This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

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

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

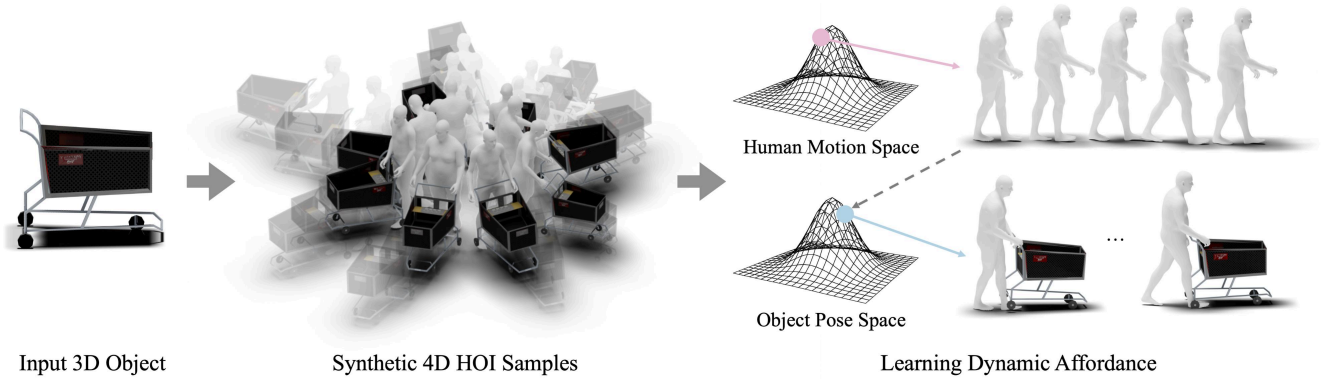

Figure 1. **DAViD for Learning Dynamic Affordance.** Given an input 3D object (left), our method, DAViD, learns Dynamic Affordance to model the dynamic patterns that occur during HOI (right). To build it, we present a framework to synthesize diverse 4D HOI samples (middle) by leveraging a pre-trained video diffusion model.

Abstract

Modeling how humans interact with objects is crucial for AI to effectively assist or mimic human behaviors. Existing studies for learning such ability primarily focus on static human-object interaction (HOI) patterns, such as contact and spatial relationships, while dynamic HOI patterns, capturing the movement of humans and objects over time, remain relatively underexplored. In this paper, we present a novel framework for learning Dynamic Affordance across various target object categories. To address the scarcity of 4D HOI datasets, our method learns the 3D dynamic affordance from synthetically generated 4D HOI samples. Specifically, we propose a pipeline that first generates 2D HOI videos from a given 3D target object using a pre-trained video diffusion model, then lifts them into 3D to generate 4D HOI samples. Leveraging these synthesized 4D HOI samples, we train DAViD, our generative 4D human-object interaction model, which is composed of two key components: (1) a human motion diffusion model (MDM) with Low-Rank Adaptation (LoRA) module to fine-tune a pre-trained MDM to learn the HOI motion concepts from limited HOI motion samples, (2) a motion diffusion model for 4D object poses conditioned by produced human interaction motions. Interestingly, DAViD can integrate newly learned HOI motion concepts with pre-trained human motions to create novel HOI motions, even for multiple HOI motion concepts, demonstrating the advantage of our pipeline with LoRA in integrating dynamic HOI concepts. Through extensive experiments, we demonstrate that DAViD outperforms baselines in synthesizing HOI motion.

1. Introduction

Humans effortlessly understand how to effectively use objects to achieve their goals within a given environment. This knowledge, known as affordance originally introduced by Gibson [22], encapsulates a comprehensive understanding of the inherent ways humans interact with objects. Modeling these patterns is crucial for AI systems to assist or mimic human behavior, leading to extensive research in AI fields [1, 11–14, 17, 37, 40, 41, 54, 56, 58, 62–66, 72, 73, 76, 79, 80, 86, 93, 98, 114, 122–124]. However, existing studies on affordance learning primarily focus on static human-object interaction (HOI) patterns. For example, some approaches address affordances by inferring contact (or interaction) region between humans and objects in 2D images [2, 50], 3D humans [91, 111], and 3D objects [110, 111]. Others [27, 38] represent affordances as spatial distributions aggregated from static 3D HOI samples.

While recent methods generate human motions including HOI [46, 51, 52, 71, 105, 107, 108] and Human-Scene Interaction (HSI) [101, 105, 113], they often focus on limited actions such as standing, sitting, or picking, without focusing on rich, dynamic nature of object-specific affordances in human motion.

In this paper, we present a novel framework for learning Dynamic Affordance across diverse and unbounded target object categories, without the need for laborious 3D HOI data capture. Our approach is motivated by the observation that human-object interactions involve not only static properties (e.g., contact regions or grasping patterns) but also object-specific characteristic movements over time. For instance, as shown in Fig. 2, humans naturally push a shopping cart and pull a suitcase while walking forward, yet rarely lift their handles vertically. Existing studies, which mainly focus on static affordance features, fail to capture these dynamic interaction patterns. To address this, we present DAViD, a generative motion model for modeling Dynamic Affordance that learns the dynamic and object-specific 4D HOI patterns. To handle the vast diversity of dynamic HOI patterns, our method considers each object-specific 4D HOI pattern as a distinct motion concept, adapting a pre-trained motion diffusion model (MDM) [89] via Low-Rank Adaptation (LoRA) [32]. Once human motions are generated by our MDM with LoRA, we further synthesize object motions conditioned by human motion using a score-based diffusion model, where we introduce a novel sampling technique, named Temporal Guidance Sampling (TGS), for enhancing motion consistency and fine details. Notably, our approach not only models dynamic affordances from limited data but also seamlessly blends newly learned HOI motion concepts with those already learnt in the pre-trained model, enabling zero-shot generalization, even across multiple HOI concepts.

To train DAViD, we also present a pipeline to synthetically produce realistic 4D HOI samples for a target object category, enabling effective model training without requiring expensive real-world 3D data collection. Most existing dynamic HOI learning [34, 99, 117] or mimicking approaches [109] rely on the 3D datasets captured in lab-controlled environments [4, 19, 29, 33, 39, 87, 120]. This significantly limits the opportunity to learn diverse and distinctive dynamic affordance patterns across a broader range of object categories. Our key insight is that pre-trained video generation models inherently capture dynamic patterns of object usage. As shown in Fig. 2, these models already know plausible dynamic interactions (*e.g.*, pushing a shopping cart or pulling a suitcase). Since it presents non-trivial challenges to leverage the synthetic 2D videos for learning 4D dynamic affordance, we propose a novel pipeline that synthesizes plausible 4D HOI samples for a given 3D object mesh. We achieve this by generating 2D videos aligned to the 3D object, and aligning the uplifted 4D human motion sequences with the object model, guided by synthetic video cues.

Figure 2. **Dynamic Patterns in HOI.** People show dynamic patterns during HOI. Pre-trained video diffusion models have the knowledge of the dynamic patterns.

We validate the efficacy of our approach by learning Dynamic Affordance on 30 3D object categories, gathered from multiple sources [4, 10, 33, 38, 61, 84, 106]. We compare our DAViD with other HOI motion synthesis approaches [51, 52, 89] against public dataset [51], demonstrating that our model outperforms competitors in generating HOI motion with plausible contact. We further demonstrate the advantage of our pipeline, by generating results for a various previously unexplored target objects, and also validating its advantage in merging multiple HOI motion concepts with known human motion from pre-trained MDM [89].

In summary, our main contributions are as follows: (1) We present a generative motion model, DAViD, to learn Dynamic Affordance; (2) We present a pipeline to synthetically generate diverse 4D HOI samples leveraging video diffusion models; (3) We address generating various novel HOI motions by fusing multiple, separately trained LoRAs with known human motions from a pre-trained model. The results and our source code are publicly available.

