

Controlling the World by Sleight of Hand

Sruthi Sudhakar, Ruoshi Liu, Basile Van Hoorick, Carl Vondrick, and Richard Zemel

Columbia University
coshand.cs.columbia.edu

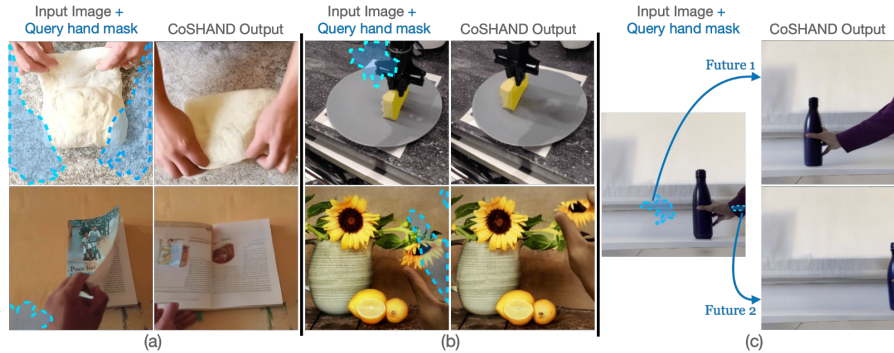


Fig. 1: **CoSHAND** synthesizes an image of a future after a specific interaction (dotted blue mask) has occurred. In (a) we demonstrate **CoSHAND**'s ability to perform complex manipulations on deformable objects such as kneading dough and opening a book. In (b) we show generalization to robot gripper interactions. In (c) we show it is possible to generate diverging futures given the same input context but different hand controls.

Abstract. Humans naturally build mental models of object interactions and dynamics, allowing them to imagine how their surroundings will change if they take a certain action. While generative models today have shown impressive results on generating/editing images unconditionally or conditioned on text, current methods do not provide the ability to perform object manipulation conditioned on actions, an important tool for world modeling and action planning. Therefore, we propose to learn an action-conditional generative models by learning from unlabeled videos of human hands interacting with objects. The vast quantity of such data on the internet allows for efficient scaling which can enable high-performing action-conditional models. Given an image, and the shape/location of a desired hand interaction, **CoSHAND**, synthesizes an image of a future after the interaction has occurred. Experiments show that the resulting model can predict the effects of hand-object interactions well, with strong generalization particularly to translation, stretching, and squeezing interactions of unseen objects in unseen environments. Further, **CoSHAND** can be sampled many times to predict multiple possible effects, modeling the uncertainty of forces in the interaction/environment. Finally, method generalizes to different embodiments, including non-human hands, i.e. robot hands, suggesting that generative video models can be powerful models for robotics.

1 Introduction

Humans and animals have an impressive ability to mentally simulate what would happen to an object if they were to interact with it [7]. Our ability to reason about our actions and the resulting dynamics of the objects around us has developed from years of interacting with our environment with our hands. Such mental models and forecasting allow us to simulate realistic experiences depending on a specific interactions. Enabling machines to similarly model the future based on interactions can be useful in a variety of tasks such as robotic planning and augmented or virtual reality.

Current generative models are often conditioned on modalities that are easy to acquire at scale, such as text [20, 38, 44]. However, these conditioning modalities are not aligned to the action space to enable machines to predict how objects and the environment change and potentially deform in physically plausible ways according to a *specific* interaction. For example, even with a detailed text conditioning of “*squeezing a pillow so that it deforms horizontally*”, the precise direction and distance of the deformation is hard to capture via text (see Fig. 2). The key question is, how do we give machines the ability to imagine interactions with their environment?

In this work, we propose to learn an action-conditional generative model by leveraging large amounts of unlabeled videos of people interacting with objects using their hands. Given an image, and the shape/location of a desired hand interaction, **CoSHAND**, (**C**ontrolling the **W**orld by **S**leight of **H**and), synthesizes an image of a future after the interaction has occurred.

We use existing off-the-shelf hand segmentation methods to obtain binary hand masks which we can then pair with before and after frames of a video. By using automatic segmentation methods and unlabeled videos (180k examples of real-world hand-object interactions from the SomethingSomethingv2 dataset), we can efficiently obtain a large-scale dataset, highlighting the potential to efficiently scale up the approach to achieve stronger performance and better generalization. Furthermore, using binary masks allows the method to be agent-agnostic enabling capabilities for non-human agents. We fine-tune **CoSHAND** from a pre-trained image diffusion model. Image generative models (such as DALL-E-2 [31]) have seen billions of images of hand-objects pairs in a variety of configurations, allowing us to leverage its strong priors about how objects and hands move with respect to each other. Furthermore, the probabilistic generative nature of diffusion models allows **CoSHAND** to model uncertainty in future states due to environment/interaction forces.

We perform several quantitative and qualitative experiments on our training dataset and an unseen (In-the-wild) test set that we create. **CoSHAND** can predict the effects of hand-object interactions well. A key achievement is the generalization ability to novel objects and scenes not in the training set. Particularly we see strong results on interactions involving translations, stretching, and compressing, where objects are clear, separable from surrounding objects, and in-frame in the input image. The model performs decently on more complex object interactions such as large rotations and precise deformations involving folding. Interestingly,

our method can also predict the effects of a robot gripper’s interactions, even though it has not been trained on robot data; a promising direction towards robotic planning. Finally, we benchmark against various unconditional and text-conditional generative models to show the benefits of our proposed approach of controlling by hand.

