Controlling the World by Sleight of Hand

by Srudthi Sudhakar et al.

ECCV 2024 (Oral Presentation) – Award Candidate

Jakub Świstak

MI2.AI Seminar, Warsaw, January 13th, 2025

Authors

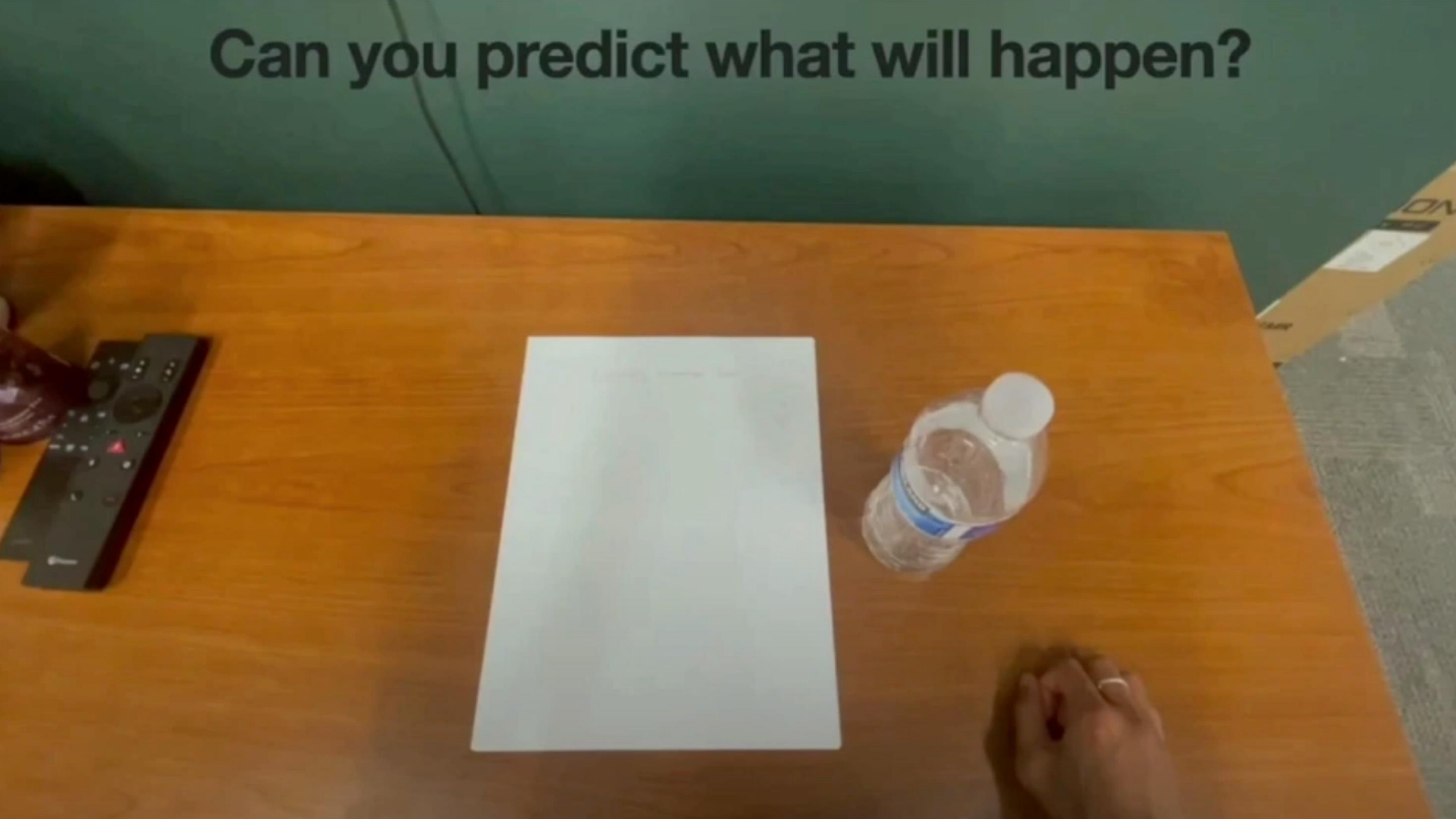
Sruthi Sudhakar

Ruoshi Liu

Basile Van Hoorick

Carl Vondrick

