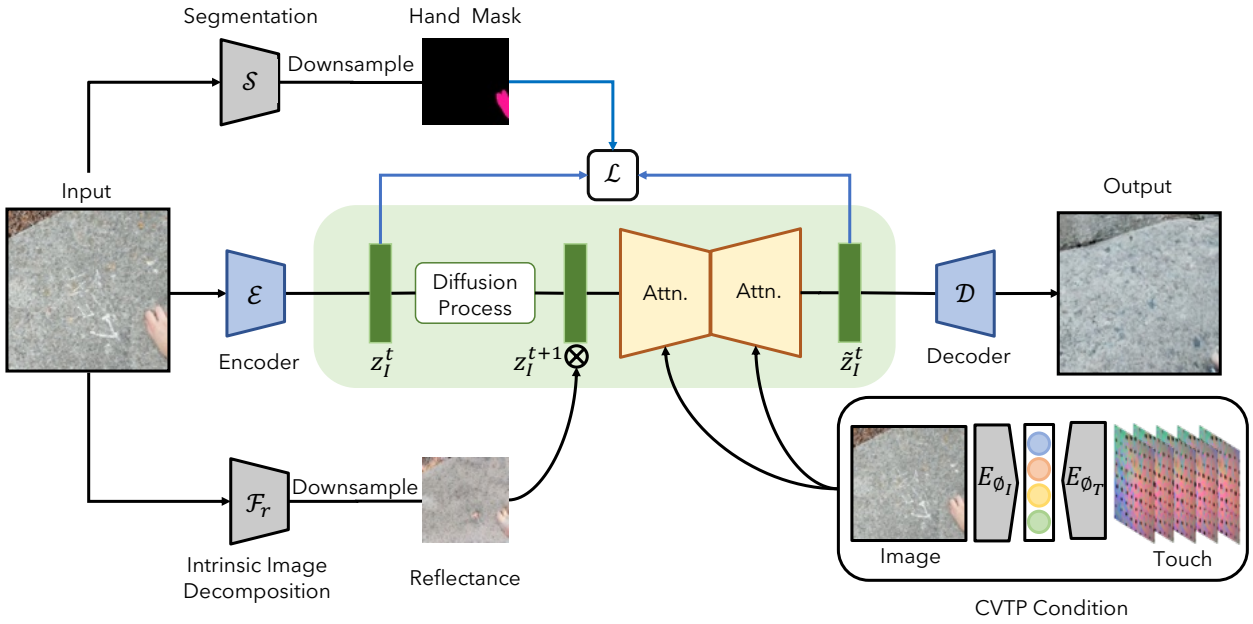


Figure 2: **Touch-to-image model**. We use a latent diffusion model to generate an image of a scene from touch. The touch signal is represented using multiple frames of video from a GelSight sensor. The model uses a segmentation mask to optionally generate only the scene content containing the pressed object (*i.e.*, without a hand or touch sensor). We also optionally condition on reflectance from a scene, in which case the model's generation task requires it to estimate shading.

3.1. Cross-Modal Synthesis of Vision and Touch

We now describe our framework for cross-modal synthesis. First, we describe a contrastive visuo-tactile model, which we use to perform conditional generation. Second, we describe our cross-modal latent diffusion model.

3.1.1 Contrastive Visuo-tactile Pretraining (CVTP)

Following other work in cross-modal synthesis [48, 50], we provide conditional information to our generation models through multimodal embeddings via contrastive learning [14, 66, 62, 59]. Our embedding-learning approach resembles that of Yang *et al*. [63] and contrastive multiview coding [59]. A key difference is that we incorporate temporal information into our visual and tactile representations. Touching an object is a dynamic process, and the information we obtain varies over time, from the moment when the tactile sensor begins touching the object, to the point when the sensor has reached it maximum deformation. Adding temporal cues provides information about material properties that may be hard to perceive from a single sample, such as the hardness or softness of a surface [65, 25].

Given the visual and tactile datasets X_I and X_T, which consist of N synchronized visual-tactile frames $\{\mathbf{x}_I^i, \mathbf{x}_T^i\}_{i=1}^N$, we denote the video clip sampled at time i with the window size $w = 2C + 1$, $v_I^i = \{\mathbf{x}_I^{i-C}, ..., \mathbf{x}_I^i, ..., \mathbf{x}_I^{i+C}\}$ and the corresponding tactile clip $v_I^t = \{\mathbf{x}_I^{i-C}, ..., \mathbf{x}_I^i, ..., \mathbf{x}_I^{i+C}\}$. We denote examples taken from the same visual-tactile recording $\{v_I^i, v_T^i\}$ as positives, and samples from different visual-tactile video pair $\{v_I^i, v_T^j\}$ as negatives.

Our goal is to jointly learn temporal visual $z_I = E_{\phi_I}(v_I)$ and tactile $z_T = E_{\phi_T}(v_T)$ encoder. We use a 2D ResNet as the architecture for both encoders. For easy comparison to static models, we incorporate temporal information into the model via early fusion (concatenating channel-wise).

Then we maximize the probability of finding the corresponding visuo-tactile video pair in a memory bank containing K samples using InfoNCE [46] loss:

$$\mathcal{L}_i^{V_I,V_T} = -\log \frac{\exp(E_{\phi_I}(v_I^i) \cdot E_{\phi_T}(v_T^i)/\tau)}{\sum_{j=1}^K \exp(E_{\phi_I}(v_I^i) \cdot E_{\phi_T}(v_T^j)/\tau)} \quad (1)$$

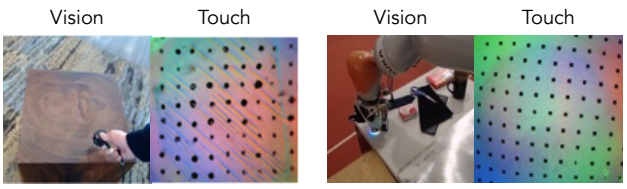
where τ is a small constant. Analogously, we get a symmetric objective $\mathcal{L}^{V_T,V_I}$ and minimize:

$$\mathcal{L}_{\text{CVTP}} = \mathcal{L}^{V_I,V_T} + \mathcal{L}^{V_T,V_I}. \quad (2)$$