2. Related Work

Reasoning Visual Affordance. Affordance is a concept defined by Gibson [22] as the set of possible interactions that an agent can perform within an environment. As the concept narrows to the primary agent (or humans) and the specific interaction target (or objects), the area of learning affordance mainly focuses on HOI. Specifically, followed by the intuition that the visual geometry of an object is closely related to its functionality, many studies leverage visual cues to reason about object's affordance. The studies are subdivided into various areas, including affordance classification [41, 58, 73, 80, 86, 93, 94], affordance detection [1, 12, 17, 62, 64–66, 79], and affordance segmentation [13, 37, 56, 76]. Such traditional studies infer affordance for the given context (mostly provided as images), as a form of action labels [6, 31, 48], and contact (or interaction) regions in 2D [2, 50] or 3D [91, 110]. Later studies model static HOI patterns (*e.g.*, contact) during interactions [15, 27, 38], going beyond the understanding of individual HOI situations. Recent studies synthesize human

motion including HOI [52, 71, 101, 105, 107, 108, 113], showing potential to learn dynamic HOI patterns, but they focus on a text-driven approach, generating consistent human motions regardless of object categories (*e.g.*, dragging a chair, holding a backpack). In contrast, our method focuses on learning rich, dynamic object-specific HOI patterns that are observable in real-world HOIs.

Learning from Synthesized Data. Generative models such as GANs [23] and diffusion models are used as a strategy to address the problem of insufficient real-world data in many fields. Starting from augmentation [3, 8, 30, 67, 90, 104], synthetic datasets are widely used in vision area including segmentation [59, 92, 119], classification [9, 57, 88], shape reconstruction [69], and neural rendering [28, 118]. While data synthesis is a powerful technique with broad applications, a key consideration when learning from generated data is its reliability compared to real-world data. Recently, a few studies in the HOI field [27, 38] leverage pre-trained 2D diffusion models to generate 2D HOI images, demonstrating the reliability of synthetic data to learn affordance knowledge. While following the approaches, we focus on learning dynamic patterns during HOI rather than learning static patterns.

Learning Concepts in Diffusion Models. Concept learning, also known as concept customization, is a strategy for extending the space of the pre-trained model to understand specific concepts (*e.g.*, subjects, styles). Some studies tune the embeddings [21, 96] corresponding to a given concept in a pre-trained diffusion model, while others [44, 77] tune both the model weights and embeddings together. Recently, LoRA [32], originally used in large language models (LLMs), shows impressive results in producing images well aligned to the given concept [78], leading to an increase of attempts [26, 81, 83] to use LoRA for concept learning. Such concepts can be provided in various forms, including images [24, 53, 55, 121], videos [26, 102], and audios [97]. However, concept learning in fields beyond images, video, and audio is relatively less explored. We apply LoRA to the human motion space of pre-trained MDM [89] to learn the dynamic patterns displayed in HOI as a single concept, extending the original human motion space in terms of HOI.

3. Method

Our approach aims to learn the Dynamic Affordance of a target object from synthetically generated 2D videos, which are produced by a video diffusion model [42]. In the first stage, we generate diverse 4D HOI samples interacting with a given 3D object mesh by leveraging pre-trained video models (Sec. 3.1). Then, our method models the collected 4D HOI samples into our DAViD, which consists of (1) a motion diffusion model along with LoRA, and (2) an object pose diffusion model. The first module synthesizes feasible human motions for the target object, and the second module synthesizes object motions paired with previously generated human motion, using our novel sampling technique, Temporal Guidance Sampling (TGS) (Sec 3.2). An overview of our method is shown in Fig. 3.

3.1. 4D HOI Sample Generation

Our pipeline begins with a 3D object mesh input, and produces diverse 2D HOI videos showing how a person interacts with (or uses) the target object.

2D HOI Video Synthesis. Given an input 3D object mesh, we produce multiple 2D HOI videos capturing human-object interactions with the object. We first render the object from cameras with multiple viewpoints. The intrinsic parameter is fixed to match the same perspective camera model and parameters of the world-grounded human mesh recovery method (GVHMR [82]) used in our later 3D lifting process, ensuring a consistent camera setup. For the extrinsic parameters, we consider diverse viewpoints based on the object's type, either ground-placed or portable objects. For relatively stationary objects (*e.g.*, cart), we consider horizontal viewpoint variations centered on the object, while for the portable objects (*e.g.*, trumpet) we sample viewpoints within a specified range from all view angles. The object type is automatically inferred via a vision-language model [68], by querying with the rendered image with a prompt: *"Is the object plausibly on the ground while being used?"*.

Once the object is rendered from each view, we synthesize a realistic human naturally interacting with it using an off-the-shelf pre-trained 2D image diffusion model [45]. Unlike the previous similar approaches [38, 43, 112] that use inpainting by providing a mask region to a model, we use ControlNet [116] to synthesize the entire image, including the human and the plausible background, by using edge cues of rendered object [7] as the condition. Our primary objective is to synthesize both plausible human-object interaction for the given object viewpoint and a realistic background, which is important for estimating dynamic cues and camera poses in a global coordinate. After synthesizing HOI images from several views, we refine the results by filtering out low-quality or undesirable images. The filtering process can be performed efficiently with minimal human effort, requiring only about three minutes per category.

Finally, we synthesize corresponding 2D HOI videos $\{\mathcal{V}_d\}_{d=1}^{N}$ from each 2D HOI images $\{\mathcal{I}_d\}_{d=1}^{N}$ by using a pre-trained image-to-video generation model [42] with the same text prompt used for generating images[1]. Notably, our synthesized 2D HOI videos have key advantages over in-the-wild 2D real videos in learning dynamic affordance: (1) our HOI videos contain diverse views and interaction scenarios for the input 3D object, (2) we have precise camera parameters and 3D object pose information for the initial

[1]As we synthesize several images at each camera view, the number of videos N is not necessary the same as camera view number, C.