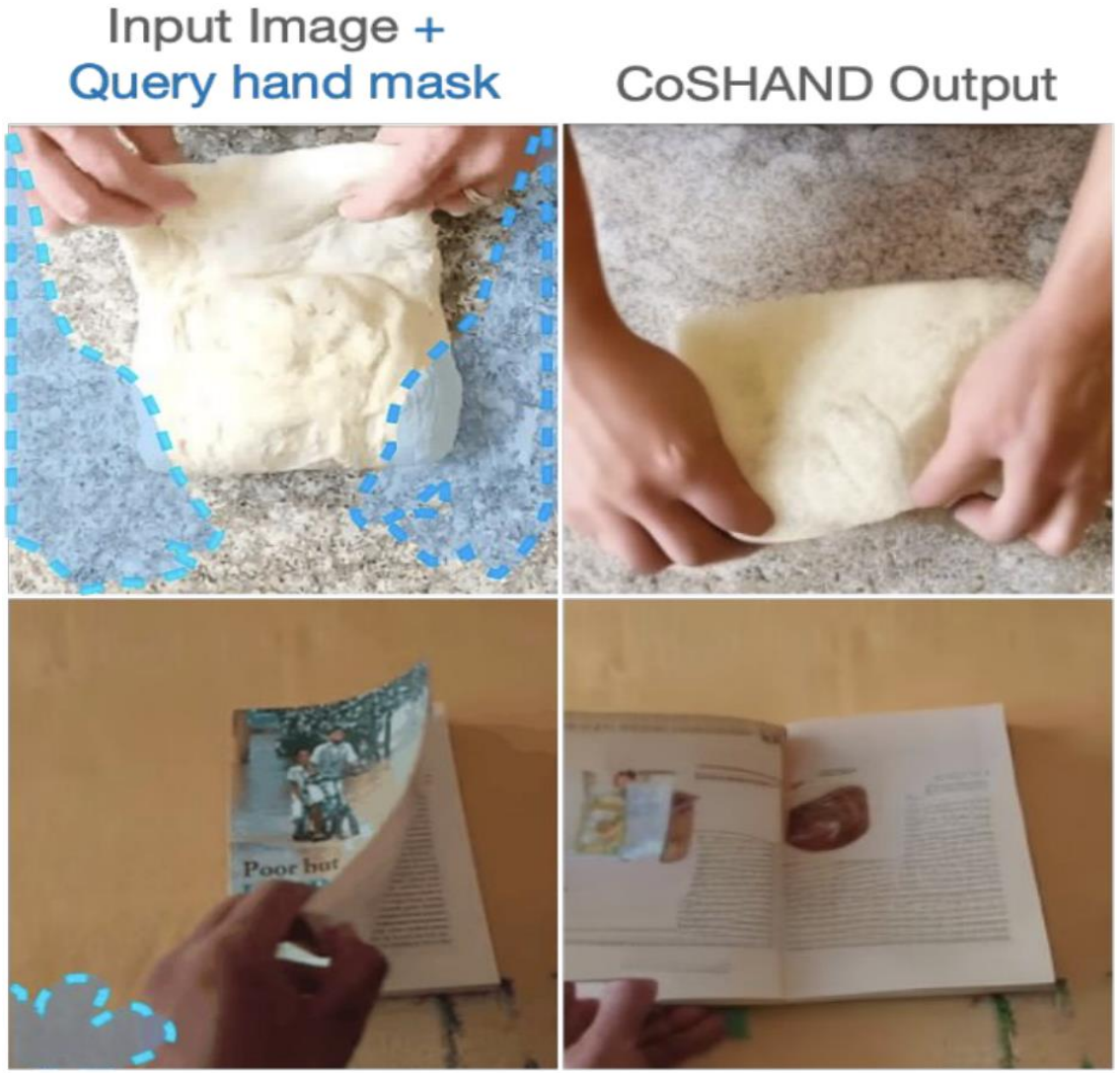
Richard Zemel

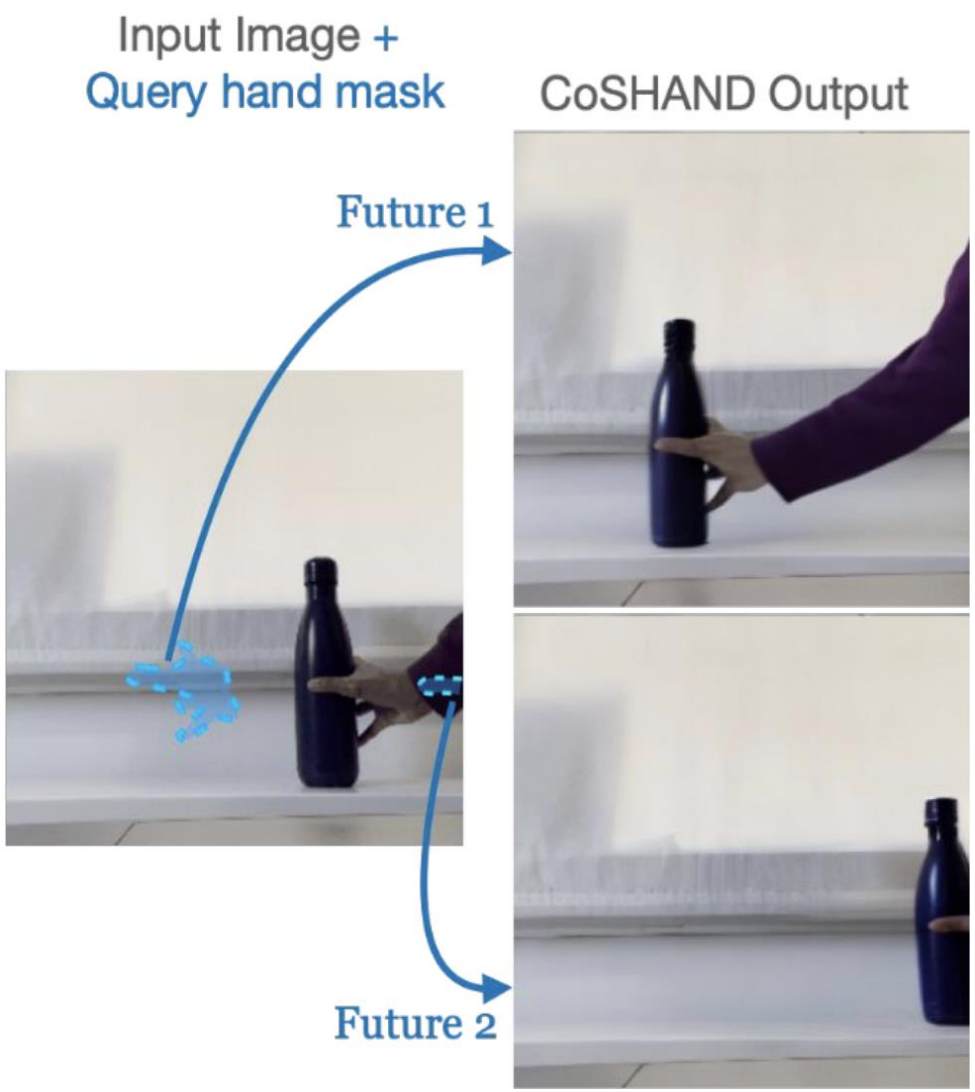

Can you predict what will happen?

Humans can model interactions with objects

Humans can build mental models of the world, reason about actions and their effects on objects.

Action-conditional generative models