3.1.2 Touch-conditioned Image Generation

We now describe the tactile-to-image generation model (an image-to-touch model can be formulated in an analogous way). Our approach follows Rombach *et al*. [50], which translates language to images, but with a variety of extensions specific to the visuo-tactile synthesis problem. Given a visuo-tactile image pair $\{\mathbf{x}_I, \mathbf{x}_T\} \in \mathbb{R}^{H \times W \times 3}$, our goal is to generate an image $\widetilde{\mathbf{x}}_I$ from tactile input $\mathbf{x}_T$. We encode the input $\mathbf{x}$ into a latent representation $\mathbf{z} = \mathcal{E}(\mathbf{x}) \in \mathbb{R}^{h \times w \times 3}$. A decoder $\mathcal{D}$ will reconstruct the image $\hat{x} = \mathcal{D}(\mathbf{z})$ from the code. The latent dimension $h \times w$ is smaller than the image dimension $H \times W$.

Training. We train a touch-to-vision diffusion generation in the latent space $\mathbf{z}_I = \mathcal{E}(\mathbf{x}_I)$. Diffusion models learn



Touch and Go [63] VisGel [37]

Figure 3: **Visuo-tactile datasets**. For our experiments, we evaluate our model on natural scenes from *Touch and Go* [63] and robot-collected data from *VisGel* [37].

to generate images by recursively denoising from a normal distribution to the desired data distribution. Specifically, given our latent representation $\mathbf{z}_I$, we uniformly sample a diffusion step $t \in \{1, ..., T\}$ and obtain the corresponding noisy image $\mathbf{z}_I^t$ by iteratively adding Gaussian noise with a variance schedule. We use a U-Net [51] network ϵ_θ as our denoising model, which is conditioned on the tactile representation encoded through the tactile encoder E_{ϕ_T} trained in Section 3.1.1. We minimize:

$$L(\theta, \phi) = \mathbb{E}_{\mathbf{z}_I, \mathbf{c}, \epsilon, t} \left[\|\epsilon_t - \epsilon_\theta(\mathbf{z}_I^t, t, E_{\phi_T}(\mathbf{v}_T))\|_2^2 \right], \quad (3)$$

where ϵ_t is the added noise at time t, and $\mathbf{v}_T$ is the tactile example. The denoising network ϵ_θ and the tactile encoder E_{ϕ_T} are jointly trained.

Inference. At test time, we first sample noise $\widetilde{\mathbf{z}}_I^T \sim \mathcal{N}(0, 1)$ at time T, and then use the trained diffusion model to iteratively predict the noise $\widetilde{\epsilon}_t$, resulting in a denoised latent representation $\widetilde{\mathbf{z}}_I^t = \widetilde{\mathbf{z}}_I^{t+1} - \widetilde{\epsilon}_{t+1}$ from $t \in \{T-1, ..., 0\}$. Following [50, 12], we use classifier-free guidance to trade off between sample quality and diversity in the conditional generation, computing the noise as:

$$\widetilde{\epsilon}_t = \epsilon_\theta(\widetilde{\mathbf{z}}_I^t, t, \emptyset) + s \cdot \left(\epsilon_\theta(\widetilde{\mathbf{z}}_I^t, t, E_{\phi_T}(\mathbf{v}_T)) - \epsilon_\theta(\widetilde{\mathbf{z}}_I^t, t, \emptyset) \right), \quad (4)$$

where $\emptyset$ denotes a zero-filled conditional example (for unconditional generation), and s is the guidance scale. Finally, we convert the latent representation $\widetilde{\mathbf{z}}_I^0$ to an image $\widetilde{\mathbf{x}}_I = \mathcal{D}(\widetilde{\mathbf{z}}_I^0) \in \mathbb{R}^{H \times W \times 3}$.

3.2. Visuo-Tactile Synthesis Models

So far, we have presented models for translating between touch and images (and vice versa). We now describe several visuo-tactile synthesis models that we build on this diffusion framework.

3.2.1 Generating realistic images without hands

One of the challenges of dealing with visuo-tactile data is that the tactile sensor typically occludes the object that is being touched (Fig. 3). Generated images will therefore contain the sensor, and potentially the arm that held it. This is not always desirable, as a major goal of touch sensing is to generate images of objects or materials that could have plausibly led to a given touch signal. We address this problem for the natural scenes from the *Touch and Go* dataset [63], which contain visible human hands and GelSight sensors [64].

To generate images containing only objects that yield a given tactile signal (without hands or touch sensors), we only compute the loss for pixels that do not overlap with hands during the training, thereby depriving the model of supervision for hand pixels. We first generate hand segmentation masks for the visual image $\mathbf{m}_I = \mathcal{S}(\mathbf{x}_I)$ and obtain the downsampled mask $\mathbf{z}_m$ of the same spatial dimension of the image latent representation. For this, we use the off-the-shelf hand segmentation model from Darkhalil et al. [11], which is a modified model from PointRend [29] instance segmentation designed specifically for segmenting hands. We then mask the diffusion loss (Eq. 6) to be:

$$\mathbb{E}_{\mathbf{z}_m, \mathbf{z}_I, \mathbf{c}, \epsilon, t} \left[\|\mathbf{z}_m \odot \left(\epsilon_t - \epsilon_\theta(\mathbf{z}_I^t, t, E_{\phi_T}(\mathbf{v}_T)) \right) \|_2^2 \right], \quad (5)$$

where $\mathbf{z}_m$ indicates whether a pixel overlaps with a hand, and $\odot$ denotes pointwise multiplication.

3.2.2 Tactile-driven Image Stylization

Tactile-driven image stylization [63] aims to manipulate the visual appearance of an object so that it looks more consistent with a given touch signal. Previous work posed the problem of editing the visual style of an image while preserving its structure [63, 36].