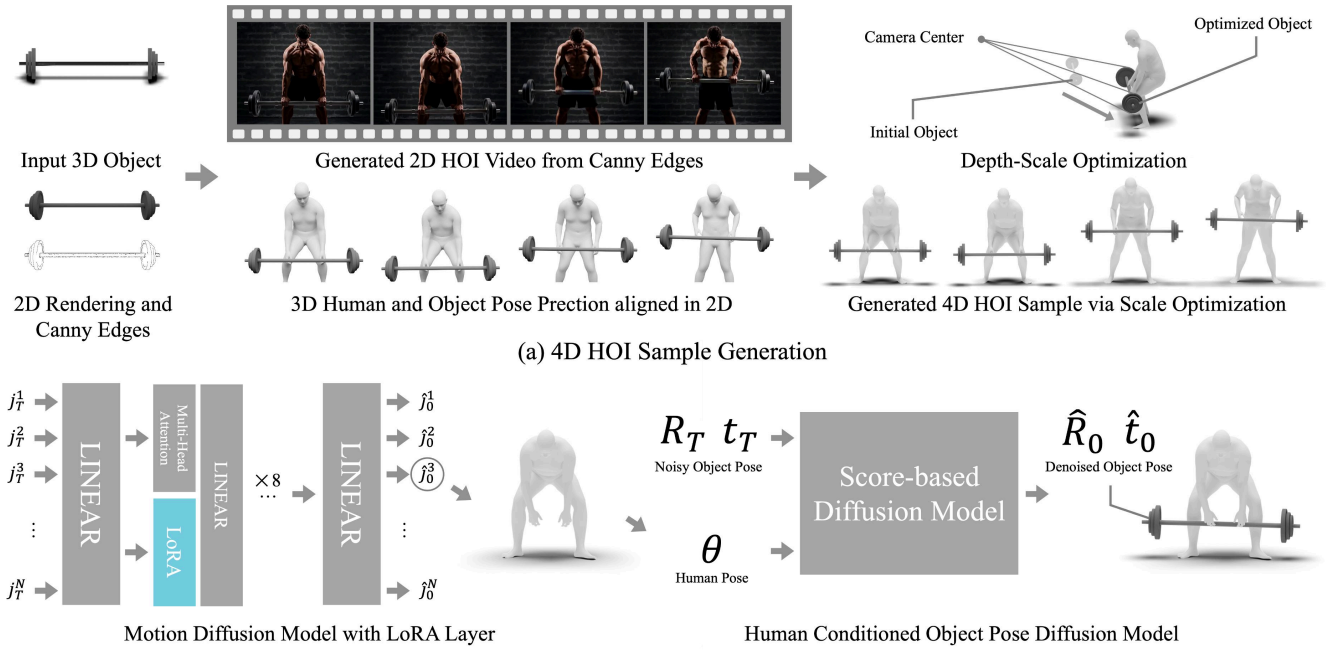

Figure 3. **Overview.** Our method consists of two parts: (1) Generating 4D HOI Samples and (2) Learning Dynamic Affordance from Generated 4D HOI Samples. First, we create 2D HOI videos, and generate 4D HOI samples with our uplifting pipeline. Then, the generated 4D HOI samples are used to train DAViD, learning the patterns of human motion and object pose.

frame of each video. These are important factors for reliable 4D HOI sample generations.

Uplifting for 4D HOI Sample Generations. For each 2D HOI video $\mathcal{V}$ [2], we uplift 3D human motions and 3D object movements to generate 4D HOI samples. Our key strategy is to uplift each component (human, object) by finding each motion relative to each camera, and co-locate them in the common camera coordinate shared by both components.

For the object side, we use the Perspective-n-Points (PnP) [20, 49] algorithm to find the relative camera pose with respect to the given 3D object mesh. Specifically, we aim to compute the extrinsic camera poses $\{\boldsymbol{R}_o^f, \boldsymbol{t}_o^f\}_{f=1}^F$, where the one for the first frame $f = 1$ is already known. We first identify visible object mesh vertices $\{\mathbf{V}_o\}_{\text{vis}}$ via raycasting [75] for the first frame $f = 1$, where $\mathbf{V}_o \in \mathbb{R}^3$ is an object vertex. The 2D projections of the visible vertices at the first frame is denoted as $\{\boldsymbol{v}_o^{f=1}\}_{\text{vis}}$, where $\boldsymbol{v}_o^{f=1} \in \mathbb{R}^2$. By applying an off-the-shelf 2D video tracker [35, 36], we obtain the corresponding 2D points $\boldsymbol{v}_o^f$ in the remaining frames $f > 1$. As we have the correspondences between 3D object vertices and matched 2D points for every frame, we apply PnP [20, 49] to compute the camera poses at frame f with respect to the object mesh in its canonical space:

$$(\boldsymbol{R}_o^f, \boldsymbol{t}_o^f) = \text{PnP}\left(\{\boldsymbol{v}_o^f\}_{\text{vis}}, \{\boldsymbol{V}_o\}_{\text{vis}}\right). \tag{1}$$

For the human side, we leverage world-grounded human motion recovery models, denoted as $\mathbf{F}_{\text{human}}$ to uplift human motions and corresponding camera motions into 3D from the synthesized 2D HOI videos. We use GVHMR [82], except in the cases where human translation occurs without walking motions (*e.g.*, riding a motorcycle). In such cases, we use TRAM [100], as GVHMR [82] typically fails to predict plausible camera motion under these conditions. Therefore,

$$\mathbf{F}_{\text{human}}(\mathcal{V}) = \{\boldsymbol{\theta}^f, \boldsymbol{\beta}^f, \boldsymbol{\phi}^f, \boldsymbol{\tau}^f, \boldsymbol{R}_h^f, \boldsymbol{t}_h^f\}_{f=1}^F, \tag{2}$$

where $\boldsymbol{\theta}^f \in \mathbb{R}^{21\times3}$, $\boldsymbol{\beta}^f \in \mathbb{R}^{10}$, $\boldsymbol{\phi}^f \in \text{SO}(3)$, $\boldsymbol{\tau}^f \in \mathbb{R}^3$ are the predicted SMPL-X [70] parameters of human pose, shape, root joint's rotation, and translation for f-th frame, respectively. The $\boldsymbol{R}_h^f \in \text{SO}(3)$, $\boldsymbol{t}_h^f \in \mathbb{R}^3$ are the extrinsic parameters of estimated camera poses. Note that all of the parameters are defined in a world coordinate for the human mesh space, which is different from object mesh space.

Intuitively, the human and object at time f are observed from the same camera, and thus we can simply co-locate them in the same space, by transferring each component to its camera coordinate, where the camera becomes the origin:

$$\hat{\boldsymbol{V}}_o^f = \boldsymbol{R}_o^f \boldsymbol{V}_o^f + \boldsymbol{t}_o^f, \tag{3}$$

$$\hat{\boldsymbol{V}}_h^f = \boldsymbol{R}_h^f \boldsymbol{V}_h^f + \boldsymbol{t}_h^f, \tag{4}$$

where $\boldsymbol{V}_h^f \in \mathbb{R}^3$ is a human 3D vertex in its global coordinate after applying all SMPL parameters, and $\hat{\boldsymbol{V}}_o^f$ and $\hat{\boldsymbol{V}}_h^f$ are the human and object 3D vertices in the shared camera

[2] we drop the video index d for simplicity.

coordinate. To this end, we align the object-side cameras to the human-side cameras, bringing cameras, humans, and objects into a unified coordinate system of human side.

Resolving Depth-Scale Ambiguity. Although we uplift and place human and object movements in a common 3D space, there still exists the depth-scale ambiguity, since each components are defined in arbitrary object scales, as shown in depth-scale optimization part in Fig. 3. To resolve the ambiguity, we optimize the scale of the human ($s_h \in \mathbb{R}$) and object ($s_o \in \mathbb{R}$) in the camera coordinate at the first frame, using (1) a weak depth cue from the depth map, and (2) a contact cue from the object movement assumption. The objective function is defined as follows:

$$\mathcal{L}_{\text{total}} = \lambda_h \mathcal{L}_h + \lambda_o \mathcal{L}_o + \lambda_{\text{HOI}} \mathcal{L}_{\text{HOI}} + \lambda_{\text{col}} \mathcal{L}_{\text{collision}}, \quad (5)$$

where λs denote the weighting factors. The $\mathcal{L}_h$ and $\mathcal{L}_o$ are the depth consistency loss for human and object, defined as:

$$\mathcal{L}_h = \sum_{i \in \text{vis}_\text{h}} \left\| s_h \hat{\boldsymbol{V}}_{h,i} - \boldsymbol{P}_{h,i} \right\|^2 \quad (6)$$

$$\mathcal{L}_o = \sum_{j \in \text{vis}_\text{o}} \left\| s_o \hat{\boldsymbol{V}}_{o,j} - \boldsymbol{P}_{o,j} \right\|^2, \quad (7)$$

where $\hat{\boldsymbol{V}}_{h,i}$ and $\hat{\boldsymbol{V}}_{o,i}$ are the i-th human and object vertices in the camera coordinate at the first frame respectively, and $\boldsymbol{P}_{h,\text{3D}}$, $\boldsymbol{P}_{o,\text{3D}}$ are the corresponding 3D points by back-projecting 2D metric depths (computed by metric depth model [5]) into the camera coordinate. Note that the indices i, j are from the set of visible vertex indices of the human and object at the first frame.