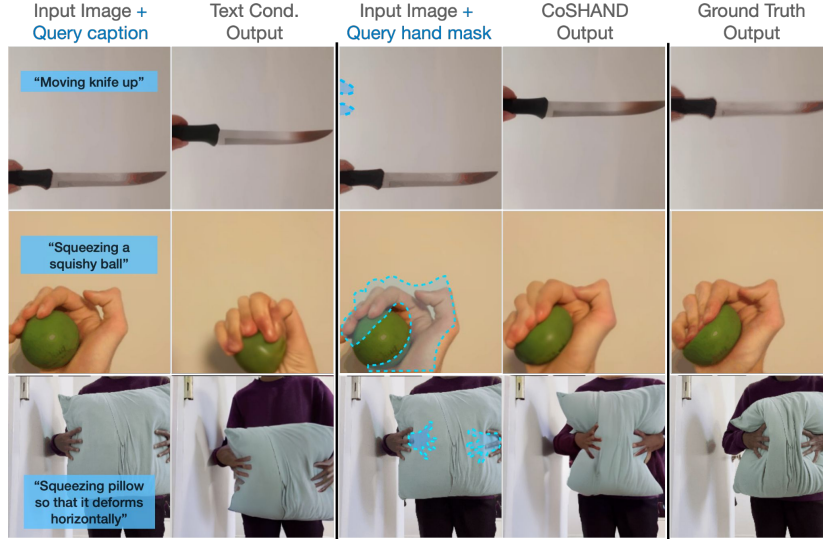


Fig. 2: We show that text-conditioning is insufficient to model interactions, whereas hands allow for better control. Columns 1 & 2 show the input image, query caption, and output of text conditional generation. Columns 3 & 4 show the input image, query hand mask, and output of CoSHAND. Column 5 shows the ground truth output. Notice that CoSHAND is able to achieve precise control (including the exact final location of the knife in row 1 and the precise squeezing motion in rows 2 & 3) which results in a output that is more consistent with the ground truth.

2 Related Work

2.1 Diffusion Models

Diffusion models have gained tremendous momentum in generative modeling in computer vision. Compared to previous generative architectures such as VAEs and GANs, diffusion models offer many advantages including better training stability, improved coverage of multi-modal distribution of training data, and scalability. Prior works [12] have shown strong experimental results that demonstrates diffusion model beats GAN [15] in image synthesis quality. In DDPM [21], authors propose an effective implementation of diffusion model for image

generation by iteratively denoising an image composed of Gaussian noise. Latent Diffusion Model (LDM) [33] proposed an efficient diffusion architecture that performs denoising in latent space of a VAE [32] instead of pixel space. Due to its fully open-source release, Stable Diffusion has been used as a strong image prior to solve many challenging task in computer vision such as semantic image editing [4, 14, 35], 3D [10, 26, 43], and segmentation [1, 28]. In this work, we leverage the strong visual priors provided by the internet-pretrained diffusion model to solve the zero-shot hand-conditioned interaction understanding problem.

2.2 Conditional Generation

While large-scale diffusion models have recently shown incredible capabilities in generating images and video scenes, methods to control these models are still emerging. Such control is imperative in using these models for extracting plausible world dynamics and making them useful for various applications. Several works have aimed to tackle this control problem from a variety of angles such as text-conditioning by manipulating the weights of the attention maps [4, 19], camera-view conditioning like in [26], and image based conditioning with a separate set of additional weights like in [23] or [50]. Different from a normal text conditioned prediction models, world model [17] is a class of models that predict the future conditioned on past observation as well as an action. World models have been applied in robotics for manipulation [9, 13, 18, 44], locomotion [18, 42], and planning [45]. We propose a unique approach of utilizing hand interaction based image conditioning to predict future states by learning from a large-scale dataset of human videos and generalizing to other embodiments.

2.3 Hand Environment Interaction

Hand-environment interaction forms a significant percentage of embodied human experience, making it an important area of study in computer vision and robotics. Prior works have widely explored this topic from various angles. A line of work in computer vision extensively studies the problem of 3D reconstruction of human hands and objects from images [5, 37, 46–48]. Another line of work investigates affordances of objects [2, 11, 27, 29, 30, 49]. Instead of recognition and reconstruction, many methods propose synthesizing hand-object grasp [3, 8, 22, 25, 49]. Other works incorporate physics simulation to generate dynamic grasps for objects that are physically plausible [6]. Different from prior work, our proposed method focuses on predicting the changes in object state such as position, appearance, and geometry (deformable objects) caused by hand motions, from a single view.

3 Method

We present our method to solve the task of generating an image depicting a future state of the scene after a certain hand interaction has occurred.

3.1 Learning From Hands at Scale

Recent large-scale models have shown impressive results on tasks such as image generation, visual question answering, etc. These can largely be attributed to the scale of data with which they have been trained on. For example, Stable Diffusion [34] was trained on the LAION-5B dataset [36] which likely includes numerous examples of people interacting with objects in diverse scenarios. We hypothesize that the model understands something about the state of objects based on a certain hand position. Furthermore, objects are often manipulated by hands in the real world, thus Stable Diffusion has likely seen more objects in different states (such as laptops ‘opened’ and ‘closed’) when the hand is present as well. Therefore the key question is how to unlock control to enable manipulation between states (such as opening a laptop). Allowing the model to understand the relationship between hand motions and object states would enable interaction.

We aim to leverage these priors by using Stable Diffusion as a starting point and finetuning a large-scale pretrained image-based model on video, as video datasets provide an easy way to extract before and after states of the world to obtain the affects of the applied interaction. Not only does this give us strong priors about how hands and objects interact, but it also allows us to generalize to objects and scenes beyond the data distribution we train on. Furthermore, the priors learned from these hand-object interactions can be transferred to other embodiments such as robot arms as seen in Sec 4.4, enabling generalization beyond hands. We emphasize that **CoSHAND** can be easily scaled to larger video datasets because we use off-the shelf models to represent hands, therefore there are no metadata/annotation requirements.