Dataset used for fine tuning

SomethingSomethingv2 - 180k videos of humans performing pre-defined, basic actions with everyday objects

Putting *a white remote* into *a cardboard box*

Pretending to put *candy* onto *chair*

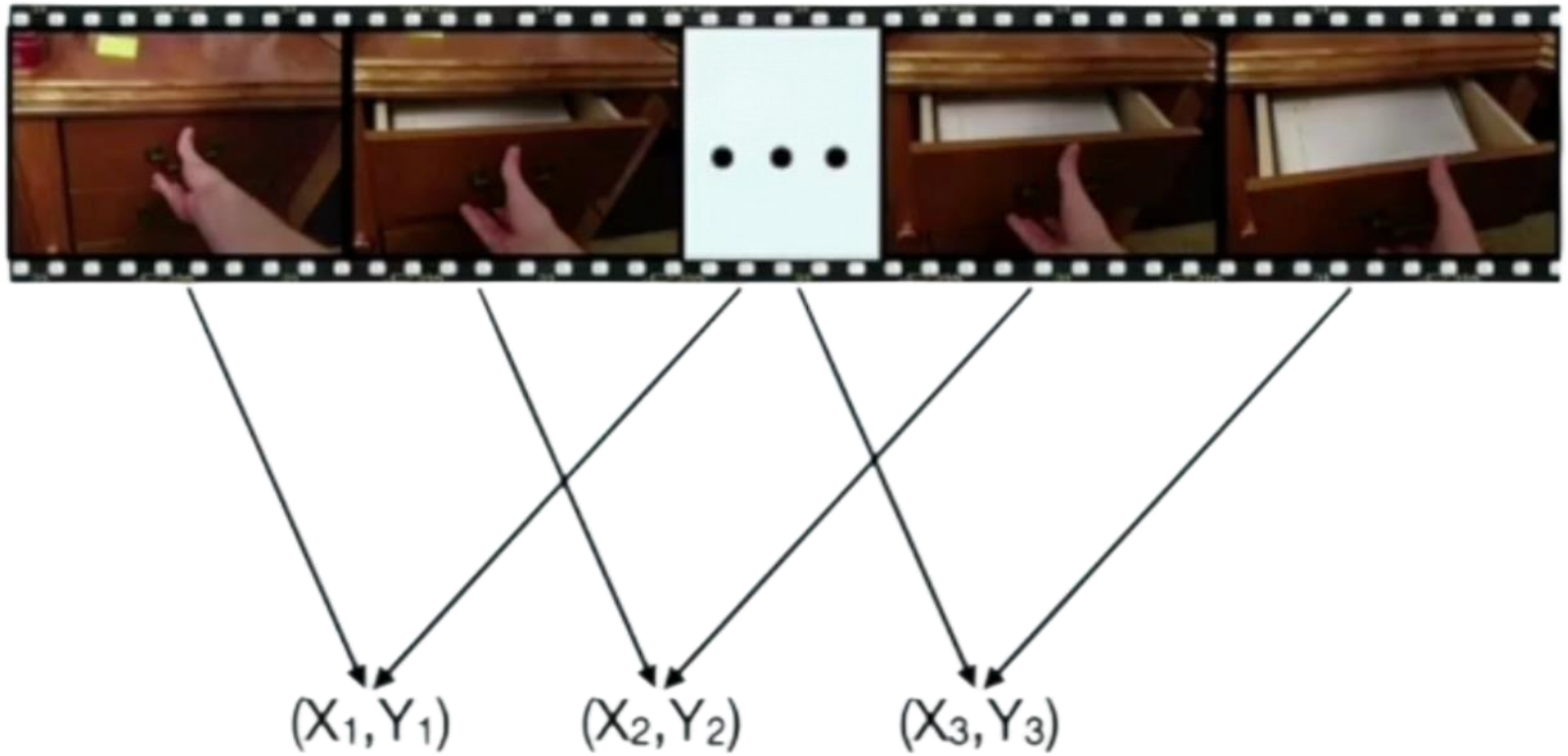
Pushing *a green chilli* so that it falls off the table