Additionally, the loss term of the weak HOI contact cue is defined as follows:

$$\mathcal{L}_{\text{HOI}} = \mathcal{D}_n \left(\{ s_h \hat{\boldsymbol{V}}_{h,i} \}, \{ s_o \hat{\boldsymbol{V}}_{o,j} \} \right), \quad (8)$$

where $\mathcal{D}_n$ is the function to compute the average distance of the n closest points between two points sets. Intuitively, the cost term encourages contact between the human and object, assuming they are in contact at the first frame. Note that n is a hyperparameter that varies across categories. We apply the collision term to avoid undesired penetration using COAP [60]. The object vertices are used as query points, reducing the collision between human and object in $\mathcal{L}_{\text{collision}}$.

3.2. Learning Dynamic Affordance

Based on the diverse dynamic 4D HOI samples we synthesized for a target 3D object, we build a generative HOI motion model, DAViD, to model dynamic affordance. Our DAViD consists of two key components: (1) a generative human motion model by fine-tuning a pre-trained MDM [89] with object-specific LoRA [32] module, and (2) a score-based diffusion model for estimating object poses conditioned by generated human poses. For the score-based diffusion model, we introduce a novel inference-time sampling technique, TGS to enhance overall quality of motion.

Human Motion Synthesis with LoRA for MDM. To capture human motion patterns interacting with a specific object, we insert a LoRA [32] layer into a pre-trained MDM [89] model and train it for each object category. By leveraging LoRA [32], our model can learn previously unseen human motion patterns with a limited amount of training samples, keeping the original motion synthesis quality of the pre-trained model. Interestingly, we find that merging LoRA [32] layers trained across different object categories enables our model to generate human motion interacting with multiple objects in a zero-shot manner (see Fig. 5 (c)).

For training, we extract only the human motion from our 4D HOI samples, excluding object motions. We preprocess the motion data following HumanML3D [25], and obtain the training data of human joints denoted as $\boldsymbol{j} \in \mathbb{R}^{F \times 22 \times 3}$. The LoRA [32] layer is then added into the multi-head attention layer of the transformer encoder block in text-conditioned MDM [89] as shown in Fig. 3. Our model is trained with the following joint recovery loss function.

$$\mathcal{L}(\Delta\phi) = \mathbb{E}_{\boldsymbol{j}_0, t} \| \boldsymbol{j}_0 - \mathcal{M}_{\phi + \Delta\phi}(\boldsymbol{j}_t, t, c) \|_2^2, \quad (9)$$

where ϕ is the original weight of pre-trained MDM [89] denoted as $\mathcal{M}$, c is the encoded text prompt using CLIP encoder [74], and $\Delta\phi$ is the weight of our LoRA [32] module we want to obtain. For the training text prompt, we use the category name and annotate the additional hand index (left or right) for categories with hand-object interactions to improve controllability. See Supp. Mat. for more details.

Object Motion Synthesis from Human Motion. After generating plausible human motion for a given 3D object, we synthesize a corresponding 3D object motion conditioned on the human motion. To achieve it, we design our object motion module with two key parts: (1) a score-based diffusion model for modeling the distribution of object pose conditioned on the corresponding 3D human pose in HOI, (2) an inference method to estimate the final motion trajectory across a sequence of frames, with a novel sampling technique, named Temporal Guidance Sampling (TGS), for enhancing motion consistency and fine contact details.

To train the score-based diffusion model, we first normalize each frame of the generated 4D HOI sample by adjusting the translation so that the pelvis joint aligns with the origin, and rotate the root orientation to align with z-axis of the world coordinate. The transformations are applied to both the human and the object. A pair of the SMPL-X [70] body pose parameters $\boldsymbol{\theta} \in \mathbb{R}^{21 \times 3}$ and the normalized object pose $\boldsymbol{p} \in \mathbb{R}^9$ (consists of 6D rotation and 3D translation) are used for training our diffusion model. Using the pre-processed data samples, we train our score-based diffusion model Ψ to output the score of the conditional object pose

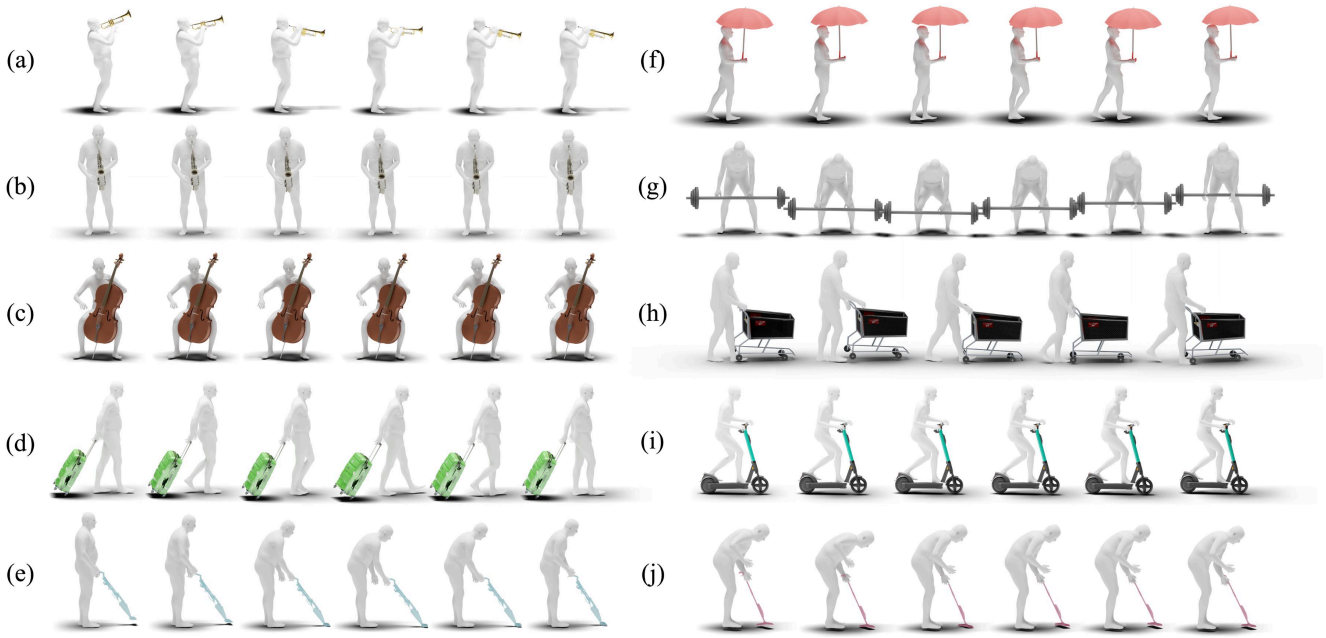

Figure 4. **Qualitative Results.** DAViD models the dynamic patterns of both a human and an object for various object categories. We visualize HOI motions for various objects generated by our DAViD, where frames are visualized in temporal order.

$(\boldsymbol{p}_t)$ distribution for all timestep t:

$$\Psi(\boldsymbol{p}, t|\boldsymbol{\theta}) = \nabla_{\boldsymbol{p}} \log p(\boldsymbol{p}_t|\boldsymbol{\theta}). \tag{10}$$

As demonstrated in Pascal et al. [95], the objective can be achieved by minimizing the following loss function.

$$\mathcal{L} = \mathbb{E}_{t\sim\mathcal{U}(\epsilon,1)}\lambda(t)\mathbb{E}_{\boldsymbol{p},t}\left\|\Psi(\boldsymbol{p}_t, t|\boldsymbol{\theta}) - \frac{\boldsymbol{p}_0 - \boldsymbol{p}_t}{\sigma(t)^2}\right\|_2^2 \tag{11}$$

$$\sigma(t) = \sigma_{\text{min}}\left(\frac{\sigma_{\text{max}}}{\sigma_{\text{min}}}\right)^t, \tag{12}$$

where $\sigma_{\text{min}} = 0.1$, $\sigma_{\text{max}} = 50$ are hyperparameters about noise perturbation. For the detailed structure of our score model, we follow the architecture of GenPose [115].

