DAViD: Modeling Dynamic Affordance of 3D Objects Using Pre-trained Video Diffusion Models (Supplementary Material)

Hyeonwoo Kim Sangwon Baik Hanbyul Joo
Seoul National University
https://snuvclab.github.io/david/

A. Implementation Details

In this section, we introduce the details of our method for modeling Dynamic Affordance. From Sec. A.1 to Sec. A.4, we cover our first pipeline, 4D HOI Sample Generation. Sec. A.5 and Sec. A.6 describe our second pipeline, learning Dynamic Affordance.

A.1. Rendering Object from Multi-Viewpoints

For camera installation, we position eight perspective cameras evenly spaced at 45°intervals around the object at a fixed elevation of 5°. The radius (distance of camera to origin) is set as a hyperparameter along with additional adjustment of camera's z-coordinate to ensure the object fits within the image frame. To have a consistent camera setup in the uplifting pipeline, we follow GVHMR [19] and set the intrinsic parameters as follows.

$$K = \begin{bmatrix} f & 0 & w/2 \\ 0 & f & h/2 \\ 0 & 0 & 1 \end{bmatrix}, \tag{1}$$

where $f = \sqrt{h^2 + w^2}$ and h, w represent the height and width of our rendering image, respectively. In practice, we use $h = 800$, $w = 1200$ for rendering. For object installation, relatively large and stationary ground-placed objects (e.g., motorcycles) are placed at the origin in a canonical state, while small and portable objects (e.g., umbrellas) are perturbed by sampling their position and rotation within a certain range. The range of the position and rotation is set as a hyperparameter.

A.2. Generating 2D HOI images

For the image rendered in Sec. A.1, we use the Canny edge detector [2] to obtain structural guidance. In practice, we use an upper threshold of 30 and a lower threshold of 25 to capture dense structures. We use the obtained Canny edges as input of ControlNet [22] and leverage the off-the-shelf pre-trained 2D diffusion model, FLUX [12], to generate the 2D HOI Image. Unlike other approaches [9] that directly use inpainting on the rendered object, maintaining a consistent background color (e.g., white, gray), our method generate background, offering the advantage of aligning with the training domain of the video diffusion model while providing motion cues to the world-grounded HMR (*e.g.*, if the background moves left, the subject moves right). For specific settings, we use a classifier-free guidance scale of 3.5, 28 inference steps, and the FlowMatchEulerDiscrete scheduler [3] for image generation. In cases where it is natural for a person to occlude an object (*e.g.*, a hand occluding the handle of a cart), strong structural guidance can lead to the generation of implausible images. Therefore, we set the ControlNet [22] conditioning guidance as 0.725 for the first 12 denoising steps, and 0.0 for the later steps. We empirically find that this approach helps generating plausible HOI image considering appropriate occlusion. For the text prompt for generating images, we use a vision-language model [15] to automatically obtain prompts that include HOI. Specifically, we obtain the text prompt using the following input.

> *Write a text prompt in two sentence. The format of the text prompt should start with "1 person" and should include word "{category}". Write a detailed text prompt focusing on human pose and the interaction between "1 person" and "{category}". The third word of the first sentence must describe the interaction.*

We add the additional tag ", full body" at the end of the obtained text prompt, which we find beneficial for expressing the holistic body in image. While we know the category of the input 3D object in many cases, we use the rendering of the object to request a prompt if the category is not available.

A.3. Generating 2D HOI Video from 2D HOI Image

We use a pre-trained video diffusion model [11] to generate 2D HOI videos from 2D HOI images. For the text prompt, we use the same one used for generating the 2D