Given an input image $\mathbf{x}_I$ and a desired tactile signal $\mathbf{x}'_T$ (obtained from a different scene), our goal is to manipulate $\mathbf{x}_I$ so that it appears to "feels" more like $\mathbf{x}'_T$. We adapt the approach of Meng *et al*. [43]. We first compute the noisy latent representation z_I^N at time $0 \leq N \leq T$, where T denotes the total number of denoising steps. We then conduct the denoising process for z_I^N from time step N to 0 conditioned on $\mathbf{x}'_T$. This allows for fine-grained control over the amount of content preserved from the input image, via the parameter N. We analyze the choice of N at Sec. 4.6.

3.2.3 Tactile-driven Shading Estimation

Touch conveys a great deal of information about a surface's microgeometry [27]. Much of this information can also be perceived through *shading* cues: intensity variations due to light interacting with surface orientation for objects with Lambertian material properties. Following classic work in intrinsic image decomposition [2, 19, 3], we assume that the image can be factorized into reflectance and shading for each pixel, i.e., we can write our image $\mathbf{x}_I = \mathbf{x}_R \odot \mathbf{x}_S$ where the two terms in the product are the per-pixel reflectance and shading.

We propose a model that deals with inferring shading from touch. Given an image's estimated reflectance map

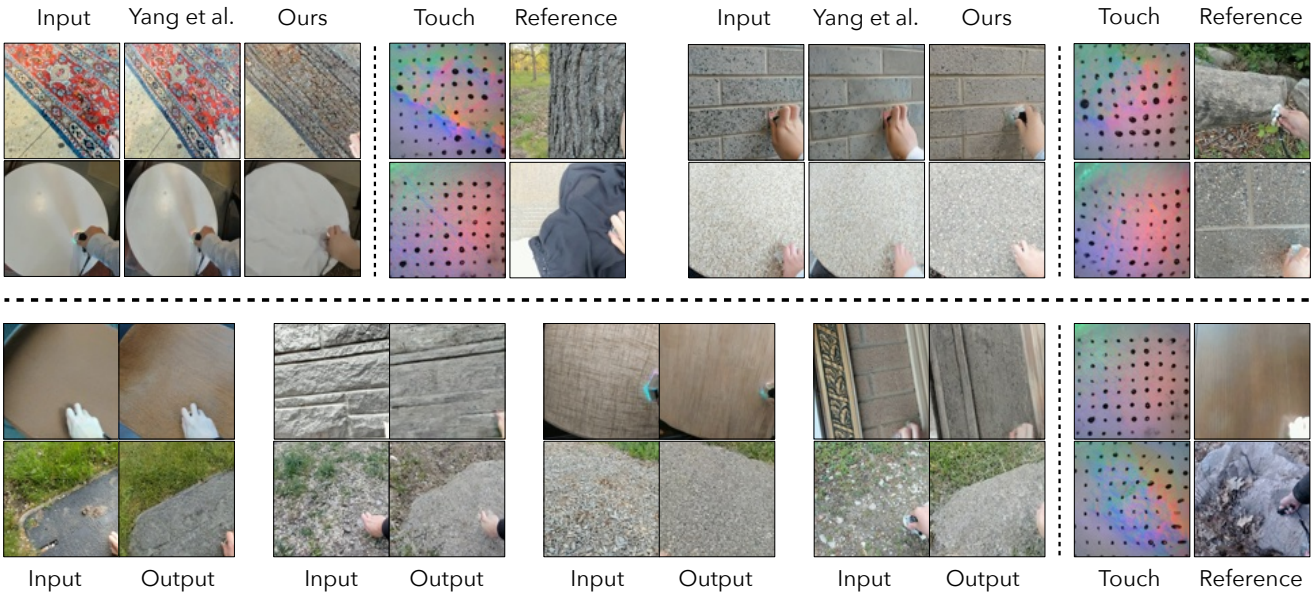


Figure 4: **Tactile-driven Image Stylization.** *(Top)* We restyle the input image using the given touch signal (reference image from scene provided for clarity). We compare our approach to Yang et al. [63]. Our approach generates images with higher quality matching more closely to the given tactile signal. *(Bottom)* We show more examples of the manipulated images. Please see supplement for more examples.

$\mathbf{x}_R$, along with a touch signal $\mathbf{x}_T$, we reconstruct the original image $\mathbf{x}_I$. This is a task that requires inferring the shading, since it is the component that is missing from the input. By formulating the problem so that we predict the original image, we can easily reuse the latent encoder/decoder from natural images.

We address this task by modifying our network so that it also takes reflectance as input (Eq. 6). We first estimate reflectance using the intrinsic image decomposition model of Liu *et al*. [40] and downsample it to the same dimensions as the latent space. We then concatenate the downsampled reflectance $\mathbf{z}_R$ to the noisy representation $\mathbf{z}_I^t$ as the input for each denoising step. Thus we modify the loss function (Eq. 6) as the following:

$$L(\theta, \phi) = \mathbb{E}_{\mathbf{z}_I, \mathbf{c}, \epsilon, t} \left[\|\epsilon_t - \epsilon_\theta(\mathbf{z}_I^t \otimes \mathbf{z}_R, t, E_{\phi_T}(\mathbf{v}_T))\|_2^2 \right], \quad (6)$$

where $\otimes$ denotes concatenation.

4. Results

We evaluate our cross-modal synthesis models through qualitative and quantitative experiments on natural scenes and robot-collected data.

4.1. Implementation details

Contrastive visuo-tactile model. Following [63], we use ResNet-18 as the backbone of contrastive model, and train on *Touch and Go* [63]. This model is trained using SGD for 240 epochs with the learning rate of 0.1 and weight decay of 10^{-4}. The ResNet takes 5 reference frames as input using early fusion (concatenated channel-wise) and we take the feature embedding from the last layer of the feature and map it to 512 dimensions. Following prior work [59], we use $\tau = 0.07$ and use a memory bank with 16,385 examples.

Visuo-tactile diffusion model. We base our latent diffusion model on Stable Diffusion [50]. We use the Adam optimizer with the base learning rate of 2×10^{-6}. Models are all trained with 30 iterations using the above learning rate policy. We train our model with the batch size of 96 on 4 RTX A40 GPUs. The conditional model is finetuned along with the diffusion model. We use the frozen, pretrained VQ-GAN [13] to obtain our latent representation, with the spatial dimension of 64×64. During the inference, we conduct denoising process for 200 steps and set the guidance scale $s = 7.5$.