Based on our conditional object pose diffusion model, we present an inference method to estimate the entire object motion sequence, rather than sampling object poses independently per frame. This process is further enhanced by our Temporal Guidance Sampling (TGS) technique, which encourages temporal consistency and human-object contact during inference. Specifically, we estimate the object pose sequence over F frames $\boldsymbol{p}^{1:F} \in \mathbb{R}^{F\times 9}$, conditioned on the corresponding body motion sequence $\boldsymbol{\theta}^{1:F} \in \mathbb{R}^{F\times 21\times 3}$. Instead of relying solely on the estimated score from the model, we introduce additional guidance term into the Probability Flow ODE [85] as follows:

$$\frac{d\boldsymbol{p}^{1:F}}{dt} = -\sigma(t)\dot{\sigma}(t)\Psi(\boldsymbol{p}^{1:F}, t|\boldsymbol{\theta}^{1:F}) + \nabla_{\boldsymbol{p}^{1:F}}\mathcal{L}_{\text{TGS}}. \tag{13}$$

The $\nabla_{\boldsymbol{p}^{1:F}}\mathcal{L}_{\text{TGS}}$ is the gradient of the object pose at time step t on the following TGS loss, $\mathcal{L}_{\text{TGS}}$:

$$\mathcal{L}_{\text{TGS}}(\boldsymbol{p}^{1:F}) = \lambda_{\text{s}}\mathcal{L}_{\text{smooth}} + \lambda_{\text{c}}\mathcal{L}_{\text{contact}}, \tag{14}$$

$$\mathcal{L}_{\text{smooth}}(\boldsymbol{p}^{1:F}) = \left\|\boldsymbol{p}^{2:F} - \boldsymbol{p}^{1:F-1}\right\|_2^2, \tag{15}$$

$$\mathcal{L}_{\text{contact}}(\boldsymbol{p}^{1:F}) = \mathcal{D}_{\text{chamfer}}\left(\mathcal{V}_h, \mathcal{V}_o(\boldsymbol{p}^{1:F})\right), \tag{16}$$

where $\mathcal{L}_{\text{smooth}}$ encourages temporal consistency between consecutive frames. The $\mathcal{L}_{\text{contact}}$ encourages proximity between human and object vertices, measured by the Chamfer Distance $\mathcal{D}_{\text{chamfer}}$ between the human vertices $\mathcal{V}_h$ and object vertices $\mathcal{V}_o$ positioned by the current pose $\boldsymbol{p}^{1:F}$. In practice, we use a certain threshold to focus on nearby points only when computing the Chamfer Distance. We utilize RK45 ODE solver [18] to solve the ODE trajectory.

4. Experiments

In this section, we conduct experiments to evaluate our method. In Section 4.3, we show 4D HOI motions generated by DAViD for various objects, demonstrating the efficacy of DAViD in learning dynamic HOI patterns. In Section 4.4, we conduct a quantitative comparison with other baselines and a quantitative ablation study on our design choices.

4.1. Datasets

For qualitative evaluation, we consider 30 input 3D object meshes from various sources, obtained from BEHAVE [4], InterCap [33], ShapeNet [10, 61], SAPIEN [106], FullBody-Manipulation Dataset [51], and SketchFab [84]. For quantitative evaluation, we use FullBodyManipulation Dataset [51], which consists of the motion of both humans and objects.

4.2. Baselines and Metric

Even though our goal is to learn dynamic HOI patterns for various objects, which is slightly different from typical text-driven HOI motion synthesis [16, 71, 103, 107, 108, 113], we conduct quantitative comparisons on existing datasets for evaluation. Specifically, we compare DAViD against OMOMO [51] and CHOIS [52] on FullBodyManipulation Dataset [51], which provide available code and evaluate the quality of HOI, unlike other stuides [16, 71, 103] that focus on evaluating human quality solely. However, OMOMO [51] and CHOIS [52] generate human motion conditioned on object motion (or waypoints), making a direct comparison with ours challenging. To address this, we compare the HOI motion generated by baselines using object motion (or waypoints) condition with the HOI motion generated by DAViD using human motion condition. Each condition is extracted from the test dataset of FullBodyManipulation [51], enabling a comparison with the ground truth. Due to the different conditions, we compare only the quality of HOI for the fairness, specifically the precision, recall, and F1 score of the contact, following baselines [51, 52]. As DAViD performs modeling of object poses during interaction, we evaluate by cropping each test dataset sequence from the start of the interaction (contact start) to the end (contact end). Additionally, to assess the human motion quality, we report the Foot Sliding (FS) and foot height (in centimeters) of the human motion generated from DAViD and the baselines, both trained on FullBodyManipulation Dataset [51].

4.3. Qualitative Results

4D HOI Generation for Various Objects. Different from previous studies, we leverage the pre-trained video diffusion model, which allows to learn dynamic patterns of HOI for unbounded object categories. Fig. 4 visualizes the HOI motion for various objects generated by DAViD. The frames are visualized in temporal order. The results show that our method effectively models the dynamic patterns in HOIs. In Fig. 4 (a), Fig. 4 (b), Fig. 4 (c) we observe that the human motion and the relative position of the object are well modeled in relatively static scenarios. In cases with more dynamic human motion, we find that various interaction patterns, including pulling−Fig. 4 (d), pushing−Fig. 4 (h), lifting−Fig. 4 (g), and holding−Fig. 4 (f) are well aligned with the typical use of objects of humans. Also, we can observe that our method can effectively capture complex human motions and their corresponding object motions, such as back and forth movement, as shown in Fig. 4 (e) and Fig. 4 (j). Importantly, unlike previous studies where motion translation typically resulted from walking, we observe that our model effectively captures cases of translation changes without foot movement, as shown in Fig. 4 (i).

Combining Multiple HOI Concepts. As we model dynamic patterns of human motion using LoRA [32] on a pre-trained MDM [89], the existing knowledge of MDM [89] can be combined with the newly learned concepts of HOI. Fig. 5 (a) and Fig. 5 (b) show the novel HOI motion generated by combining the pre-trained MDM's previously known jumping and walking concepts with the newly learned HOI concept; umbrella, and coffee. By simply adding LoRA tags along with the text prompt, we can combine the newly learned HOI concept with the existing motions known by pre-trained MDM to create a novel human motion. The corresponding object motion is then generated based on the human motion, forming the novel HOI motion.

This approach goes beyond combining a single HOI motion concept as shown in Fig. 5 (c), and allows to combine multiple HOI motion concepts which share a similar body structure. We merge two separately trained LoRA (suitcase, umbrella) parameters into the original network by a weighted summation, combining the two concepts into a single network. The object motions are then generated similarly based on the human motion.

Ablation on Temporal Guidance Sampling. To verify the efficacy of our TGS, we conduct an ablation study. Fig. 6 shows six continuous frames of HOI motion with a suit-

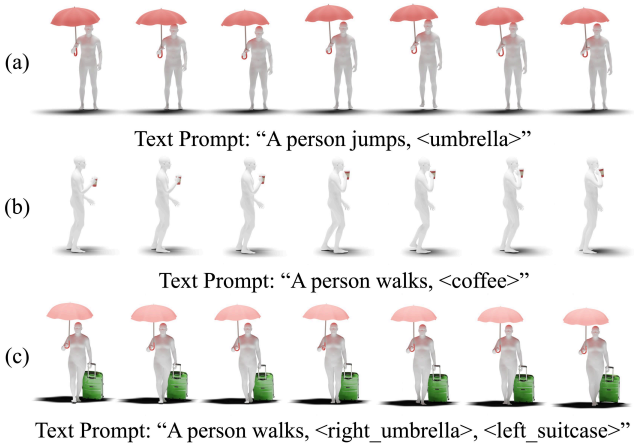

Figure 5. **Combining Concepts.** Our HOI motion concepts can be combined with original knowledge of pre-trained MDM, even for multiple HOI motion concepts.