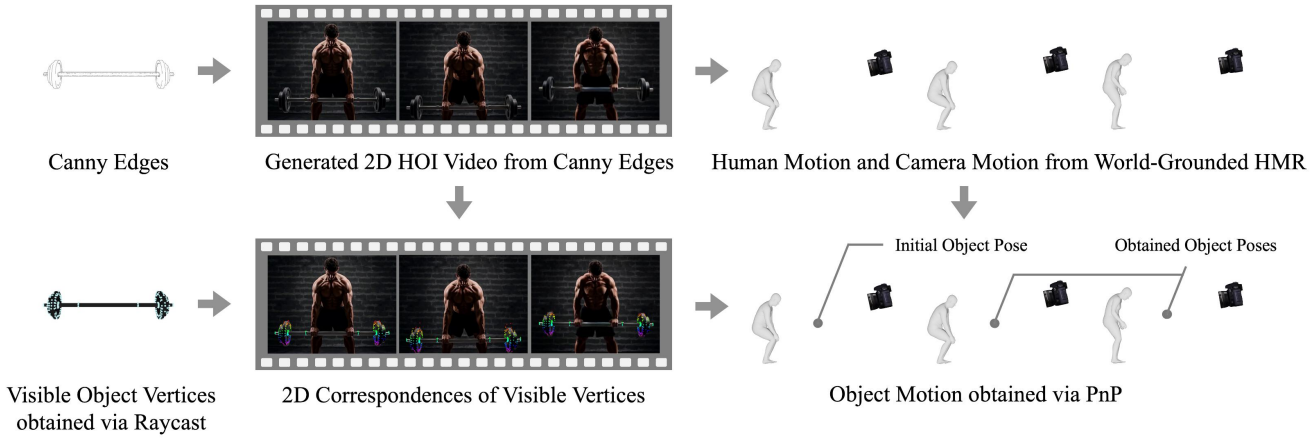

Figure 1. **Obtaining Object Motion.** We first leverage off-the-shelf world-grounded HMR to obtain human motion and corresponding camera motion. Then, for the object vertices visible in our rendering camera, we find the 2D correspondences across the video. Using the 2D-3D correspondence of the vertices and camera pose for every frame, we compute the object pose for each frame via PnP.

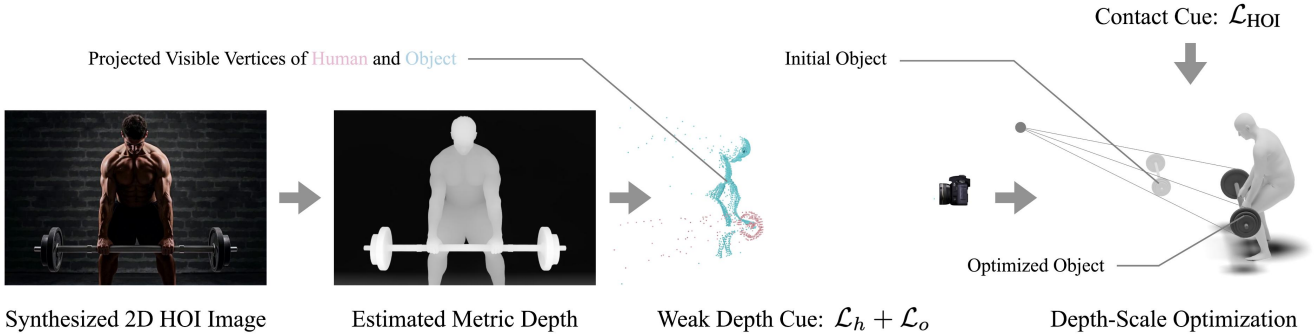

Figure 2. **Resolving Depth Ambiguity.** To resolve the depth ambiguity between human and object motion, we leverage weak depth cues obtained from a metric depth model and contact cues, based on the intuition that object movement is driven by human contact. By optimizing the human and object scales using these cues, we obtain the 4D HOI sample.

HOI image. As the video diffusion model support only specific resolution conditions, we resize both the input image and the output video.

A.4. Lifting 2D HOI Videos to 4D HOI Samples

We detail the process of (1) computing object motion and (2) resolving depth ambiguity which are used to lift 2D HOI Videos into 4D HOI samples with additional figures (Fig. 1, Fig. 2).

Obtaining Object Motion. We leverage an off-the-shelf world-grounded HMR, GVHMR [19] to obtain both human motion and the corresponding camera motion in world coordinates. The core idea for obtaining the remaining object motion is to find 2D-3D correspondences for each frame. As we use a camera model same with GVHMR [19] for rendering, it is possible to transform (rotation and translation of) the rendering camera to the first frame camera of GVHMR's output. Using the same transformation, we obtain the initial (first frame) object pose aligned with the human and camera motion. At the same time, we obtain the vertices of the object visible in the rendered camera through raycasting [18], and find the correspondences of 2D projection points across the generated 2D HOI Video via video tracking [7, 8]. Through this, we establish the 2D-3D correspondences of the vertices for each frame with known camera motion, allows PnP [4, 13] to compute object pose for each frame, as shown in Fig. 1.