3.2 Problem Formulation

We introduce **CoSHAND**, and show an overview in Fig. 3. Given an RGB image $x_t \in \mathbb{R}^{H \times W \times 3}$, the corresponding binary hand-mask $h_t \in \mathbb{R}^{H \times W}$ which marks the pixels belonging to the hand in the input image, and a query hand-mask $h_{t+1} \in \mathbb{R}^{H \times W}$ which marks an action taken, our goal is to learn a function f such that

$$f(x_t, h_t, h_{t+1}) = \hat{x}_{t+1}$$

where $\hat{x}_{t+1} \in \mathbb{R}^{H \times W \times 3}$ is the estimated image and should be perceptually similar to the true but unobserved future x_{t+1} .

3.3 Conditioning Diffusion Models on Context

Image diffusion models have become increasingly popular due to the impressive quality of produced images and the ease and stability of training in comparison to GANs. In DDPM [21], a model is trained to learn to reverse the diffusion process, which gradually denoises data to produce an image. In latent diffusion models [33], rather than learning to denoise images in the pixel space, the model is

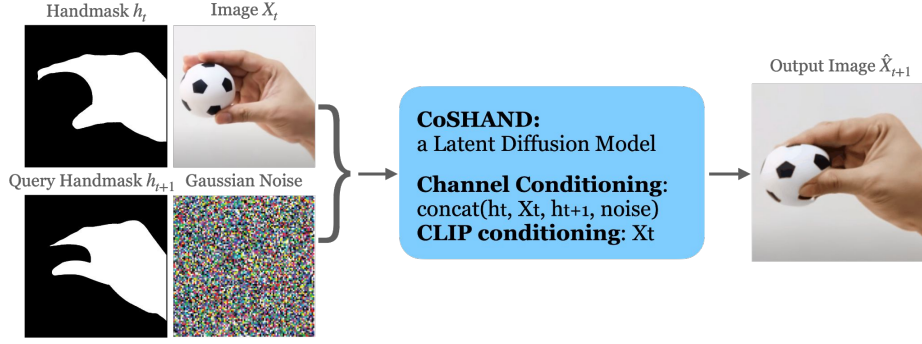


Fig. 3: CoSHAND Method. We propose a novel approach of controlling by hands to enable manipulating objects in an image. Given an image, the corresponding hand mask, and a query hand mask of the desired interaction, **CoSHAND** synthesizes an image with the interaction applied. Such visual conditioning allows for object interaction.

trained to denoise in the latent space of a VAE as this reduces the computational complexity, and increases the expressiveness of the latent code that is learned.

Similar to LDM [33], we use a fixed autoencoder, \mathcal{E} , which first encodes an image $x \in \mathbb{R}^{H \times W \times 3}$, into its latent representation $z = \mathcal{E}(x)$. The fixed decoder, \mathcal{D} reconstructs the image from the latent: $\hat{x} = \mathcal{D}(z) = \mathcal{D}(\mathcal{E}(x))$. Both the encoder and decoder are initialized from the pre-trained Stable Diffusion image model [34], allowing us to take advantage of the priors that have been learned. Because we care about modeling an interaction on a *specific* scene, we provide the current image as context to the model. We encode each image x_t , the corresponding hand masks h_t , and the future query hand mask h_{t+1} , and perform channel-wise concatenation to obtain our full ‘context’ latents $c_i \in \mathbb{R}^{h \times w \times 3c}$. We concatenate the context latents with the latent embedding of the image we are aiming to denoise $z_i \in \mathbb{R}^{h \times w \times c}$ along the channel dimension (where i indicates the diffusion time step and z_i indicates the encoded latent with i steps of added noise).

We learn the diffusion reverse process in this latent space with a U-Net parameterized by θ , ϵ_θ , trained to iteratively denoise the input image by predicting the noise vector that is subtracted at each time step. We want to model the conditional distribution $p(z|x)$ such that the synthesized image looks semantically similar to the input image. This is because the task entails retaining much of the semantics of the objects and backgrounds in the scene. Therefore, along with channel conditioning, we apply cross-attention to condition the model on a CLIP embedding of the input image $\tau(x_t)$. We use a frozen CLIP image encoder and randomly drop the conditioning signals c_i and $\tau(x_t)$ with a probability of 5% during training. At inference time, we start from Gaussian noise and condition on c_i and $\tau(x_t)$ using ϵ_θ . Our learning objective is:

$$\min_{\theta} \mathbb{E}_{z, c \sim \mathcal{E}(x), i, \epsilon \sim \mathcal{N}(0,1)} \|\epsilon - \epsilon_\theta(z_i, c_i, \tau(x_t), i)\|_2^2.$$