Moving *puncher* closer to *scissor*

Dataset used for fine tuning

SomethingSomethingv2 - 180k videos of humans performing pre-defined, basic actions with everyday objects

Sample pairs of frames



Dataset used for fine tuning

SomethingSomethingv2 - 180k videos of humans performing pre-defined, basic actions with everyday objects

Sample pairs of frames

Generate hand masks

Problem definition

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}.

Problem definition

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:

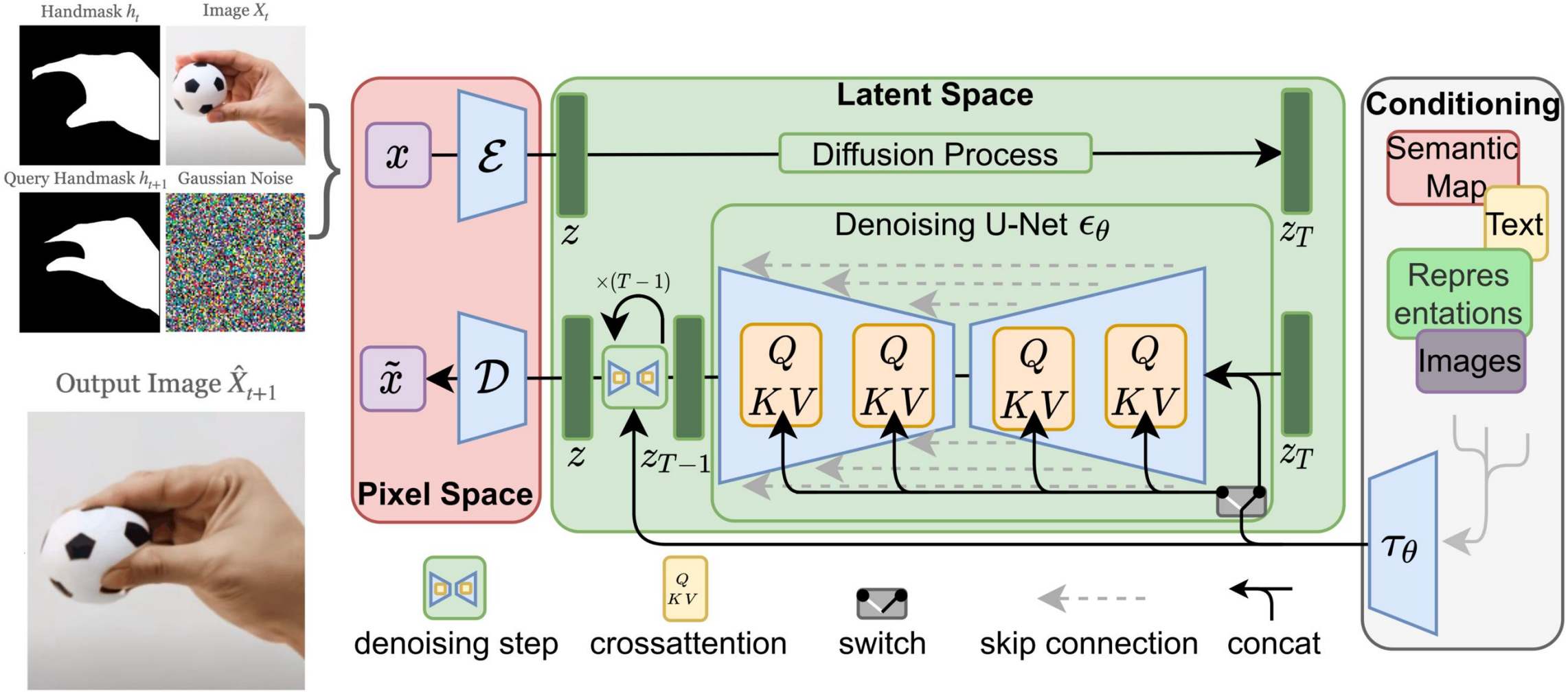
$$f(x_t, h_t, h_{t+1}) = \hat{x}_{t+1}$$