Resolving Depth Ambiguity. Even after obtaining the human motion, camera motion, and object motion aligned on 2D, the human motion and object motion do not interact with each other in 3D space. To resolve the depth ambiguity that occurs on perspective camera rays, we optimize the object's scale in the first frame using (1) weak depth cues and (2) contact cues. First, we use a publicly available depth estimation model [1] to predict the metric depth from the generated images. As shown in Fig. 2, the visible vertices of the human and the object, obtained through raycasting [18] are projected into 3D space to construct a point cloud. The MSE distance between the human point cloud and the corresponding visible 3D vertices of the human is defined as

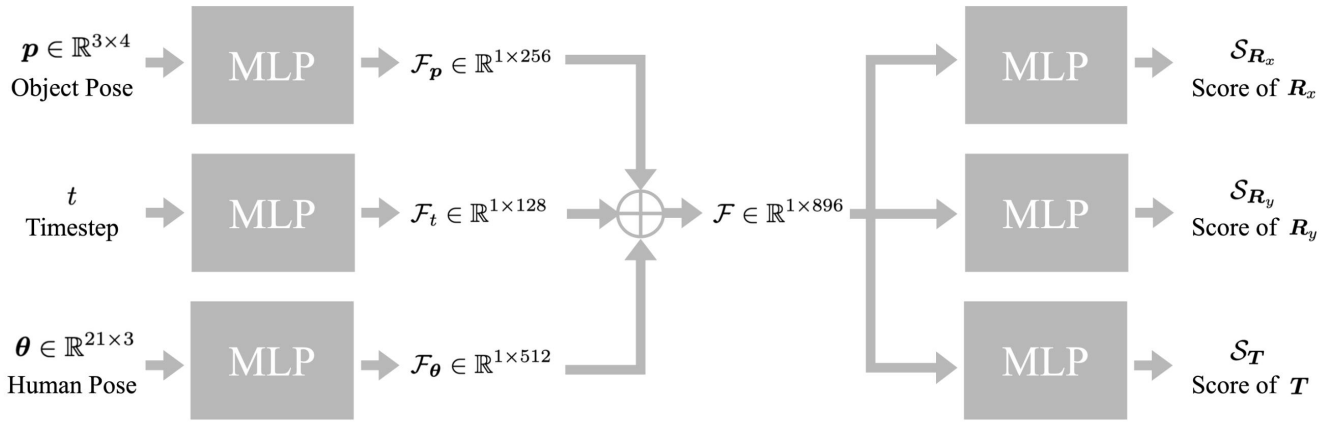

Figure 3. **Architecture of Human Conditioned Object Pose Diffusion Model.** We design a diffusion model that generates a plausible object pose for interacting with a given human pose. Each object pose, time step, and human pose are encoded by MLP. The concatenated features then pass through different MLPs, producing an object pose output consisting of 6D rotation and translation.

$\mathcal{L}_h$, and we define $\mathcal{L}_o$ similarly. Additionally, based on the intuition that the object must be in contact with the human to have movement, we define $\mathcal{L}_{\text{HOI}}$ as the loss, calculated as the average distance of the n closest 3D vertices of the object to the 3D vertices of the human. In practice, we set n to one-third of the total number of vertices in the object mesh. We define the final loss as $\mathcal{L}_{\text{total}} = \mathcal{L}_h + \mathcal{L}_o + \mathcal{L}_{\text{HOI}}$ and optimize the scales of the human and object, s_h, and s_o, to obtain to s_h^*, and s_o^*. To preserve the real-world scale of the human, we fix the human scale and only adjust the object's scale by s_o^*/s_h^*.

A.5. Network Architecture

We describe the network architecture of (1) LoRA for MDM and (2) the Human-Conditioned Object Pose Diffusion Model, which form our DAViD.

LoRA for MDM. To learn concepts through LoRA [6], we model the concepts represented by the samples using text prompts. To ensure that the text effectively models the concepts demonstrated by the given samples, we add LoRA [6] layers to the multi-head attention within the transformer encoder layer of the pre-trained MDM. Specifically, we add four 2-layer MLPs for query, key, value, and output projection, respectively for a single transformer encoder layer, allocating them as a space to learn additional knowledge. We add this to all 8 transformer encoder layers stacked in the transformer encoder.