4.2. Experimental Setup

Dataset. We conduct our experiments on two real-world visuo-tactile datasets:

- ***Touch and Go* dataset.** The *Touch and Go* dataset is a recent, real-world visuo-tactile dataset in which humans probe a variety of objects in both indoor and outdoor scenes. There are 13,900 touches from roughly 4000 different object instances and 20 material categories. Since this is the only available dataset with zoomed-in images and clearly visible materials, we use it for all three tasks.
- ***VisGel* dataset.** The *VisGel* dataset contains synchronized videos of a robot arm equipped with a GelSight sensor interacting with 195 household objects. The dataset includes 195 objects from a wide range of indoor scenes of food items, tools, kitchen items, to fabrics and stationery. In total, the dataset contains 12k touches and around 3M frames.

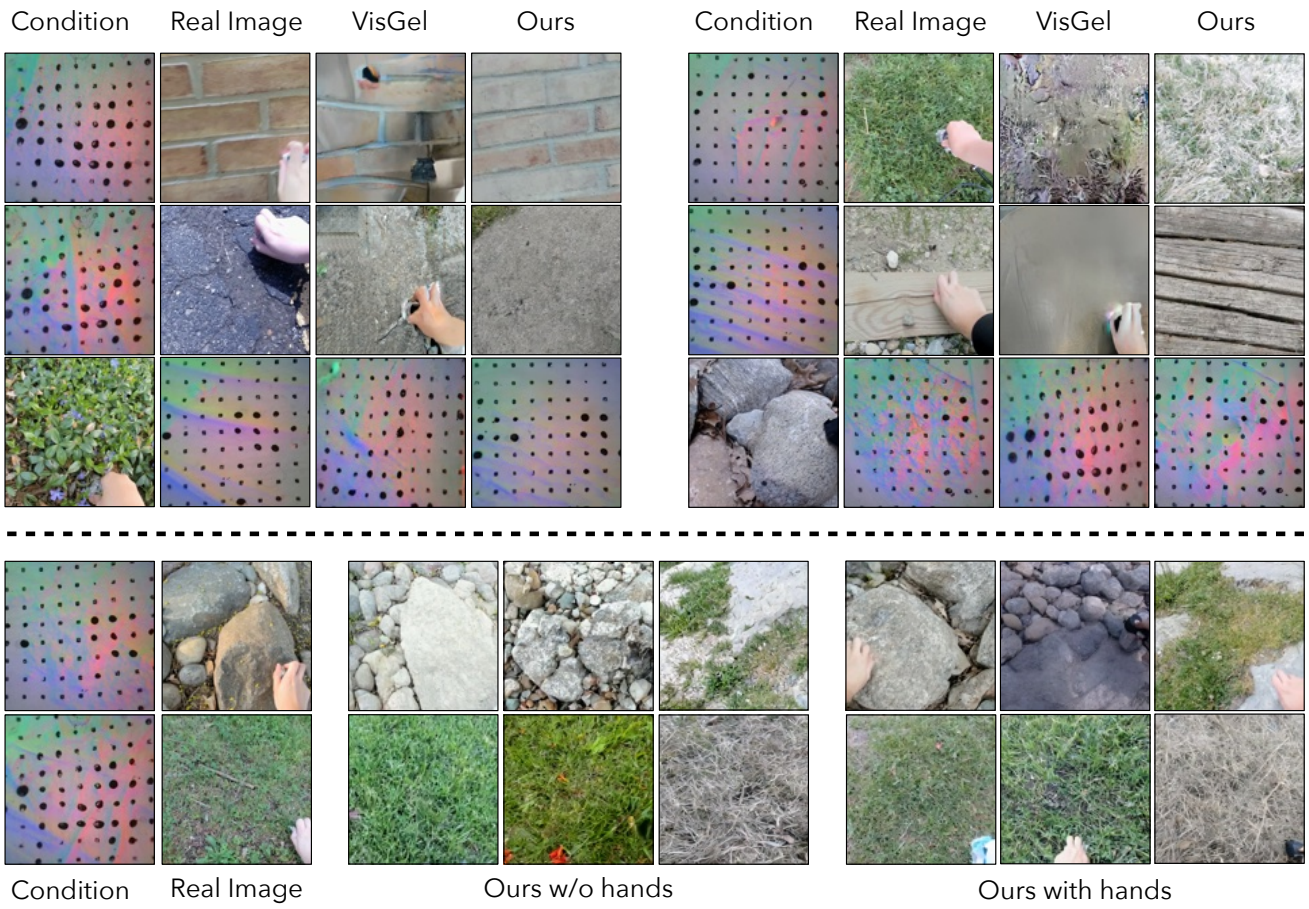


Figure 5: **Visuo-tactile Cross Generation on *Touch and Go* dataset.** *(Top)* We compare our approach to state-of-the-art method Visgel [37]. *(Bottom)* We show more results of our generated images with and without hands. In both case our approach is able to generate realsitic images with high fidelity.

Evaluation metrics. We use several quantitative metrics to evaluate the quality of our generated images or tactile signals. We use **Frechet Inception Distance (FID)**, which compares the distribution of real and generated image activations using trained network. Following Yang *et al*. [63] and CLIP [48], we take the cosine similarity between our learned visual and tactile embeddings for the generated images and conditioned tactile signals, a metric we call **Contrastive Visuo-Tactile Pre-Training (CVTP)**. A higher score indicates a better correlation between touch and images. It is worth noting that the CVTP metric only takes one frame of touch input. Following [63], we measure **Material Classification Consistency**: we use the material classifier from Yang *et al*. [63] to categorize the predicted and ground truth images, and measure the rate at which they agree. Finally, following [16], we evaluate standard **Structural Similarity Index Measure (SSIM)** and **Peak Signal to Noise Ratio (PSNR)** [60] metrics.