Figure 6. **Ablation on Temporal Guidance Sampling.** Our guidance during the inference stage reduces object jitter and non-contact interactions, enabling the generation of plausible HOI motion.

Methods	C_{prec} ↑	C_{rec} ↑	F1 Score ↑	FS ↓	H_{feet} ↓
CHOIS [52]	0.760	0.221	0.330	0.472	3.47
$DAViD_{w/o\ contact\ guidance}$	0.848	0.389	0.524	**0.234**	**2.30**
$DAViD_{Full}$	**0.850**	**0.517**	**0.634**		
OMOMO [51]	0.830	0.434	0.552	0.255	2.58
$DAViD_{w/o\ contact\ guidance}$	**0.891**	0.350	0.490	**0.247**	**2.35**
$DAViD_{Full}$	0.867	**0.488**	**0.616**		
MDM [89]	N/A	N/A	N/A	0.260	2.48

Table 1. **Quantitative Results.** We evaluate the quality of generated HOI motion against FullBodyManipulation Dataset.

Methods	Jitter →
DAViD	**30.9**
$DAViD_{w/o\ smoothness\ guidance}$	2.95×10^3
FullBodyManip DB [51]	10.7

Methods	FID ↓	Diversity →
DAViD	**1.30**	**8.90**
$DAViD_{w/o\ LoRA}$	2.93	8.59
Real [25]	0.00160	9.48

Table 2. **Quantitative Ablation.** Our smoothness guidance reduces the jitter level of the object motion (left), and our LoRA module better preserves the pre-trained MDM knowledge (right).

case, along with an overlay of objects in the subsequent two frames and a magnified view of the hand-object interaction. Without smoothness guidance, object poses are independently sampled for the human pose, causing motion jitters between adjacent frames and reducing the plausibility of the motion. Additionally, without contact guidance, fine details such as hand contact are lacking, resulting in floating objects. This shows the importance of TGS as it finds a path that continuously connects plausible object poses for each frame, creating a motion that is frame-wise plausible, temporally smooth, and contacted during interaction.

4.4. Quantitative Results

Comparison with Baselines. To verify the efficacy of our model, we compare DAViD against OMOMO [51] and CHOIS [52] on the FullBodyManipulation Dataset [51]. Specifically, we measure the performance of DAViD on each of the test datasets of the FullBodyManipulation [51], selected from baselines. For the metrics about HOI quality, hand contact is considered when the average distance between the 25 hand joints of SMPL-X [70] and the object vertices is smaller than a threshold, 0.05. As shown in Tab. 1, DAViD outperforms the baselines for all contact metrics. This shows that DAViD generates plausible HOI motion, demonstrating the efficacy of our approach in modeling the object pose distribution from the human pose, and sampling object motion using TGS. Additionally, we report FS and foot height of the human motions generated by DAViD, where we generate the same number of samples as in the test dataset. We demonstrate that DAViD generates more natural human motions compared to the baselines, with lower FS and foot height, as shown in Tab. 1.

Ablation Study. We conduct a quantitative ablation study on our key design choices, TGS and LoRA [32]. As shown in Tab. 1, we find that our contact guidance in TGS is crucial for expressing detailed hand contact in HOI motions. Note that this is not obvious, as contact guidance automatically detects potential contact region and guides them to get closer, and not provide information on where the contact should occur. The result shows that the contact guidance in TGS performs fine refinement of object motion within a learned distribution. We also perform an ablation study on our smoothness guidance in TGS by generating object motion using human motion from FullBodyManipulation [51] and measure the jitter. We define jitter following MocapEvery [47] and compute it for object translation. As shown in Tab. 2 (left), our smoothness guidance in TGS significantly reduces the object jitter that occurs when sampling object poses frame-wise, bringing it to a level similar to the jitter found in real-world capture data.

We conduct an ablation study on LoRA to verify its ability to maintain the pre-trained model's knowledge while extending the model's space. We train both models for 1000 epochs on the FullBodyManipulation dataset [51] and compare the FID and diversity on HumanML3D dataset [25] (which MDM is trained on) for text-conditioned motion generation. As shown in Tab. 2 (right), LoRA enables the model to learn new concepts without significantly altering the motion distribution of the pre-trained model. We demonstrate that our design, which adds LoRA to MDM is important to preserve the knowledge of pre-trained model, allowing to combine various HOI concepts with pre-trained human motion.

5. Discussion

In this paper, we present DAViD, a method to learn Dynamic Affordance, addressing both static patterns (*e.g.*, contact, spatial relation) and dynamic motion patterns in HOIs. While obtaining a 4D HOI dataset including diverse objects is challenging, we observe that the video diffusion model has prior knowledge of the dynamic patterns for general objects. Compared to learning affordance from videos generated from text-to-video models, our pipeline, which first renders the 3D object to generate HOI images and then uses an image-to-video model, has the advantage of lifting affordances in video into 3D space. Using the 4D HOI samples generated through our lifting pipeline, we train DAViD, which consists of (1) the LoRA module of MDM for learning human motion patterns and (2) an object pose diffusion model with human pose condition. We qualitatively show that the 4D HOI samples generated by DAViD effectively model the dynamic pattern of human pose and corresponding object pose during HOIs. Quantitative results show that our DAViD models human motion and interactions that are more similar to real-world HOI than the baseline. In addition, we demonstrate the advantage of using LoRA to model the dynamic patterns of using specific objects by generating HOI motion that includes multi-object interactions.

Acknowledgements

This work was supported by NAVER Webtoon, NRF grant funded by the Korean government (MSIT) (No. RS-2022-NR070498, No. RS-2023-00218601, and RS-2025-25396144), and IITP grant funded by the Korea government (MSIT) (No. RS-2024-00439854, RS-2021-II211343, and RS-2025-25442338). H. Joo is the corresponding author.

References

[1] Aitor Aldoma, Federico Tombari, and Markus Vincze. Supervised learning of hidden and non-hidden 0-order affordances and detection in real scenes. In *ICRA*, 2012. 1, 2

[2] Shikhar Bahl, Russell Mendonca, Lili Chen, Unnat Jain, and Deepak Pathak. Affordances from human videos as a versatile representation for robotics. In *CVPR*, 2023. 1, 2

[3] Sangwon Baik, Hyeonwoo Kim, and Hanbyul Joo. Learning 3d object spatial relationships from pre-trained 2d diffusion models. In *ICCV*, 2025. 3

[4] Bharat Lal Bhatnagar, Xianghui Xie, Ilya Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *CVPR*, 2022. 2, 6

[5] 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. 5

[6] Minjie Cai, Kris M. Kitani, and Yoichi Sato. Understanding hand-object manipulation with grasp types and object attributes. In *RSS*, 2016. 2

[7] John Canny. A computational approach to edge detection. In *IEEE TPAMI*, 1986. 3

[8] Hyunsoo Cha, Inhee Lee, and Hanbyul Joo. Perse: Personalized 3d generative avatars from a single portrait. In *CVPR*, 2025. 3

[9] Lucy Chai, Jun-Yan Zhu, Eli Shechtman, Phillip Isola, and Richard Zhang. Ensembling with deep generative views. In *CVPR*, 2021. 3