Human Conditioned Object Pose Diffusion Model. To model the conditional object pose based on the given human pose, we design a score-based diffusion model. We encode the object pose, timestep, and human pose using each MLP, concatenating the feature vectors to construct the total feature. The feature is then fed into three different MLPs, which output the scores for $\boldsymbol{R}_x$, $\boldsymbol{R}_y$, and $\boldsymbol{T}$, where $\boldsymbol{R}_x$ and $\boldsymbol{R}_y$ constitute the 6D rotation representation, and $\boldsymbol{T}$ represents the translation. The overall architecture is shown in Fig. 3.

A.6. Training Details

In this section, we describe the training details of (1) LoRA for MDM and (2) the Human Conditioned Object Pose Diffusion Model, which form our DAViD.

LoRA for MDM. For training the LoRA [6] layer in the pre-trained MDM [20], we create a dataset by extracting only the human motion from previously generated 4D HOI samples and processing it following HumanML3D [5]. The number of training samples varies by object category, ranging from 5 to 50, and we figure out that this amount is sufficient for learning the concept of human motion through LoRA [6]. During training, we freeze all other weights and train only the weights of the LoRA [6] layer. As our concepts are represented in the form of text, we use object category as a text prompt for training our LoRA. For motions with multiple modes (e.g., left and right hand-object interactions), the text prompt is modified by adding tags such as "left_" or "right_" before the main tag. We found that simply adding these additional tags gives controllability to model. We train a total of 500 to 3000 steps (depending on categories) using the Adam [10] optimizer with a learning rate of 1×10^{-4} without decay.

Human Conditioned Object Pose Diffusion Model. For training human conditioned object pose diffusion model, we extract pairwise human pose and object pose from each frame of the 4D HOI Sample and use them as training data. The number of data samples used varies by object category, ranging from 765 to 7,650. We train total of 1000 to 5000 steps (depending on categories) using the Adam [10] optimizer with a learning rate of 5×10^{-3} and a weight decay of 0.99.

Figure 4. **Additional Qualitative Results.** We showcase additional results of our method. We present diverse samples generated from our DAViD, with each frame visualized in temporal order.

Figure 5. **Qualitative Results on FullBodyManip Datatset.** We showcase additional qualitative results of DAViD trained on the FullBodyManip dataset.

B. Experimental Details

B.1. Additional Qualitative Results

We showcase additional qualitative results in Fig. 4 and Fig. 5. In Fig. 4, we show the results of generating various HOI motions using our trained DAViD. Through the results of generating various HOI motions, we demonstrate that our LoRA [6] faithfully learns the dynamic patterns during HOI. In Fig. 5, we show the qualitative results generated from DAViD, trained on the FullBodyManip [14] dataset. We demonstrate that DAViD is not only able to learn coherent and simple HOI patterns, but also capable of generating relatively complex HOI motions.

B.2. Additional Quantitative Results

We report additional comparisons with our baselines for each category of the FullBodyManip dataset in Tab. 1. As our TGS automatically detects potential contact points and guides them closer during sampling, we vary the threshold of potential contact ρ to examine the effect of our guidance. Note that the potential contact threshold ρ used in TGS is lower than the threshold used for evaluating the metric (0.05), to ensure that points not considered as contact by the model are not forced into contact. As shown in Tab. 1, our contact guidance in TGS significantly improves recall and consequently the F1 score. We demonstrate that our contact guidance allows to sample fine-grained hand-object contact on the coarse distribution learned by our object pose diffusion model. In contrast, precision tends to remain stable or decrease as the threshold level increases, which appears to be a side effect of unintended potential contacts detected in the early stage of the denoising preocess. Empirically, we find that maintaining a low threshold around 0.02 minimizes side effects and is effective for sampling fine-grained hand-object contact.

B.3. Scale of the Object

As our human conditioned object pose diffusion model generates plausible object pose for a given human pose, it does not provide information about the object's scale. Since the output human of MDM and the human in the training data both have a uniform scale of 1.0, we automatically determine the appropriate object scale in the generated HOI motion by sampling between the minimum and maximum scales of objects existing in our 4D HOI Samples.