4.3. Cross-modal Generation

We perform cross-modal generation, *i.e*., generating an image from touch and vice versa, on both in-the-wild *Touch and Go* dataset and robot-collected dataset *VisGel*. For straightforward comparison to prior work [37], on *VisGel* we provide a *reference* photo of the scene as an input to the model. Thus, successfully predicting the ground truth image amounts to inserting imagery of the robotic arm to the correct location in the scene. For *Touch and Go*, we do not condition the model on a visual input: instead, we simply translate one modality to the other.

For evaluation metrics, we use CVTP, material classification consistency, and FID score for touch-to-image generation and SSIM and PSNR for image-to-touch generation. For *VisGel* dataset we leverage SSIM and PSNR as the evaluation metric for both tasks. We only use CVTP, material classification consistency and FID only on touch-to-image generation task on *Touch and Go*, since these evaluation metrics rely on a pretrained neural network from datasets of natural images, which may not generalize well on a different modality or to robot-collected data.

We compare our model to the prior state-of-the-art visuo-tactile generation method [37], which is adapted from pix2pix [23] and is specifically designed to bridge the large

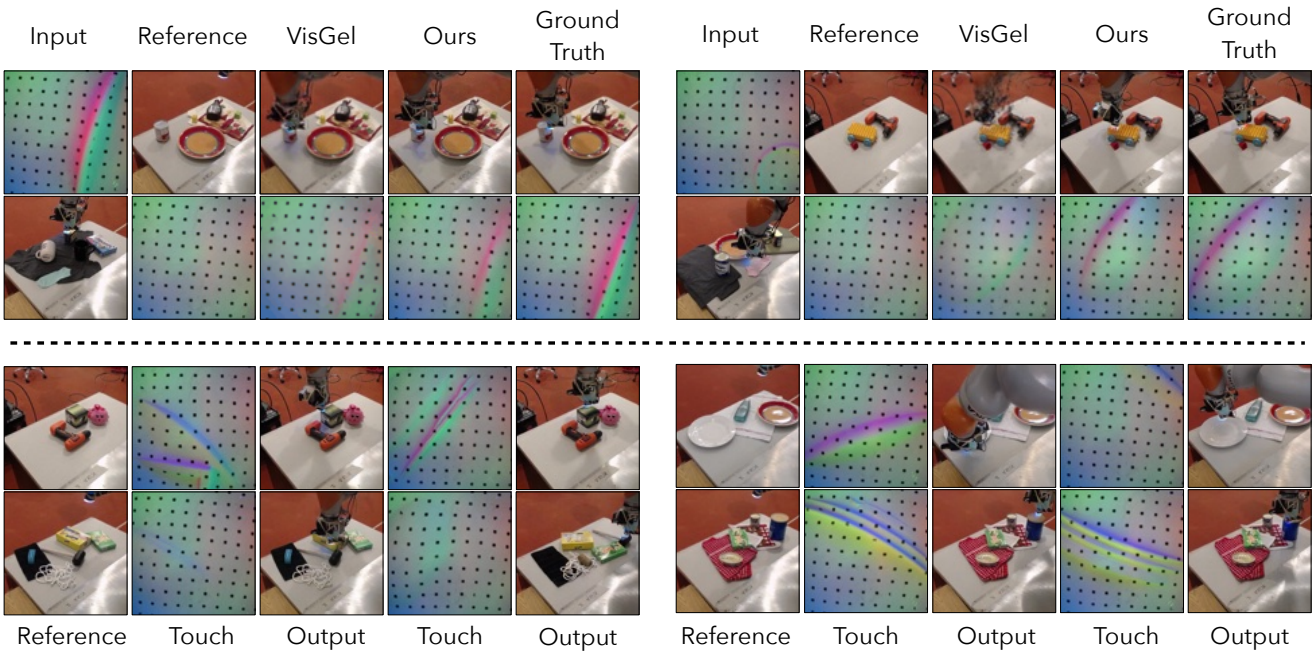


Figure 6: **Visuo-tactile Cross Generation on *VisGel* dataset.** *(Top)* We compare our approach to state-of-the-art method VisGel [37]. *(Bottom)* Our approach is able to generate robotic hands touching reasonable locations of objects given the same reference image but different tactile signals.

Table 1: Evalutation of cross-modal generation on *Touch and Go*.

Method	Touch → Image			Image → Touch	
	CVTP (↑)	Material(↑)	FID(↓)	SSIM(↑)	PSNR(↑)
Pix2Pix [23]	0.08	0.15	136.4	0.43	14.3
VisGel [37]	0.07	0.15	128.3	0.45	15.0
Ours w/ hands	**0.12**	0.22	**48.7**	**0.50**	**15.4**
Ours w/o hands	**0.12**	**0.24**	81.5	**0.50**	**15.4**

Table 2: Evaluation of cross-modal generation on *VisGel* (and conditioning on another photo from the scene).

Method	Touch → Image		Image → Touch	
	SSIM(↑)	PSNR(↑)	SSIM (↑)	PSNR (↑)
Pix2Pix [23]	0.50	15.1	0.71	20.7
VisGel [37]	0.59	17.9	0.76	26.2
Ours	**0.76**	**21.5**	**0.85**	**27.6**

domain gap between modalities by adding a reference image and temporal condition. As it is not possible to find a reference image in the natural image dataset, we remove the reference image while keeping everything else the same.

We show quantitative results for both tasks on *Touch and Go* and *VisGel* in Table 1 and Table 2 respectively. Our methods outperform existing state-of-the-art methods by a large margin for all evaluation metrics. We note that the variation of our model that removes hands from images obtains a worse FID score compared to those with hands, due to the discrepancy of hands between the original dataset and our generated images. Interestingly, the presence of hands does not does not affect the performance of CVTP and material classification consistency. We provide qualitative results from both models in Figure 5 *(bottom)*.

4.4. Tactile-Driven Image Stylization

Following [63], we evaluate the performance of tactile-driven image stylization on *Touch and Go* [63] using CVTP and material classification metrics. We also calculate the FID score between the set of generated images and the set of real images associated with the given tactile signals, which measures the fidelity of the output. We compare our model to a modified version of CycleGAN [67] and the state-of-the-art method of Yang et al. [63]. From the quantitative comparisons in Table 3, our method demonstrates a significant improvement over existing methods. We also show qualitative comparisons in Figure 3, where the generated images more closely match the tactile signal, and we are able to generate styles that existing methods fail to capture.