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}.

Similar to LDM authors are using autoencoder ε, which first encodes an image $x \in \mathbb{R}^{H\times W\times 3}$, into its latent representation $z = \varepsilon(x)$. The fixed decoder, $\mathcal{D}$ reconstructs the image from the latent $\hat{x} = \mathcal{D}(\mathrm{z}) = \mathcal{D}(\varepsilon(x))$. To care about the current image we will provide it as a context to the model. To obtain full 'context' latents authors perform channel-wise concatenation $c_i \in \mathbb{R}^{h\times w\times 3c}$. The context is then concatenated with the latent embedding of the image we are aiming to denoise $z_i \in \mathbb{R}^{h\times w\times c}$

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

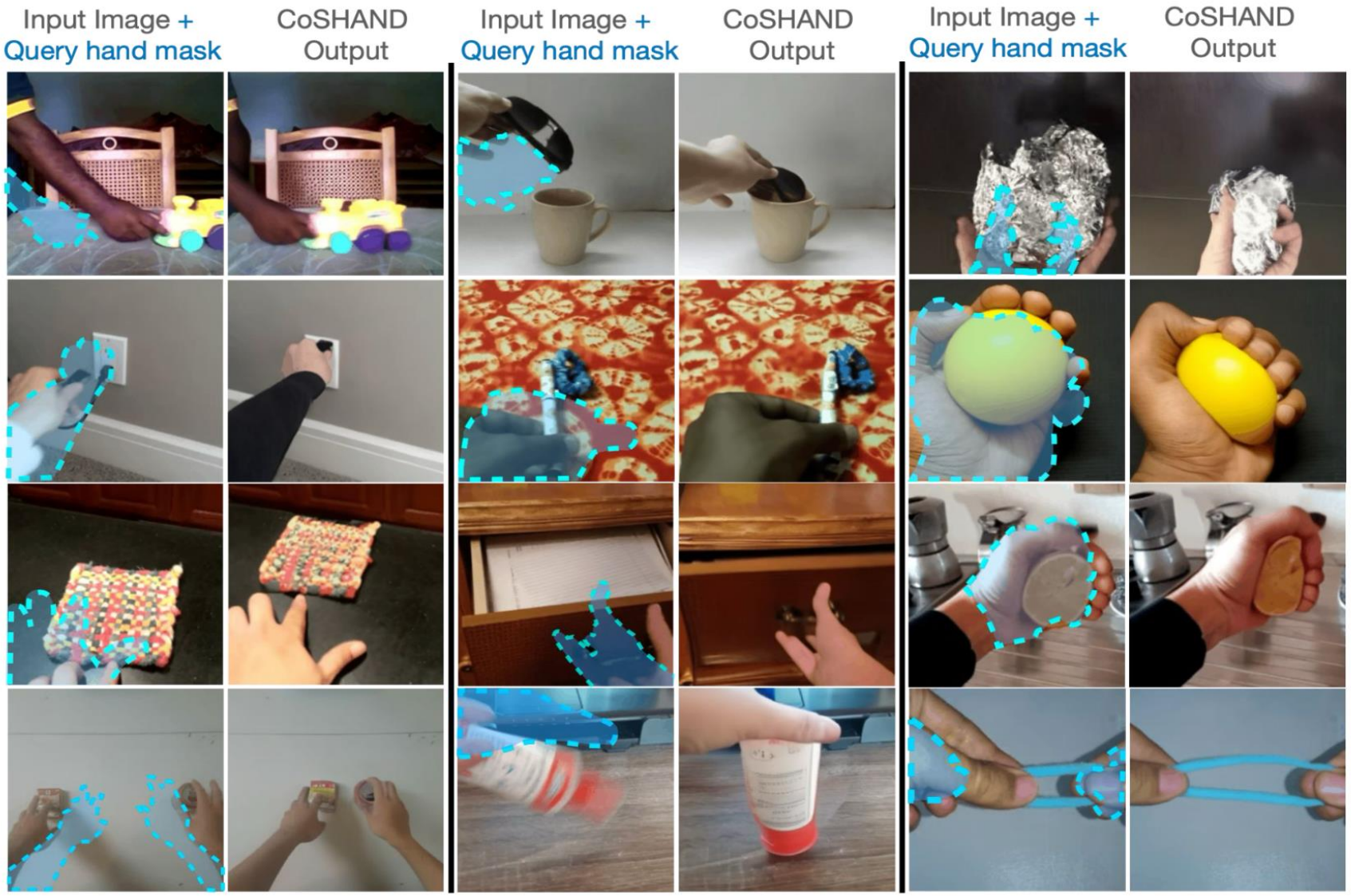
Conditional Generative Models



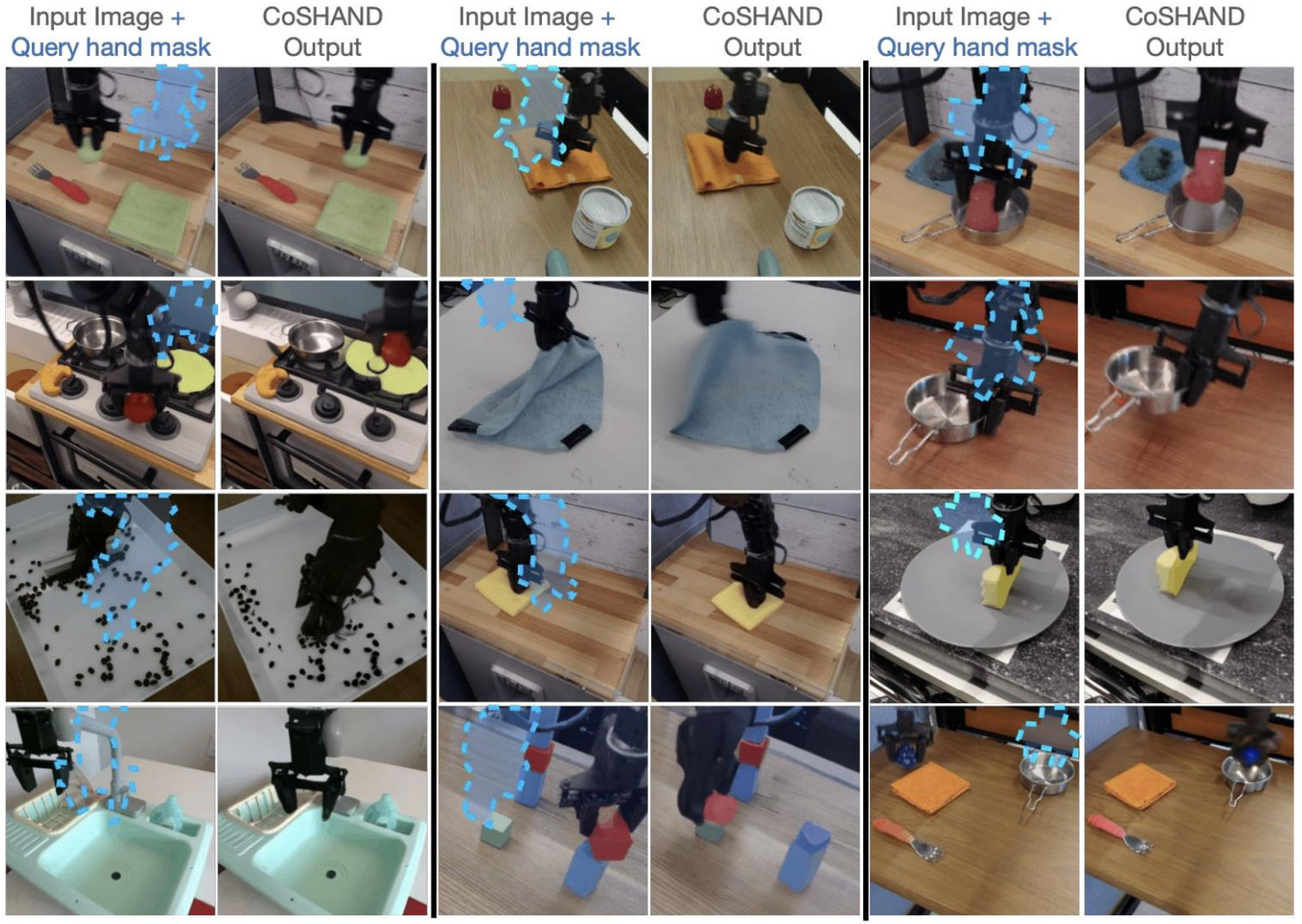
Hypothesis

“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).”