B.4. Generalizability Across Input 3D Objects.

By leveraging pre-trained 2D diffusion models, our 4D HOI sample generation pipeline is scalable to various object

Methods	Clothesstand			Floorlamp			Largebox			Largetable			Monitor			Plasticbox			Smallbox		
	C_{prec} ↑	C_{rec} ↑	F1 ↑	C_{prec} ↑	C_{rec} ↑	F1 ↑	C_{prec} ↑	C_{rec} ↑	F1 ↑	C_{prec} ↑	C_{rec} ↑	F1 ↑	C_{prec} ↑	C_{rec} ↑	F1 ↑	C_{prec} ↑	C_{rec} ↑	F1 ↑	C_{prec} ↑	C_{rec} ↑	F1 ↑
DAViD$_{\rho=0.00}$	0.667	0.088	0.156	0.611	0.297	0.400	0.894	0.578	0.702	0.917	0.500	0.647	0.807	0.38	0.517	0.934	0.463	0.619	0.988	0.464	0.632
DAViD$_{\rho=0.01}$	0.913	0.309	0.462	0.500	0.486	0.493	0.862	0.311	0.457	0.743	0.197	0.311	0.847	0.485	0.617	0.848	0.275	0.415	0.985	0.515	0.676
DAViD$_{\rho=0.02}$	0.765	0.382	0.510	0.531	0.459	0.493	0.889	0.497	0.637	0.929	0.598	0.728	0.799	0.442	0.570	0.865	0.553	0.675	0.986	0.508	0.671
DAViD$_{\rho=0.03}$	0.536	0.221	0.313	0.600	0.243	0.346	0.863	0.547	0.669	0.837	0.545	0.661	0.823	0.523	0.640	0.821	0.471	0.599	0.982	0.545	0.701
CHOIS	0.615	0.353	0.449	0.667	0.378	0.483	0.773	0.211	0.332	0.783	0.136	0.232	0.827	0.156	0.263	0.674	0.119	0.202	0.957	0.239	0.382
DAViD$_{\rho=0.00}$	0.667	0.098	0.171	0.784	0.138	0.235	0.951	0.485	0.642	0.808	0.371	0.509	0.849	0.345	0.491	0.881	0.349	0.500	0.991	0.528	0.689
DAViD$_{\rho=0.01}$	0.429	0.197	0.270	0.615	0.114	0.193	0.979	0.539	0.695	0.839	0.173	0.287	0.895	0.488	0.632	0.824	0.371	0.512	0.990	0.610	0.755
DAViD$_{\rho=0.02}$	0.405	0.279	0.330	0.667	0.276	0.391	0.954	0.488	0.645	0.873	0.456	0.599	0.911	0.602	0.725	0.800	0.446	0.572	0.989	0.649	0.784
DAViD$_{\rho=0.03}$	0.571	0.328	0.417	0.472	0.119	0.190	0.959	0.588	0.729	0.868	0.533	0.661	0.877	0.564	0.687	0.790	0.522	0.629	0.990	0.645	0.781
OMOMO	0.432	0.311	0.362	0.828	0.114	0.201	0.917	0.617	0.738	0.868	0.603	0.711	0.776	0.432	0.555	0.831	0.342	0.484	0.982	0.639	0.774