Table 3: Quantitative results of of tactile-driven image stylization.

Method	Evaluation Metrics		
	CVTP (↑)	Material (↑)	FID (↓)
Cycle GAN [67]	0.09	0.15	24.6
Yang et al. [63]	0.10	0.20	22.5
Ours	**0.13**	**0.22**	**15.8**

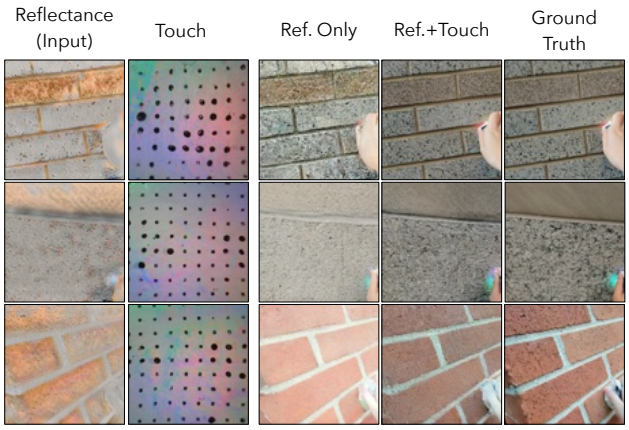


Figure 7: **Tactile-driven shading estimation.** We compare our approach to a model without a tactile signal (only reflectance), finding that the tactile-driven model better captures subtle material properties, such as roughness.

Table 4: Quantitative results for tactile-driven shading estimation.

Method	Reflectance → Image		
	SSIM(↑)	PSNR(↑)	FID(↓)
Touch Only	0.27	11.6	48.7
Reflectance Only	0.46	14.5	40.7
Reflectance + Touch	**0.48**	**15.4**	**36.9**

4.5. Tactile-driven Shading Estimation

We hypothesize that the tactile signal conveys information about the microgeometry of an image, and thus allows a model to produce more accurate images than a reflectance-to-image model that does not have access to touch. We evaluated both models on *Touch and Go* (Table 4) and found that adding touch indeed improves performance on all evaluation metrics. We also show qualitative comparisons in Figure 7. We found that tactile signals are especially informative for predicting roughness and smoothness of Lambertian surfaces, such as bricks.

4.6. Analysis

Importance of temporal information. We first study the effect of adding multiple GelSight frames to the contrastive visuo-tactile embedding (Figure 9). We compare our method with the unconditional generation and material class conditional generation on *Touch and Go*. We found that conditioned generation provides a large improvement in performance compared to the unconditional generation. We also observed that the generation conditioned on the pre-trained model is significantly better than that without pre-training. Interestingly, the model conditioned on the material class outperforms the variation of the model that only observes a single GelSight frame, suggesting that perceiving a touch signal from only a single moment in time may be less informative than the material category. Providing the model with additional frames significantly improves the model, with the 5-frame model obtaining the overall best performance.

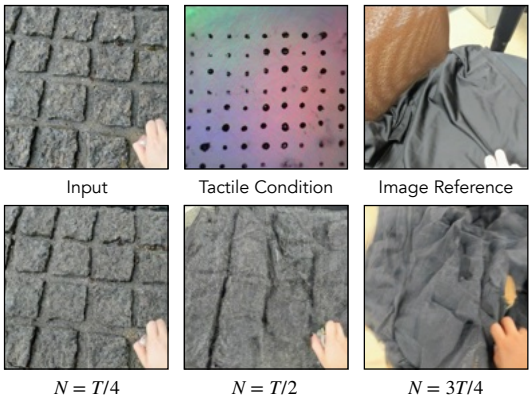


Figure 8: **Controlling the amount of preserved image content.** Manipulated images of tactile-driven image stylization using different values of N.

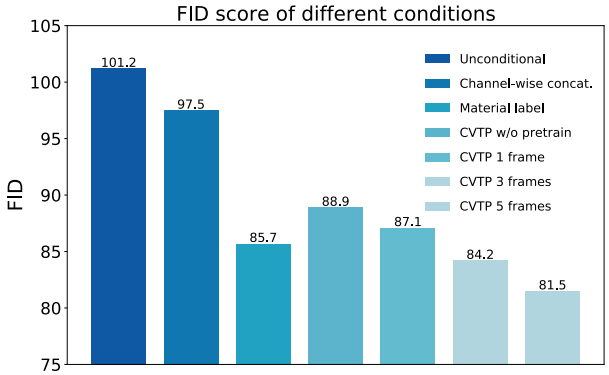


Figure 9: **Effect of different types of tactile conditioning**.

Controllable Image Stylization Our method allows us to control over the amount of image content that is preserved from the original image by changing the denoising staring point N (Sec. 3.2.2) [43]. From Figure 8, we observe that if we select the larger N, the generated image will be changed more drastically where the visual appearance will be changed completely to match the tactile signal while ruining the original image structure. In extreme case, where $N = T$ the manipulated result will be equal to the touch-to-image generation result, while small N will result in little overall change. We empirically found that selecting $N = T/2$ obtains a good trade-off between these factors.

5. Conclusion

We proposed a visuo-tactile diffusion model that unifies previous cross-modal synthesis tasks, and allows us to address novel problems. We are the first to generate realistic images in the natural scenes from touch (and vise versa) without any image-based conditioning. We also show the ability to generate realistic "hand-less" images and solve a

novel tactile-driven shading estimation task. Finally, we obtain significantly more realistic results on the tactile-driven stylization task than prior work. We see our work as being a step toward integrating the fields of tactile sensing and generative modeling.

Limitations. Since our work has applications in creating fake imagery, a potential issue is that it could be used to create disinformation. Also, as touch mainly conveys material properties and microgeometry, the generated image will often differ semantically from the ground truth.