Results


Input Image +
Query hand mask
CoSHAND
Output
Input Image +
Query hand mask
CoSHAND
Output
Input Image +
Query hand mask
CoSHAND
Output

Results (BridgeDatav2)



Baselines comparison

Method	SSIM ↑	PSNR ↑	LPIPS ↓
MCVD	0.231	8.75	0.307
UCG	0.340	12.08	0.124
IPix2Pix	0.289	9.53	0.296
TCG	0.234	9.05	0.221
Ours	**0.414**	**13.72**	**0.116**

SomethingSomethingv2

Method	SSIM ↑	PSNR ↑	LPIPS ↓
MCVD	0.373	11.487	0.282
UCG	0.458	13.858	0.210
IPix2Pix	0.498	13.594	0.275
TCG	0.454	14.201	0.207
Ours	**0.576**	**18.156**	**0.125**

In-the-wild test set

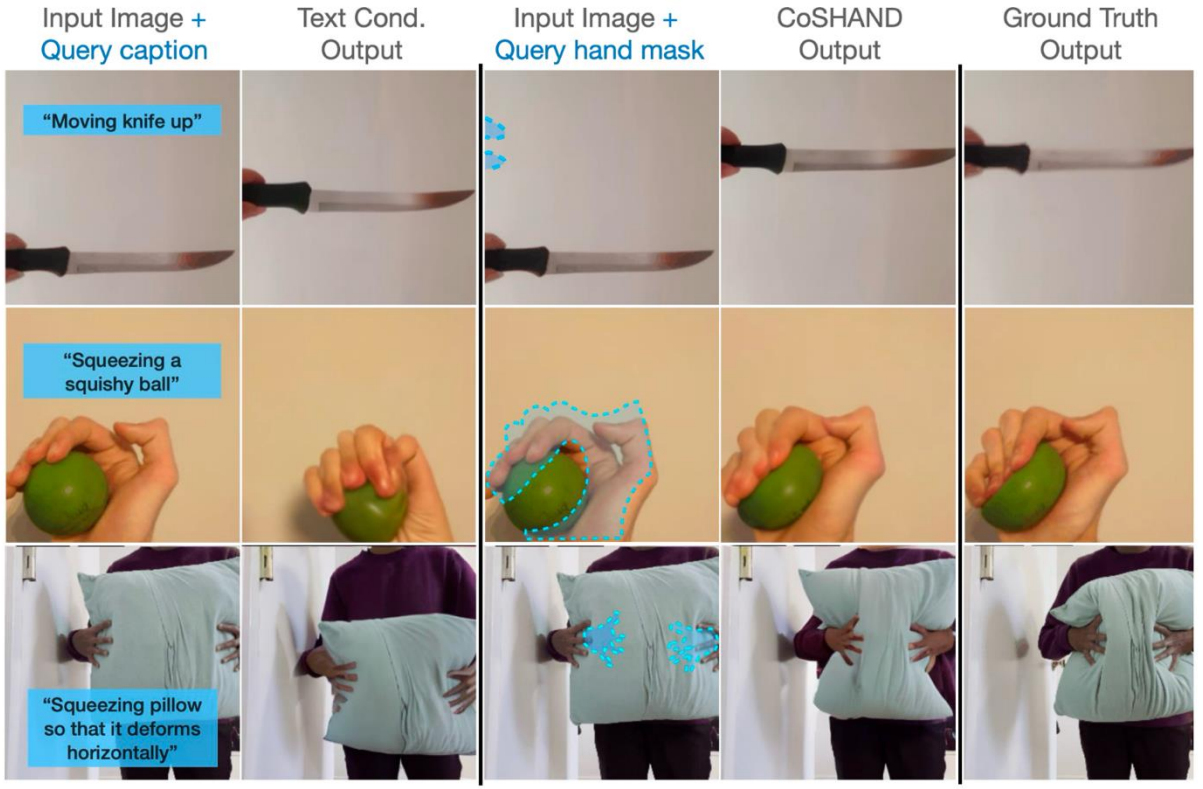


Fig. 2: We show that text-conditioning is insufficient to model interactions, whereas hands allow for better control. Columns 1 & 2 show the input image, query caption, and output of text conditional generation. Columns 3 & 4 show the input image, query hand mask, and output of `CosHand`. Column 5 shows 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.