[10] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. Shapenet: An information-rich 3d model repository. In *arXiv:1512.03012*, 2015. 2, 6

[11] Yu-Wei Chao, Zhan Wang, Rada Mihalcea, and Jia Deng. Mining semantic affordances of visual object categories. In *CVPR*, 2015. 1

[12] Chenyi Chen, Ari Seff, Alain Kornhauser, and Jianxiong Xiao. Deepdriving: Learning affordance for direct perception in autonomous driving. In *ICCV*, 2015. 2

[13] Fu-Jen Chu, Ruinian Xu, and Patricio A. Vela. Learning affordance segmentation for real-world robotic manipulation via synthetic images. In *IEEE Robot. Autom. Lett.*, 2019. 2

[14] Ching-Yao Chuang, Jiaman Li, Antonio Torralba, and Sanja Fidler. Learning to act properly: Predicting and explaining affordances from images. In *CVPR*, 2018. 1

[15] Shengheng Deng, Xun Xu, Chaozheng Wu, Ke Chen, and Kui Jia. 3d affordancenet: A benchmark for visual object affordance understanding. In *CVPR*, 2021. 2

[16] Christian Diller and Angela Dai. Cg-hoi: Contact-guided 3d human-object interaction generation. In *CVPR*, 2024. 7

[17] Thanh-Toan Do, Anh Nguyen, and Ian Reid. Affordancenet: An end-to-end deep learning approach for object affordance detection. In *ICRA*, 2018. 1, 2

[18] J.R. Dormand and P.J. Prince. A family of embedded runge-kutta formulae. In *J. Comput. Appl. Math.*, 1980. 6

[19] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. Arctic: A dataset for dexterous bimanual hand-object manipulation. In *CVPR*, 2023. 2

[20] 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. 4

[21] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *ICLR*, 2023. 3

[22] James J. Gibson. The ecological approach to visual perception. In *Houghton Mifflin*, 1979. 1, 2

[23] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 3

[24] Yuchao Gu, Xintao Wang, Jay Zhangjie Wu, Yujun Shi, Chen Yunpeng, Zihan Fan, Wuyou Xiao, Rui Zhao, Shuning Chang, Weijia Wu, Yixiao Ge, Shan Ying, and Mike Zheng Shou. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. In *NeurIPS*, 2023. 3

[25] 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. 5, 8

[26] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In *ICLR*, 2024. 3

[27] Sookwan Han and Hanbyul Joo. Learning canonicalized 3d human-object spatial relations from unbounded synthesized images. In *ICCV*, 2023. 1, 2, 3

[28] Zekun Hao, Arun Mallya, Serge Belongie, and Ming-Yu Liu. Gancraft: Unsupervised 3d neural rendering of minecraft worlds. In *ICCV*, 2021. 3

[29] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *ICCV*, 2019. 2

[30] Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and Xiaojuan Qi. Is synthetic data from generative models ready for image recognition? In *arXiv:2210.07574*, 2022. 3

[31] Tucker Hermans, James M Rehg, and Aaron Bobick. Affordance prediction via learned object attributes. In *ICRA*, 2011. 2

[32] 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. 2, 3, 5, 7, 8

[33] Yinghao Huang, Omid Taheri, Michael J. Black, and Dimitrios Tzionas. Intercap: Joint markerless 3d tracking of humans and objects in interaction from multi-view rgb-d images. In *IJCV*, 2024. 2, 6

[34] Nan Jiang, Tengyu Liu, Zhexuan Cao, Jieming Cui, Zhiyuan Zhang, Yixin Chen, He Wang, Yixin Zhu, and Siyuan Huang. Full-body articulated human-object interaction. In *ICCV*, 2023. 2

[35] 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. 4

[36] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together. In *ECCV*, 2024. 4

[37] David Inkyu Kim and Gaurav S. Sukhatme. Semantic labeling of 3d point clouds with object affordance for robot manipulation. In *ICRA*, 2014. 1, 2

[38] 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, 2, 3

[39] Jeonghwan Kim, Jisoo Kim, Jeonghyeon Na, and Hanbyul Joo. Parahome: Parameterizing everyday home activities towards 3d generative modeling of human-object interactions. In *arXiv:2401.10232*, 2024. 2

[40] Taeksoo Kim, Shunsuke Saito, and Hanbyul Joo. Ncho: Unsupervised learning for neural 3d composition of humans and objects. In *ICCV*, 2023. 1

[41] Hedvig Kjellström, Javier Romero, and Danica Kragić. Visual object-action recognition: Inferring object affordances from human demonstration. In *CVIU*, 2011. 1, 2

[42] Kling AI. https://klingai.com/, (accessed Jul 20th, 2025). 3

[43] Sumith Kulal, Tim Brooks, Alex Aiken, Jiajun Wu, Jimei Yang, Jingwan Lu, Alexei A. Efros, and Krishna Kumar Singh. Putting people in their place: Affordance-aware human insertion into scenes. In *CVPR*, 2023. 3

[44] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *CVPR*, 2023. 3

[45] Black Forest Labs. Flux. https://bfl.ai/, (accessed Jul 20th, 2025). 3

[46] Jiye Lee and Hanbyul Joo. Locomotion-action-manipulation: Synthesizing human-scene interactions in complex 3d environments. In *ICCV*, 2023. 2

[47] Jiye Lee and Hanbyul Joo. Mocap everyone everywhere: Lightweight motion capture with smartwatches and a head-mounted camera. In *CVPR*, 2024. 8

[48] Yong Jae Lee and Kristen Grauman. Predicting important objects for egocentric video summarization. In *IJCV*, 2015. 2

[49] Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. Epnp: An accurate o(n) solution to the pnp problem. In *IJCV*, 2009. 4

[50] Gen Li, Varun Jampani, Deqing Sun, and Laura Sevilla-Lara. Locate: Localize and transfer object parts for weakly supervised affordance grounding. In *CVPR*, 2023. 1, 2

[51] Jiaman Li, Jiajun Wu, and C Karen Liu. Object motion guided human motion synthesis. In *ACM TOG*, 2023. 2, 6, 7, 8

[52] Jiaman Li, Alexander Clegg, Roozbeh Mottaghi, Jiajun Wu, Xavier Puig, and C. Karen Liu. Controllable human-object interaction synthesis. In *ECCV*, 2024. 2, 3, 7, 8

[53] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. In *CVPR*, 2024. 3

[54] Wei Liang, Yibiao Zhao, Yixin Zhu, and Song-Chun Zhu. What is where: Inferring containment relations from videos. In *IJCAI*, 2016. 1

[55] Zhiheng Liu, Ruili Feng, Kai Zhu, Yifei Zhang, Kecheng Zheng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Cones: Concept neurons in diffusion models for customized generation. In *arXiv:2303.05125*, 2023. 3

[56] Timo Lüddecke and Florentin Wörgötter. Learning to segment affordances. In *ICCVW*, 2017. 1, 2

[57] Chengzhi Mao, Augustine Cha, Amogh Gupta, Hao Wang, Junfeng Yang, and Carl Vondrick. Generative interventions for causal learning. In *CVPR*, 2021. 3

[58] Tanis Mar, Vadim Tikhanoff, Giorgio Metta, and Lorenzo Natale. Self-supervised learning of tool affordances from 3d tool representation through parallel som mapping. In *ICRA*, 2017. 1, 2

[59] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Finding an unsupervised image segmenter in each of your deep generative models. In *arXiv:2105.08127*, 2021. 3

[60] Marko Mihajlovic, Shunsuke Saito, Aayush Bansal, Michael Zollhoefer, and Siyu Tang. Coap: Compositional articulated occupancy of people. In *CVPR*, 2022. 5

[61] Kaichun Mo, Shilin Zhu, Angel X. Chang, Li Yi, Subarna Tripathi, Leonidas J. Guibas, and Hao Su. Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding. In *CVPR*, 2019. 2, 6

[62] Bogdan Moldovan and Luc De Raedt. Occluded object search by relational affordances. In *ICRA*, 2014. 1, 2

[63] Roozbeh Mottaghi, Connor Schenck, Dieter Fox, and Ali Farhadi. See the glass half full: Reasoning about liquid containers, their volume and content. In *ICCV*, 2017.