Methods	Smalltable			Suitcase			Trashcan			Tripod			Whitechair			Woodchair			Average		
	C_{prec} ↑	C_{rec} ↑	F1 ↑	C_{prec} ↑	C_{rec} ↑	F1 ↑	C_{prec} ↑	C_{rec} ↑	F1 ↑	C_{prec} ↑	C_{rec} ↑	F1 ↑	C_{prec} ↑	C_{rec} ↑	F1 ↑	C_{prec} ↑	C_{rec} ↑	F1 ↑	C_{prec} ↑	C_{rec} ↑	F1 ↑
DAViD$_{\rho=0.00}$	0.926	0.482	0.634	0.990	0.564	0.718	0.961	0.508	0.665	0.714	0.195	0.306	0.706	0.213	0.328	0.913	0.328	0.483	0.848	0.389	0.524
DAViD$_{\rho=0.01}$	0.953	0.527	0.679	0.992	0.609	0.755	0.952	0.496	0.652	0.727	0.312	0.436	0.808	0.187	0.303	0.716	0.276	0.398	0.847	0.383	0.511
DAViD$_{\rho=0.02}$	0.915	0.691	0.788	0.979	0.609	0.751	0.916	0.628	0.745	0.731	0.247	0.369	0.825	0.609	0.701	0.779	0.495	0.605	**0.850**	**0.517**	**0.634**
DAViD$_{\rho=0.03}$	0.907	0.717	0.801	0.967	0.649	0.777	0.947	0.893	0.919	0.667	0.208	0.317	0.826	0.591	0.689	0.711	0.448	0.55	0.812	0.508	0.614
CHOIS	0.944	0.431	0.592	0.942	0.251	0.397	0.580	0.165	0.257	0.917	0.143	0.247	0.483	0.062	0.110	0.717	0.224	0.341	0.760	0.221	0.330
DAViD$_{\rho=0.00}$	0.936	0.457	0.614	0.989	0.420	0.590	0.970	0.561	0.711	0.839	0.173	0.287	1.00	0.255	0.407	0.922	0.371	0.529	**0.891**	0.350	0.490
DAViD$_{\rho=0.01}$	0.931	0.502	0.653	0.986	0.501	0.664	0.963	0.630	0.761	0.902	0.306	0.457	0.963	0.265	0.416	0.810	0.406	0.540	0.856	0.392	0.526
DAViD$_{\rho=0.02}$	0.920	0.639	0.755	0.976	0.536	0.692	0.957	0.760	0.847	0.949	0.347	0.508	1.00	0.378	0.548	0.865	0.490	0.618	0.867	0.488	0.616
DAViD$_{\rho=0.03}$	0.945	0.706	0.809	0.985	0.552	0.708	0.956	0.699	0.807	0.967	0.321	0.482	0.974	0.388	0.555	0.774	0.516	0.619	0.856	**0.499**	**0.621**
OMOMO	0.870	0.470	0.610	0.974	0.660	0.787	0.739	0.500	0.596	0.824	0.225	0.354	0.967	0.296	0.453	0.779	0.433	0.557	0.830	0.434	0.552

Table 1. **Additional Quantitative Results.** We report additional quantitative results for each category of the FullBodyManip dataset by varying the contact loss threshold used in our TGS.

categories and instances. As shown in Fig. 7, our pipeline allows to generate 4D HOI samples not only from 3D objects in existing datasets, but also from those reconstructed from in-the-wild images. For the given in-the-wild image, we first reconstruct 3D objects from the image via TRELLIS [21], using them as input to our pipeline to generate the 4D HOI sample interacting with the object.

C. Limitations and Future Work

C.1. Spatial Bias on 2D HOI Image Generation

Due to the internal spatial bias of the pre-trained 2D diffusion model, the model may fail to generate plausible images when structural guidance is introduced in locations that do not align with this bias, leading to collapse or hallucination. For example, if an umbrella, which should be held by a hand, is rendered at the bottom of the image and Canny edges [2] are extracted from it to generate an image, the model may create and use a new umbrella in a different location, rather than in the rendered region. As a future direction, we can consider a new form of conditional image generation that is guided only by the structure of the given object, without guidance on its location in the image. This approach is expected to remove the human labor we used for filtering malicious images.

C.2. Limits of Smoothness Guidance Sampling

Although our human-conditioned object pose diffusion model is trained to generate plausible object pose during interactions given a human pose, our smoothness guidance sampling allows to generate plausible object motion for input sequential human poses. In many cases, the assumption that the object trajectory should be smooth while HOI is valid, but in situations where the object vibrates within a small range (e.g., when drilling a hole with an electric drill) collding with other object, the assumption can be problematic. To naturally model motions involving such collisions, physics information is required, and understanding such physics during HOI can be considered as potential future work.

C.3. Modeling Dexterous Hand-Object Interaction

Although we model the human with SMPL-X [16], and recent 2D diffusion models demonstrate impressive quality in representing detailed hands, the pre-trained video diffusion model and 3D human estimator struggle to uplift the 2D hand from HOI images to high-quality 4D. This hinders the modeling of dexterous hand-object interactions in both our 4D HOI samples and the learned Dynamic Affordance. As a future direction, we can explore separately learning the hand patterns and merging them with the dynamic patterns we learned, with the expectation of improving the hand quality of the sampled HOI motion. As shown in Fig. 6, we show that hand poses can be extended to our 4D HOI samples using a hand pose estimator [17] in a simple scenario.