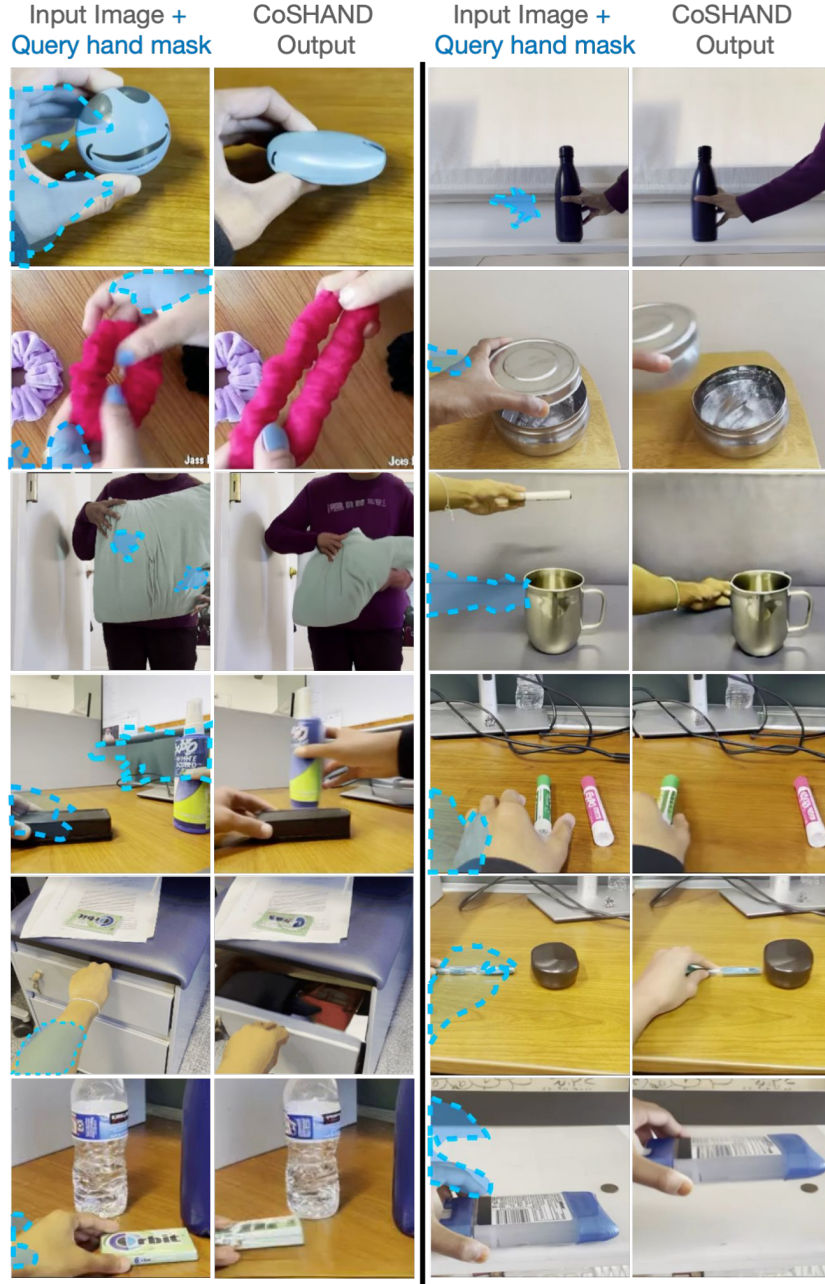


Fig. 4: Examples In-the-wild (from our lab/home environment). We test CoSHAND against challenging In-the-wild collected in our home/lab environments. CoSHAND remains robust in these scenarios, showcasing its strong generalization ability.

4 Experiments

The central hypothesis of our paper is that by **(a)** training large-scale diffusion models with our proposed hand conditioning for action control, and **(b)** learning from a large dataset of unlabeled human videos, we can predict future states of the objects across different scenes and embodiments (human and robot hands). The goal of the experiments section is to test this hypothesis by presenting a set of rigorous and controlled experiments.

4.1 Datasets

SomethingSomethingv2 (Human Video Dataset): We have seen strong results when training large models with large-scale datasets, as it leads to impressive results in open-world settings with strong performance and generalization. To leverage this, we finetune our model on the SomethingSomethingv2 dataset [16]. The dataset contains over 180k videos of humans performing pre-defined, basic actions with everyday objects, for example ‘moving something up’, ‘putting something next to something’, or ‘pushing something from left to right’. To obtain hand masks, we use Segment Anything [24] with the provided bounding boxes as prompts. We split each video into frames at a frame-rate of 12 FPS and sample before and after images at intervals of 3. At training time, we sample one frame of context and one later target frame.

In-the-wild (Self-Recorded Video Dataset): We are able to achieve generalization to different environments and objects not present in the original dataset. To showcase examples beyond the test set of the SomethingSomethingv2 dataset, we record 45 videos in our labs/homes of interactions between our hands (single and bi-manual) and a variety of objects in and around our lab/home environments. We aim to find object categories outside the dataset such as dough and white boards (however due to the great diversity of SomethingSomethingv2 object categories do overlap but of course instances are unique). Some samples and results are visualized in our qualitative results (see Fig. 4).

BridgeDatav2 (Robot Dataset): Because many high-level object dynamics and interactions are similar across embodiments, we show promising zero-shot (no further training/fine-tuning) generalization results. We use select examples from various environments in the BridgeData V2 [40] which contains over 60k trajectories of the WidowX 250 6DOF arm interacting with a variety of objects in 24 different environments.

4.2 Baselines and Metrics

Central to object deformations and interaction planning is understanding the effects of manipulating of a scene. Therefore, we compare our method of using hand conditioning against baselines, which perform image generation with different forms of control:

Method	SSIM \uparrow	PSNR \uparrow	LPIPS \downarrow	Method	SSIM \uparrow	PSNR \uparrow	LPIPS \downarrow
MCVD	0.231	8.75	0.307	MCVD	0.373	11.487	0.282
UCG	0.340	12.08	0.124	UCG	0.458	13.858	0.210
IPix2Pix	0.289	9.53	0.296	IPix2Pix	0.498	13.594	0.275
TCG	0.234	9.05	0.221	TCG	0.454	14.201	0.207
Ours	0.414	13.72	0.116	Ours	0.576	18.156	0.125

(a) **Results on validation subset of SomethingSomethingv2.** All metrics demonstrate that our method can generalize to examples not in the training distribution and that our method outperforms the baseline by a significant margin.
 (b) **Results on In-the-wild test set.** We show in the training distribution and that our method outperforms baselines.

Table 1: Quantitative Results on SomethingSomethingv2 and In-the-wild datasets.