Acknowledgements. We thank Chao Feng, Ziyang Chen and Shaokai Wu for the helpful discussions and help for visualizations. This work was supported in part by Cisco Systems.

References

[1] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18187–18197, 2021. 2

[2] Harry Barrow, J Tenenbaum, A Hanson, and E Riseman. Recovering intrinsic scene characteristics. *Comput. vis. syst*, 2(3-26):2, 1978. 4

[3] Sean Bell, Kavita Bala, and Noah Snavely. Intrinsic images in the wild. *ACM Trans. on Graphics (SIGGRAPH)*, 33(4), 2014. 2, 4

[4] Roberto Calandra, Andrew Owens, Manu Upadhyaya, Wenzhen Yuan, Justin Lin, Edward H Adelson, and Sergey Levine. The feeling of success: Does touch sensing help predict grasp outcomes? *Conference on Robot Learning (CoRL)*, 2017. 2

[5] Arkadeep Narayan Chaudhury, Timothy Man, Wenzhen Yuan, and Christopher G Atkeson. Using collocated vision and tactile sensors for visual servoing and localization. *IEEE Robotics and Automation Letters*, 7(2):3427–3434, 2022. 2

[6] Zehua Chen, Xu Tan, Ke Wang, Shifeng Pan, Danilo P. Mandic, Lei He, and Sheng Zhao. Infergrad: Improving diffusion models for vocoder by considering inference in training. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8432–8436, 2022. 2

[7] Ze Chen, Yihan Wu, Yichong Leng, Jiawei Chen, Haohe Liu, Xuejiao Tan, Yang Cui, Ke Wang, Lei He, Sheng Zhao, Jiang Bian, and Danilo P. Mandic. Resgrad: Residual denoising diffusion probabilistic models for text to speech. *ArXiv*, abs/2212.14518, 2022. 2

[8] Shin-I Cheng, Yu-Jie Chen, Wei-Chen Chiu, Hung-Yu Tseng, and Hsin-Ying Lee. Adaptively-realistic image generation from stroke and sketch with diffusion model. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2023. 2

[9] James J Cox, Frank Reimann, Adeline K Nicholas, Gemma Thornton, Emma Roberts, Kelly Springell, Gulshan Karbani, Hussain Jafri, Jovaria Mannan, Yasmin Raashid, et al. An scn9a channelopathy causes congenital inability to experience pain. *Nature*, 444(7121):894–898, 2006. 1

[10] Mark R. Cutkosky, Robert D. Howe, and William R. Provancher. Force and tactile sensors. In *Springer Handbook of Robotics*, 2008. 2

[11] Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, Sanja Fidler, David Fouhey, and Dima Damen. Epic-kitchens visor benchmark: Video segmentations and object relations. In *Proceedings of the Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*, 2022. 4

[12] Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat GANs on image synthesis. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. 2, 4

[13] Mucong Ding, Kezhi Kong, Jingling Li, Chen Zhu, John P. Dickerson, Furong Huang, and Tom Goldstein. Vq-gnn: A universal framework to scale up graph neural networks using vector quantization. In *Neural Information Processing Systems*, 2021. 5

[14] Chao Feng, Ziyang Chen, and Andrew Owens. Self-supervised video forensics by audio-visual anomaly detection. *Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[15] Ruohan Gao, Yen-Yu Chang, Shivani Mall, Li Fei-Fei, and Jiajun Wu. Objectfolder: A dataset of objects with implicit visual, auditory, and tactile representations. In *CoRL*, 2021. 2

[16] Ruohan Gao, Yiming Dou, Hao Li, Tanmay Agarwal, Jeannette Bohg, Yunzhu Li, Li Fei-Fei, and Jiajun Wu. The objectfolder benchmark: Multisensory learning with neural and real objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17276–17286, June 2023. 2, 6

[17] Ruohan Gao, Zilin Si, Yen-Yu Chang, Samuel Clarke, Jeannette Bohg, Li Fei-Fei, Wenzhen Yuan, and Jiajun Wu. Objectfolder 2.0: A multisensory object dataset for sim2real transfer. In *CVPR*, 2022. 2

[18] Sang gil Lee, Heeseung Kim, Chaehun Shin, Xu Tan, Chang Liu, Qi Meng, Tao Qin, Wei Chen, Sung-Hoon Yoon, and Tie-Yan Liu. Priorgrad: Improving conditional denoising diffusion models with data-dependent adaptive prior. In *International Conference on Learning Representations*, 2021. 2

[19] Roger Grosse, Micah K Johnson, Edward H Adelson, and William T Freeman. Ground truth dataset and baseline evaluations for intrinsic image algorithms. In *2009 IEEE 12th International Conference on Computer Vision*, pages 2335–2342. IEEE, 2009. 4

[20] Jonathan Ho, Ajay Jain, and P. Abbeel. Denoising diffusion probabilistic models. 2020. 2

[21] Jonathan Ho, Chitwan Saharia, William Chan, David J. Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *J. Mach. Learn. Res.*, 23:47:1–47:33, 2021. 2

[22] Fabian Hutmacher. Why is there so much more research on vision than on any other sensory modality? *Frontiers in psychology*, 10:2246, 2019. 1

[23] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *CVPR*, 2017. 2, 6, 7

[24] Wei Ji, Long Chen, Yinwei Wei, Yiming Wu, and Tat-Seng Chua. Mrtnet: Multi-resolution temporal network for video sentence grounding. *arXiv preprint arXiv:2212.13163*, 2022. 2

[25] Wei Ji, Xi Li, Fei Wu, Zhijie Pan, and Yueting Zhuang. Human-centric clothing segmentation via deformable semantic locality-preserving network. volume 30, pages 4837–4848. IEEE, 2019. 3

[26] Wei Ji, Xiangyan Liu, An Zhang, Yinwei Wei, and Xiang Wang. Online distillation-enhanced multi-modal transformer for sequential recommendation. In *Proceedings of the 31th ACM international conference on Multimedia*, 2023. 2

[27] Micah K Johnson and Edward H Adelson. Retrographic sensing for the measurement of surface texture and shape. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1070–1077. IEEE, 2009. 2, 4