Image editing

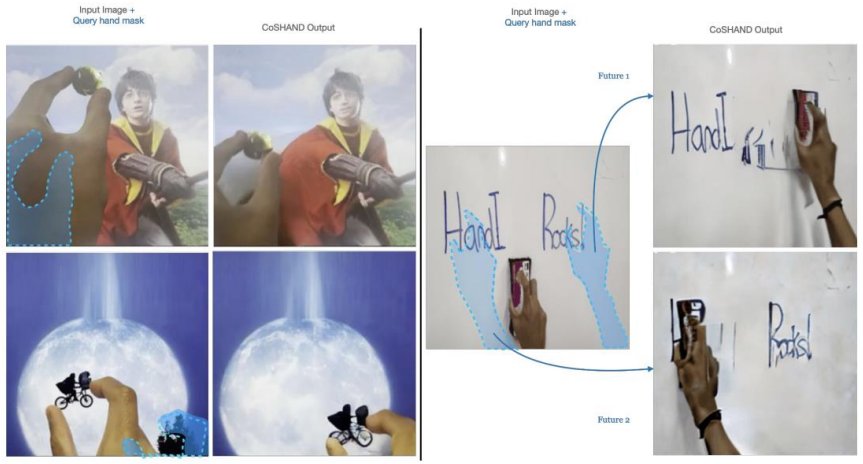


Fig. 7: CosHand can edit images and produce different futures. We can move objects around in famous movie scenes such as the snitch from Harry Potter and the bike from E.T. Furthermore, we show that conditioned on the same input context but different hand mask trajectories, CosHand predicts an **alternate future**, while maintaining the photorealism of the predicted future frames.

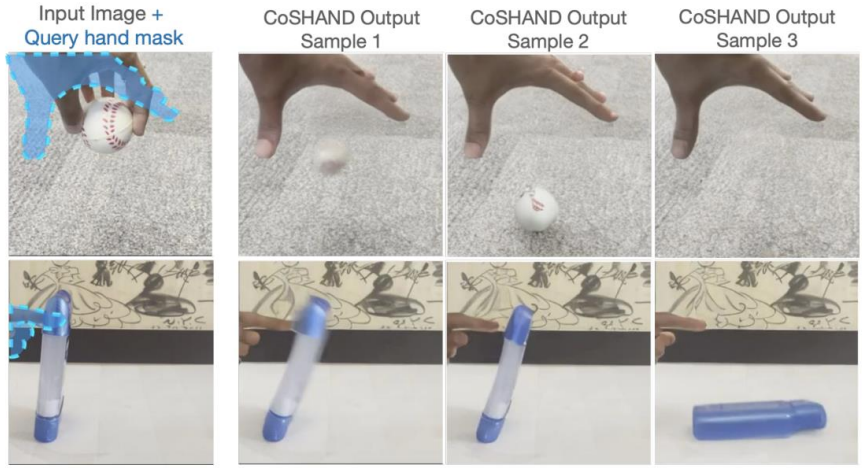
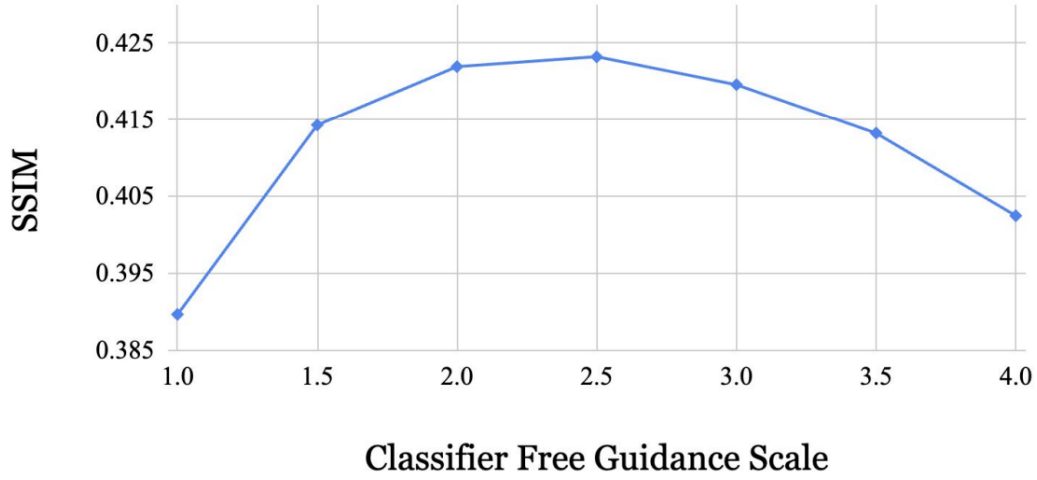


Fig. 8: Examples where forces are ambiguous: The force of an interaction or the environment affecting the interaction may be ambiguous and therefore there may be many possible futures. In column 2-4 we show three different samples taken from CosHand showing the diversity in the outputs when there is uncertainty from interaction.

Ablation Study

Method	SSIM ↑	PSNR ↑	LPIPS ↓
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.

Thank You for your attention!

Jakub Swistak

MI2.AI Seminar, Warsaw, January 13th, 2025