- **Masked Conditional Video Diffusion (MCVD)** [39]: Prior work has investigated predicting the future conditioned on previous frames. MCVD proposes a masked conditional video diffusion model that can be used for video prediction (forward and backward), unconditional generation, and interpolation. For comparison, we trained MCVD for conditional future prediction.
- **Unconditional Generation (UCG)**: While MCVD is a strong baseline, for a fair architecture comparison we also finetune the pre-trained Stable Diffusion image model with the input frame as conditioning.
- **InstructPix2Pix (IPix2Pix)**: is a strong text editing method. We evaluate InstructPix2Pix on this dataset (with captions provided in the SSv2 metadata) to show the difficulty of text based editing.
- **Text-Conditional Generation (TCG)**: We also finetune the pre-trained Stable Diffusion model to condition on the input image along with our captions (provided in the SSv2 metadata) fed in through cross-attention layers. This allows for a more fair architectural comparison.
- **Ours (CoSHAND)**: Our method is trained with an image, its corresponding handmask, and a query handmask (indicating the action), and is expected to predict the target image. Note that during evaluation, the mask can be replaced for a robot arm mask as shown later in section 4.4.

We evaluate our method and baselines on three metrics that capture various aspects of image similarity, including: PSNR which measures a scaled mean-squared error, SSIM which measures the structural similarity [41], and LPIPS which measures perceptual similarity [51].

4.3 Quantitative Analysis

We show numerical results on the test set of SomethingSomethingv2 in Table 1a. We see that **CoSHAND** outperforms all other baselines in all three metrics including PSNR, SSIM, and LPIPS. We notice that MCVD performs poorly compared to other baselines likely due to the lack of pretrained priors or hand controls.

UCG performs better however it still lacks hand controls and therefore cannot synthesize specific futures. Text conditioning allows for high-level controls, however most text-conditioning methods are unable to perform drastic object manipulations therefore IPix2Pix and TCG do not show great performance. Finally, **CoSHAND** shows the strongest performance as it uses Stable Diffusion priors and is trained on a large scale manipulation dataset.

In Table 1b we show numerical results on our In-the-wild dataset, showing strong generalization capabilities. **CoSHAND** beats all baselines. Similar trends hold in terms of large image model priors being key for generalizability. Furthermore, we support these metrics with qualitative comparisons as shown in Fig. 2. We show that **CoSHAND** is able to accurately model the manipulation while TCG only vaguely follows the desired interaction.

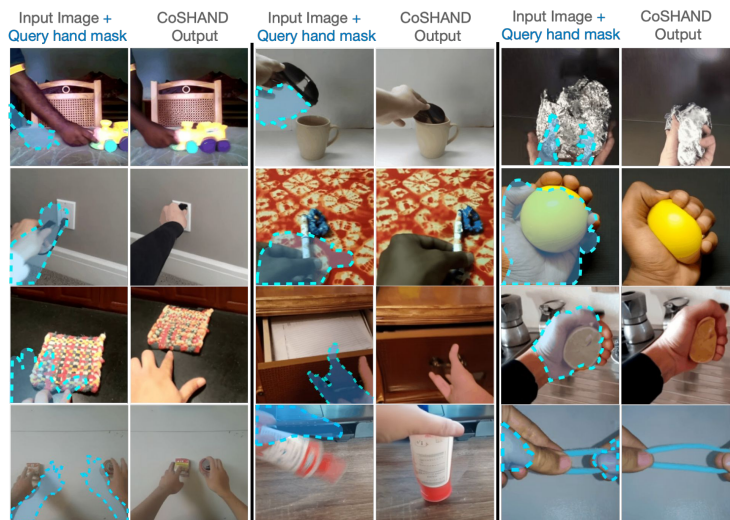


Fig. 5: We show that CoSHAND can perform complex manipulations on a variety of rigid and deformable objects. We show interactions such as squeezing a lemon, closing a drawer, rotating a bottle, and placing items inside cups, which requires understanding of deformable and articulated objects, as well as occlusion. In columns 1, 3 & 5 we visualize the input image and the query hand mask of the desired interaction. Columns 2, 4 & 6 portray the respective outputs of the applied hand interaction.

4.4 Qualitative Analysis

We conduct a qualitative analysis to understand what are success/failure cases of **CoSHAND** and its generalization capabilities. Note that since **CoSHAND** is deterministic, given the same mask and context different samples produce slightly different outputs. Therefore, during our qualitative analysis, we sample the model 4 times

and manually assess the outputs and visualize the best of 4. We notice that the variations for most samples are natural and usually correspond to uncertainty of interaction/environment forces.

Results on SomethingSomethingv2: We visualize predictions on the validation set of SomethingSomethingv2 in Fig. 5. Due to the diversity in actions that CoSHAND was trained on, it is able to understand interactions such as pushing, pulling, stretching, and compressing rigid/deformable objects very well. We also see decent success for more complex interactions such as putting objects inside other objects, opening/closing things, and folding. The training dataset includes bi-manual manipulation data, so we are able to accurately model one and two hand interactions, a strong ability that can aid in downstream robotics tasks. Furthermore, we can model secondary contact forces (e.g. when someone pushes pencil which in-turn pushes a block that it is in contact with, see Fig. 4 row 5 column 3/4).

In-the-wild results: We investigate the models ability to generalize to open-world setting. As seen in Fig. 4, the method works on unseen objects/environments likely due to the large diversity in the training dataset. Furthermore, CoSHAND is able to model situations where the hand does not remain in contact with the object. In Fig. 8 the last three columns depict three different samples of CoSHAND, reasoning about various possible futures when the forces of the interaction are ambiguous and therefore many futures are possible.