[28] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Hui-Tang Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. *ArXiv*, abs/2210.09276, 2022. 2

[29] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross B. Girshick. Pointrend: Image segmentation as rendering. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9796–9805, 2019. 4

[30] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. *ArXiv*, abs/2009.09761, 2020. 2

[31] Max W. Y. Lam, Jun Wang, Dan Su, and Dong Yu. BDDM: Bilateral denoising diffusion models for fast and high-quality speech synthesis. In *International Conference on Learning Representations*, 2022. 2

[32] Susan J. Lederman and Roberta L. Klatzky. Hand movements: A window into haptic object recognition. *Cognitive Psychology*, 19:342–368, 1987. 2

[33] Susan J. Lederman and R. L. Klatzky. Tutorial review haptic perception: A tutorial. 2009. 2

[34] Yichong Leng, Zehua Chen, Junliang Guo, Haohe Liu, Jiawei Chen, Xu Tan, Danilo P. Mandic, Lei He, Xiang-Yang Li, Tao Qin, Sheng Zhao, and Tie-Yan Liu. Binauralgrad: A two-stage conditional diffusion probabilistic model for binaural audio synthesis. *ArXiv*, abs/2205.14807, 2022. 2

[35] Nathan F. Lepora, Yijiong Lin, Ben Money-Coomes, and John Lloyd. Digitac: A digit-tactip hybrid tactile sensor for comparing low-cost high-resolution robot touch. *IEEE Robotics and Automation Letters*, 7:9382–9388, 2022. 2

[36] Tingle Li, Yichen Liu, Andrew Owens, and Hang Zhao. Learning visual styles from audio-visual associations. In *ECCV*, 2022. 4

[37] Yunzhu Li, Jun-Yan Zhu, Russ Tedrake, and Antonio Torralba. Connecting touch and vision via cross-modal prediction. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10601–10610, 2019. 1, 2, 4, 6, 7

[38] Justin Lin, Roberto Calandra, and Sergey Levine. Learning to identify object instances by touch: Tactile recognition via multimodal matching. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 3644–3650. IEEE, 2019. 1, 2

[39] David J Linden. *Touch: The science of the hand, heart, and mind*. Penguin Books, 2016. 1

[40] Yunfei Liu, Yu Li, Shaodi You, and Feng Lu. Unsupervised learning for intrinsic image decomposition from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3248–3257, 2020. 2, 5

[41] Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021. 2

[42] Paul R Manske. The sense of touch. *Journal of Hand Surgery*, 24(2):213–214, 1999. 1

[43] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022. 2, 4, 8

[44] Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. *ICML*, abs/2102.09672, 2021. 2

[45] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*, 2021. 2

[46] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3

[47] Brian O'Shaughnessy. The sense of touch. *Australasian journal of philosophy*, 67(1):37–58, 1989. 1

[48] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3, 6

[49] Tanzila Rahman, Hsin-Ying Lee, Jian Ren, S. Tulyakov, Shweta Mahajan, and Leonid Sigal. Make-a-story: Visual memory conditioned consistent story generation. *ArXiv*, abs/2211.13319, 2022. 2

[50] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, 2022. 2, 3, 4, 5

[51] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, 2015. 4

[52] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, Seyedeh Sara Mahdavi, Raphael Gontijo Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. *ArXiv*, abs/2205.11487, 2022. 2

[53] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence*, PP, 2021. 2

[54] Ruizhi Shao, Zerong Zheng, Hongwen Zhang, Jingxiang Sun, and Yebin Liu. Diffustereo: High quality human reconstruction via diffusion-based stereo using sparse cameras. In *ECCV*, 2022. 2

[55] Uriel Singer, Adam Polyak, Thomas Hayes, Xiaoyue Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. *ArXiv*, abs/2209.14792, 2022. 2

[56] Abhishek Sinha, Jiaming Song, Chenlin Meng, and Stefano Ermon. D2c: Diffusion-denoising models for few-shot conditional generation. *ArXiv*, abs/2106.06819, 2021. 2

[57] Linda Smith and Michael Gasser. The development of embodied cognition: Six lessons from babies. *Artificial life*, 2005. 1

[58] Ian Taylor, Siyuan Dong, and Alberto Rodriguez. Gelslim 3.0: High-resolution measurement of shape, force and slip in a compact tactile-sensing finger. *2022 International Conference on Robotics and Automation (ICRA)*, pages 10781–10787, 2021. 2

[59] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European conference on computer vision*, pages 776–794. Springer, 2020. 3, 5

[60] L.-T. Wang, Nathan E. Hoover, Edwin H. Porter, and John J. Zasio. Ssim: A software levelized compiled-code simulator. *24th ACM/IEEE Design Automation Conference*, pages 2–8, 1987. 6

[61] Akihiko Yamaguchi and Christopher G Atkeson. Implementing tactile behaviors using fingervision. In *2017 IEEE-RAS 17th International Conference on Humanoid Robotics (Humanoids)*, pages 241–248. IEEE, 2017. 2

[62] Fengyu Yang and Chenyang Ma. Sparse and complete latent organization for geospatial semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1809–1818, 2022. 3

[63] Fengyu Yang, Chenyang Ma, Jiacheng Zhang, Jing Zhu, Wenzhen Yuan, and Andrew Owens. Touch and go: Learning from human-collected vision and touch. *Neural Information Processing Systems (NeurIPS) - Datasets and Benchmarks Track*, 2022. 1, 2, 3, 4, 5, 6, 7

[64] Wenzhen Yuan, Siyuan Dong, and Edward H. Adelson. Gelsight: High-resolution robot tactile sensors for estimating geometry and force. *Sensors (Basel, Switzerland)*, 17, 2017. 2, 4

[65] Wenzhen Yuan, Chenzhuo Zhu, Andrew Owens, Mandayam A Srinivasan, and Edward H Adelson. Shape-independent hardness estimation using deep learning and a gelsight tactile sensor. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 951–958. IEEE, 2017. 2, 3

[66] Chenhao Zheng, Ayush Shrivastava, and Andrew Owens. Exif as language: Learning cross-modal associations between images and camera metadata. *Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[67] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017. 7