C.4. Concept Conflict

As we show that our LoRA [6] has an advantage for modeling multiple concepts (*e.g.*, combining existing knowledge of pre-trained model, and combining the knowledge of two individual LoRAs [6]), the concept conflict may appear when combining two different concepts, similar to what occurs in image diffusion models. When the two learned concepts show totally different human motion patterns (*e.g.*, lifting a barbell, pushing a cart), we empirically observe that the result converges into two cases: (1) a motion is interpolated between two concepts, resulting implausible motion or (2)

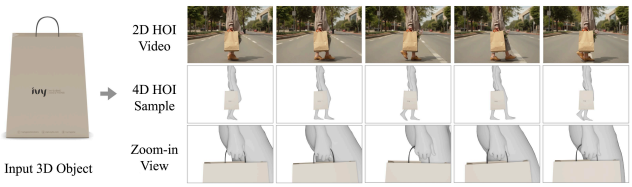

Figure 6. **Hand Pose Extension.** DAVID uses SMPL-X as the human model, allowing hand pose extension in our 4D HOI samples.

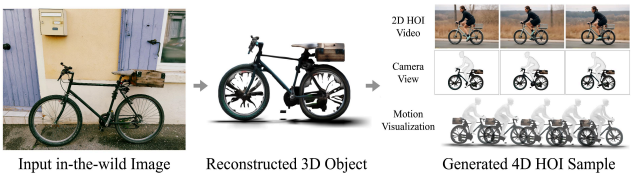

Figure 7. **Generalizability Across Input 3D Objects.** Our 4D HOI sample generation pipeline is generalizable to any input 3D object, including those reconstructed from images.

one motion is performed followed by the other. Instead, when the two concepts are reasonably similar (*e.g.*, holding an umbrella, riding a scooter), their motions can be combined to generate multi-object interactions. However, we find that the relatively less coherent patterns (*e.g.*, the position of the hand while riding a scooter) are removed while combining the concepts, which is the limitation of our application.

References

[1] Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R. Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. In *arXiv:2410.02073*, 2024. 2

[2] John Canny. A computational approach to edge detection. In *IEEE TPAMI*, 1986. 1, 5

[3] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *arXiv:2403.03206*, 2024. 1

[4] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. In *Commun. ACM*, 1981. 2

[5] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *CVPR*, 2022. 3

[6] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022. 3, 4, 5

[7] Nikita Karaev, Iurii Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker3: Simpler and better point tracking by pseudo-labelling real videos. In *arXiv:2410.11831*, 2024. 2

[8] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together. In *ECCV*, 2024. 2

[9] Hyeonwoo Kim, Sookwan Han, Patrick Kwon, and Hanbyul Joo. Beyond the contact: Discovering comprehensive affordance for 3d objects from pre-trained 2d diffusion models. In *ECCV*, 2024. 1

[10] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *arXiv:1412.6980*, 2014. 3

[11] Kling AI. https://klingai.com/, (accessed Jul 20th, 2025). 1

[12] Black Forest Labs. Flux. https://bfl.ai/, (accessed Jul 20th, 2025). 1

[13] Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. Epnp: An accurate o(n) solution to the pnp problem. In *IJCV*, 2009. 2

[14] Jiaman Li, Jiajun Wu, and C Karen Liu. Object motion guided human motion synthesis. In *ACM TOG*, 2023. 4

[15] OpenAI. Gpt-4o system card. https://openai.com/research/gpt-4v-system-card, (accessed Jul 20th, 2025). 1

[16] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, 2019. 5

[17] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3d with transformers. In *CVPR*, 2024. 5

[18] Scott D. Roth. Ray casting for modeling solids. In *Comput. Graph. Image Process.*, 1982. 2

[19] Zehong Shen, Huaijin Pi, Yan Xia, Zhi Cen, Sida Peng, Zechen Hu, Hujun Bao, Ruizhen Hu, and Xiaowei Zhou. World-grounded human motion recovery via gravity-view coordinates. In *Proc. ACM SIGGRAPH Asia*, 2024. 1, 2

[20] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *ICLR*, 2023. 3

[21] Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. In *CVPR*, 2025. 5

[22] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 1