Predicting two distinct futures: We show the capabilities of CoSHAND to predict two futures given different conditioning. For example, in column 2 of Fig. 1, the model initially sees the hand holding a bottle, but then the person decides to move their hand either right or left, and the model is able to accurately synthesize both futures. Similarly, in Fig. 7 the CoSHAND is able to model different motions to erase a whiteboard. Providing a visual cue to this controllable model allows us to detail two possible futures which can be highly useful for example in planning or simulation.

Robustness to Hand Masks: While an off-the-shelf Segmentation model can be used to obtain query handmasks as done in this work, a question that could arise is ‘what if a user were to provide hand-drawn approximate hand masks?’. We explore this questions by comparing segmentation masks obtained via off-the-shelf models (‘SegAny’) to manually-drawn ‘high quality’ query handmasks and manually-drawn ‘low quality’ query hand-masks (where details of hand/fingers are not captured in mask). To obtain manually-drawn handmasks, we used Keynote’s line-drawing tool. CoSHAND is robust to even noisy hand masks. However, CoSHAND is better able to reason about the interaction when given more detailed handmasks from ‘SegAny’ & ‘high qual’, indicating that ‘higher quality’ conditioning produces more fine grained outputs. Therefore, one should strive to select the best practically available hand mask. See figure in Supplemental.

Zero-Shot Generalization to Robot Interactions There is very little data in robotics and much of the data that exists is constrained to research lab based environments. Therefore, finding ways to transfer knowledge from humans to robots is an important step towards generalizable robotics models. We

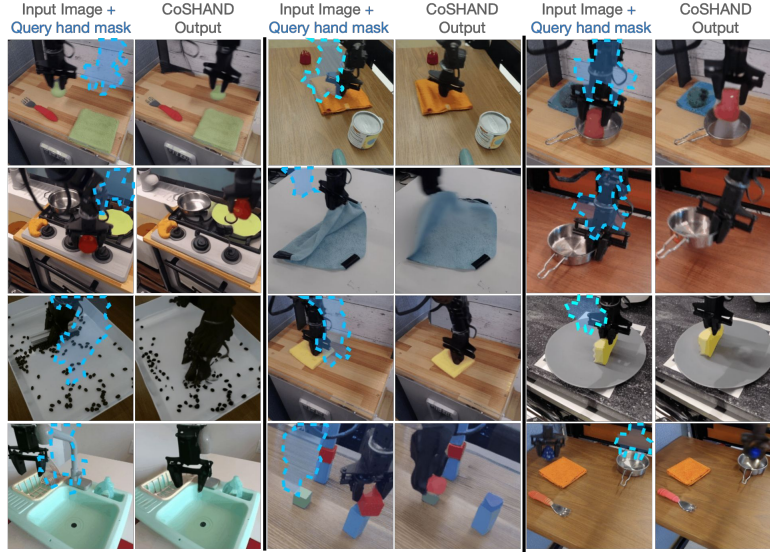


Fig. 6: While CoSHAND is only trained on hands, it can generalize to robot arms for simple actions. For example, CoSHAND can predict reasonably the result of robotic actions including: moving objects around, picking up objects, unfolding cloth, and sweeping granular particles.

hypothesize that CoSHAND has learned many key aspects of interaction dynamics which though trained on human videos, can be leveraged for other embodiments such as a robot arm.

We aim to see how CoSHAND performs for robot datasets where the interaction is executed by a robot arm. Similar to the setting with hands, we use off-the shelf bounding box and segmentation models to obtain segmentation masks of the robot arm. We input a current frame, the corresponding end effector mask, and the query end effector mask to CoSHAND, which then produces the future state of the scene. We evaluate on a small subset of data from BridgeDataV2, and as seen in Fig. 6, we see good performance on simple object/action pairs such as pushing a towel or picking up and moving a cup. This indicates a promising path towards enabling robots with the ability to build mental models of objects by watching humans interact with the world.

Image Editing: Controllable content creation is a key aspect for many designers in the arts and gaming industries, and CoSHAND allows for a very natural way to edit scenes through use of the hand. To perform this edit, we capture an image of our hand overlaying and manipulating the image as input to CoSHAND. In Fig. 1, we show an edit on a painting where the user lifts up a wilting flower. Furthermore, in Fig. 7 we show editing the location of the snitch from Harry Potter, and the bike from the movie E.T.

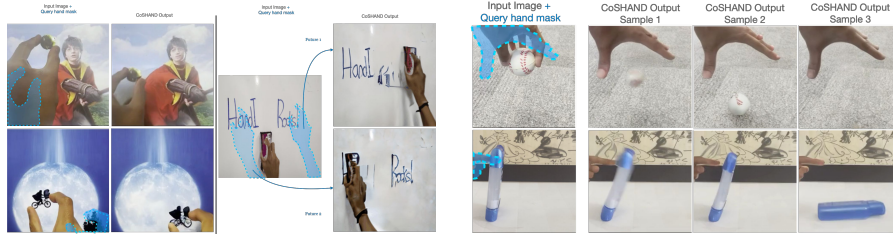


Fig. 7: CoSHAND can edit images and Fig. 8: Examples where forces are ambiguous: The force of an interaction or the move objects around in famous movie environment affecting the interaction may be scenes such as the snitch from Harry Potter and the bike from E.T. Furthermore, possible futures. In column 2-4 we show three we show that conditioned on the same input context but different hand mask trajectories, CoSHAND predicts an **alternate future, while maintaining the photorealism of the predicted future frames.**

4.5 Ablation Study

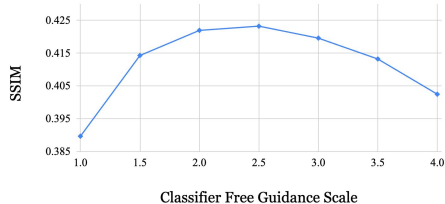
We perform several rigorous ablations to test our hypothesis (see Table 9a & Fig. 9b for quantitative results):

- **No SD prior:** As hypothesized, training CoSHAND from scratch without leveraging Stable Diffusion pretraining, results in poor performance, likely due to the lack of prior knowledge of hand-object interactions.
- **No CLIP conditioning:** We remove CLIP conditioning signal and retrain the model to verify CLIP embedding effectiveness. Performance drops on all metrics, likely because without the clip embedding, the overall semantics of the image are lost, making reconstructing details difficult.
- **Less Training Data:** When we train CoSHAND with 10% of the data, performance and generalization suffer. This indicates that as the amount of training data increases, we will see increase in the model’s capabilities.
- **Ours + Context:** We hypothesize that providing more frames of context as input to CoSHAND would result in better performance as a temporal understanding is induced. We test this hypothesis by training CoSHAND with 4 frames of context and their corresponding handmasks, and querying one step into the future. As expected, this improves performance. However, we note that this setting is less common as multiple context frames are often unavailable, therefore we do not delve deeply into this setting.
- **Varying CFG scale:** We notice that increasing the guidance scale beyond a certain threshold, 2.5 in our case, causes the image becomes less dynamic likely due to negative correlation between variance and cfg as seen in Fig. 9b. This is because a higher cfg causes the model to more closely replicate the input image x_t . On the other hand, guidance scales less than 2.0 result

in the original image being ignored, hence decreasing performance. See Fig. 9b for visualization of numerical analysis.

Method	SSIM \uparrow	PSNR \uparrow	LPIPS \downarrow
No SD prior	0.376	12.36	0.116
No CLIP Cond	0.366	11.76	0.173
Less Data	0.369	12.45	0.127
Ours	<u>0.423</u>	<u>14.00</u>	<u>0.108</u>
Ours + Context	0.448	14.76	0.088

(a) **Table 9: We perform ablations on our method.** Note that the stable-diffusion priors and size of the dataset contribute to significant performance gains. Furthermore, when more context is available, the model is able to better reason about the next state (bolded last row).



(b) **Fig 9: Classifier-Free Guidance Scale Analysis.** Performance peaks at a cfg value of around 2.5, as too high guidance decreases the variety of the possible generations, while too low of a guidance ignores the input frame.

4.6 Limitations

Our approach aims to learn plausible manipulations between hands and objects, therefore we find our model does not perform well in very unrealistic settings (e.g.: using hands to change shapes of clouds, pushing buildings, etc). Furthermore, we find that when its not obvious if two objects are separated, sometimes interactions with one object will result in ambiguous surrounding objects also being altered.

5 Discussion

In this work, we propose hand conditioning as a new method for interacting with a scene, allowing machines to build a mental model of the future conditioned on a specific action. We introduce **CoSHAND**, a method that leverages large-scale pretrained models and learns from a large source of unlabeled data to build a representation of interaction. We show experiments on the Something-Something v2 dataset, real world examples, zero-shot generalization to robot grippers, and interesting image edits. We note that **CoSHAND** does not require any metadata and can therefore be easily scaled to improve results on robot datasets and unrealistic image edits. This opens up exciting applications in robot planning, controllable image generation, and augmented or virtual reality.

Acknowledgments: This project is based on research partially supported by NSF #2202578, NSF #1925157, and the National AI Institute for Artificial and Natural Intelligence (ARNI). SS is supported by the NSF GRFP.

References

1. Amit, T., Shaharabany, T., Nachmani, E., Wolf, L.: Segdiff: Image segmentation with diffusion probabilistic models. arXiv preprint arXiv:2112.00390 (2021)
2. Brahmabhatt, S., Ham, C., Kemp, C.C., Hays, J.: Contactdb: Analyzing and predicting grasp contact via thermal imaging. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8709–8719 (2019)
3. Brahmabhatt, S., Handa, A., Hays, J., Fox, D.: Contactgrasp: Functional multi-finger grasp synthesis from contact. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 2386–2393. IEEE (2019)
4. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18392–18402 (2023)
5. Chen, Y., Tu, Z., Kang, D., Bao, L., Zhang, Y., Zhe, X., Chen, R., Yuan, J.: Model-based 3d hand reconstruction via self-supervised learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10451–10460 (2021)
6. Christen, S., Kocabas, M., Aksan, E., Hwangbo, J., Song, J., Hilliges, O.: D-grasp: Physically plausible dynamic grasp synthesis for hand-object interactions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20577–20586 (2022)
7. Craik, K.J.W.: The nature of explanation, vol. 445. CUP Archive (1967)
8. Cutkosky, M.R.: On grasp choice, grasp models, and the design of hands for manufacturing tasks. IEEE Trans. Robotics Autom. **5**, 269–279 (1989), <https://api.semanticscholar.org/CorpusID:34048367>
9. Dai, Y., Yang, M., Dai, B., Dai, H., Nachum, O., Tenenbaum, J., Schuurmans, D., Abbeel, P.: Learning universal policies via text-guided video generation. arXiv preprint arXiv:2302.00111 (2023)
10. Deitke, M., Liu, R., Wallingford, M., Ngo, H., Michel, O., Kusupati, A., Fan, A., Laforte, C., Voleti, V., Gadre, S.Y., et al.: Objaverse-xl: A universe of 10m+ 3d objects. arXiv preprint arXiv:2307.05663 (2023)
11. Detry, R., Kraft, D., Buch, A.G., Krüger, N., Piater, J.: Refining grasp affordance models by experience. In: 2010 IEEE International Conference on Robotics and Automation. pp. 2287–2293. IEEE (2010)
12. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
13. Du, Y., Yang, M., Florence, P., Xia, F., Wahid, A., Ichter, B., Sermanet, P., Yu, T., Abbeel, P., Tenenbaum, J.B., et al.: Video language planning. arXiv preprint arXiv:2310.10625 (2023)
14. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618 (2022)
15. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
16. Goyal, R., Kahou, S.E., Michalski, V., Materzyńska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., Hoppe, F., Thureau, C., Bax, I., Memisevic, R.: The "something something" video database for learning and evaluating visual common sense (2017)
17. Ha, D., Schmidhuber, J.: World models. arXiv preprint arXiv:1803.10122 (2018)

18. Hafner, D., Lillicrap, T., Ba, J., Norouzi, M.: Dream to control: Learning behaviors by latent imagination. arXiv preprint arXiv:1912.01603 (2019)
19. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control (2022)
20. Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J., Salimans, T.: Imagen video: High definition video generation with diffusion models (2022)
21. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. *Advances in neural information processing systems* **33**, 6840–6851 (2020)
22. Hsiao, K., Lozano-Perez, T.: Imitation learning of whole-body grasps. In: 2006 IEEE/RSJ international conference on intelligent robots and systems. pp. 5657–5662. IEEE (2006)
23. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models (2021)
24. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything (2023)
25. Liu, M., Pan, Z., Xu, K., Ganguly, K., Manocha, D.: Generating grasp poses for a high-dof gripper using neural networks. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 1518–1525. IEEE (2019)
26. Liu, R., Wu, R., Van Hoorick, B., Tokmakov, P., Zakharov, S., Vondrick, C.: Zero-1-to-3: Zero-shot one image to 3d object. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9298–9309 (2023)
27. Liu, S., Tripathi, S., Majumdar, S., Wang, X.: Joint hand motion and interaction hotspots prediction from egocentric videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3282–3292 (2022)
28. Ozguroglu, E., Liu, R., Surís, D., Chen, D., Dave, A., Tokmakov, P., Vondrick, C.: pix2gestalt: Amodal segmentation by synthesizing wholes (2024)
29. Pan, J., Ferrer, C.C., McGuinness, K., O’Connor, N.E., Torres, J., Sayrol, E., Giro-i Nieto, X.: Salgan: Visual saliency prediction with generative adversarial networks. arXiv preprint arXiv:1701.01081 (2017)
30. Puhlmann, S., Heinemann, F., Brock, O., Maertens, M.: A compact representation of human single-object grasping. In: 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 1954–1959. IEEE (2016)
31. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents (2022)
32. Rezende, D.J., Mohamed, S., Wierstra, D.: Stochastic backpropagation and approximate inference in deep generative models. In: International conference on machine learning. pp. 1278–1286. PMLR (2014)
33. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
34. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models (2022)
35. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22500–22510 (2023)
36. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy,

- S., Crowson, K., Schmidt, L., Kaczmarczyk, R., Jitsev, J.: Laion-5b: An open large-scale dataset for training next generation image-text models (2022)
37. Shan, D., Geng, J., Shu, M., Fouhey, D.: Understanding human hands in contact at internet scale. In: CVPR (2020)
 38. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., Parikh, D., Gupta, S., Taigman, Y.: Make-a-video: Text-to-video generation without text-video data (2022)
 39. Voleti, V., Jolicoeur-Martineau, A., Pal, C.: Mcvd: Masked conditional video diffusion for prediction, generation, and interpolation. *Advances in Neural Information Processing Systems* **35**, 23371–23385 (2022)
 40. Walke, H., Black, K., Lee, A., Kim, M.J., Du, M., Zheng, C., Zhao, T., Hansen-Estruch, P., Vuong, Q., He, A., Myers, V., Fang, K., Finn, C., Levine, S.: Bridge-data v2: A dataset for robot learning at scale. In: Conference on Robot Learning (CoRL) (2023)
 41. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* **13**(4), 600–612 (2004). <https://doi.org/10.1109/TIP.2003.819861>
 42. Wu, P., Escontrela, A., Hafner, D., Abbeel, P., Goldberg, K.: Daydreamer: World models for physical robot learning. In: Conference on Robot Learning. pp. 2226–2240. PMLR (2023)
 43. Wu, R., Liu, R., Vondrick, C., Zheng, C.: Sin3dm: Learning a diffusion model from a single 3d textured shape. *arXiv preprint arXiv:2305.15399* (2023)
 44. Yang, M., Du, Y., Ghasemipour, K., Tompson, J., Schuurmans, D., Abbeel, P.: Learning interactive real-world simulators. *arXiv preprint arXiv:2310.06114* (2023)
 45. Yang, S., Walker, J., Parker-Holder, J., Du, Y., Bruce, J., Barreto, A., Abbeel, P., Schuurmans, D.: Video as the new language for real-world decision making (2024)
 46. Ye, Y., Gupta, A.K., Tulsiani, S.: What’s in your hands? 3d reconstruction of generic objects in hands. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 3885–3895 (2022), <https://api.semanticscholar.org/CorpusID:248178083>
 47. Ye, Y., Hebbar, P., Gupta, A., Tulsiani, S.: Diffusion-guided reconstruction of everyday hand-object interaction clips. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 19717–19728 (October 2023)
 48. Ye, Y., Hebbar, P., Gupta, A., Tulsiani, S.: Diffusion-guided reconstruction of everyday hand-object interaction clips. In: ICCV (2023)
 49. Ye, Y., Li, X., Gupta, A., De Mello, S., Birchfield, S., Song, J., Tulsiani, S., Liu, S.: Affordance diffusion: Synthesizing hand-object interactions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22479–22489 (2023)
 50. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023)
 51. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric (2018)