[64] Austin Myers, Ching L. Teo, Cornelia Fermüller, and Yiannis Aloimonos. Affordance detection of tool parts from geometric features. In *ICRA*, 2015. 2

[65] Anh Nguyen, Dimitrios Kanoulas, Darwin G. Caldwell, and Nikos G. Tsagarakis. Detecting object affordances with convolutional neural networks. In *IROS*, 2016.

[66] Anh Nguyen, Dimitrios Kanoulas, Darwin G. Caldwell, and Nikos G. Tsagarakis. Object-based affordances detection with convolutional neural networks and dense conditional random fields. In *IROS*, 2017. 1, 2

[67] Hyun-Jic Oh and Won-Ki Jeong. Diffmix: Diffusion model-based data synthesis for nuclei segmentation and classification in imbalanced pathology image datasets. In *arXiv:2306.14132*, 2023. 3

[68] OpenAI. Gpt-4o system card. https://openai.com/research/gpt-4v-system-card, (accessed Jul 20th, 2025). 3

[69] Xingang Pan, Bo Dai, Ziwei Liu, Chen Change Loy, and Ping Luo. Do 2d gans know 3d shape? unsupervised 3d shape reconstruction from 2d image gans. In *arXiv:2011.00844*, 2020. 3

[70] 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. 4, 5, 8

[71] Xiaogang Peng, Yiming Xie, Zizhao Wu, Varun Jampani, Deqing Sun, and Huaizu Jiang. Hoi-diff: Text-driven synthesis of 3d human-object interactions using diffusion models. In *arXiv:2312.06553*, 2023. 2, 3, 7

[72] Cody Phillips, Matthieu Lecce, and Kostas Daniilidis. Seeing glassware: from edge detection to pose estimation and shape recovery. In *RSS*, 2016. 1

[73] Alessandro Pieropan, Carl Henrik Ek, and Hedvig Kjellström. Functional object descriptors for human activity modeling. In *ICRA*, 2013. 1, 2

[74] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *PMLR*, 2021. 5

[75] Scott D. Roth. Ray casting for modeling solids. In *Comput. Graph. Image Process.*, 1982. 4

[76] Anirban Roy and Sinisa Todorovic. A multi-scale cnn for affordance segmentation in rgb images. In *ECCV*, 2016. 1, 2

[77] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, 2023. 3

[78] Simo Ryu. Low-rank adaptation for fast text-to-image diffusion fine-tuning. https://github.com/cloneofsimo/lora, (accessed Jul 20th, 2025). 3

[79] Johann Sawatzky and Juergen Gall. Adaptive binarization for weakly supervised affordance segmentation. In *ICCVW*, 2017. 1, 2

[80] Markus Schoeler and Florentin Wörgötter. Bootstrapping the semantics of tools: Affordance analysis of real world objects on a per-part basis. In *IEEE Trans. Cog. Develop. Sys.*, 2016. 1, 2

[81] Viraj Shah, Nataniel Ruiz, Forrester Cole, Erika Lu, Svetlana Lazebnik, Yuanzhen Li, and Varun Jampani. Ziplora: Any subject in any style by effectively merging loras. In *ECCV*, 2024. 3

[82] 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. 3, 4

[83] Yujun Shi, Chuhui Xue, Jiachun Pan, Wenqing Zhang, Vincent YF Tan, and Song Bai. Dragdiffusion: Harnessing diffusion models for interactive point-based image editing. In *CVPR*, 2024. 3

[84] SketchFab. https://sketchfab.com/, (accessed Jul 20th, 2025). 2, 6

[85] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021. 6

[86] Jie Sun, Joshua Moore, Aaron Bobick, and James Rehg. Learning visual object categories for robot affordance prediction. In *IJRR*, 2010. 1, 2

[87] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *ECCV*, 2020. 2

[88] Fabio Henrique Kiyoiti dos Santos Tanaka and Claus Aranha. Data augmentation using gans. In *arXiv:1904.09135*, 2019. 3

[89] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *ICLR*, 2023. 2, 3, 5, 7, 8

[90] Brandon Trabucco, Kyle Doherty, Max Gurinas, and Ruslan Salakhutdinov. Effective data augmentation with diffusion models. In *arXiv:2302.07944*, 2023. 3

[91] Shashank Tripathi, Agniv Chatterjee, Jean-Claude Passy, Hongwei Yi, Dimitrios Tzionas, and Michael J. Black. Deco: Dense estimation of 3d human-scene contact in the wild. In *ICCV*, 2023. 1, 2

[92] Nontawat Tritrong, Pitchaporn Rewatbowornwong, and Supasorn Suwajanakorn. Repurposing gans for one-shot semantic part segmentation. In *CVPR*, 2021. 3

[93] Emre Ugur, Sandor Szedmak, and Justus Piater. Bootstrapping paired-object affordance learning with learned single-affordance features. In *ICDL*, 2014. 1, 2

[94] Karthik Mahesh Varadarajan and Markus Vincze. Parallel deep learning with suggestive activation for object category recognition. In *ICVS*, 2013. 2

[95] Pascal Vincent. A connection between score matching and denoising autoencoders. In *Neural Comput.*, 2011. 6

[96] Andrey Voynov, Qinghao Chu, Daniel Cohen-Or, and Kfir Aberman. P+: Extended textual conditioning in text-to-image generation. In *arXiv:2303.09522*, 2023. 3

[97] Chenglong Wang, Jiangyan Yi, Xiaohui Zhang, Jianhua Tao, Le Xu, and Ruibo Fu. Low-rank adaptation method for wav2vec2-based fake audio detection. In *arXiv:2306.05617*, 2023. 3

[98] Hanqing Wang, Wei Liang, and Lap-Fai Yu. Transferring objects: Joint inference of container and human pose. In *ICCV*, 2017. 1

[99] Yinhuai Wang, Jing Lin, Ailing Zeng, Zhengyi Luo, Jian Zhang, and Lei Zhang. Physhoi: Physics-based imitation of dynamic human-object interaction. In *arXiv:2312.04393*, 2023. 2

[100] Yufu Wang, Ziyun Wang, Lingjie Liu, and Kostas Daniilidis. Tram: Global trajectory and motion of 3d humans from in-the-wild videos. In *arXiv:2403.17346*, 2024. 4

[101] Zan Wang, Yixin Chen, Baoxiong Jia, Puhao Li, Jinlu Zhang, Jingze Zhang, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Move as you say, interact as you can: Language-guided human motion generation with scene affordance. In *CVPR*, 2024. 2, 3

[102] Jianzong Wu, Xiangtai Li, Yanhong Zeng, Jiangning Zhang, Qianyu Zhou, Yining Li, Yunhai Tong, and Kai Chen. Motionbooth: Motion-aware customized text-to-video generation. In *NeurIPS*, 2024. 3

[103] Qianyang Wu, Ye Shi, Xiaoshui Huang, Jingyi Yu, Lan Xu, and Jingya Wang. Thor: Text to human-object interaction diffusion via relation intervention. In *arXiv:2403.11208*, 2024. 7

[104] Weijia Wu, Yuzhong Zhao, Hao Chen, Yuchao Gu, Rui Zhao, Yefei He, Hong Zhou, Mike Zheng Shou, and Chunhua Shen. Datasetdm: Synthesizing data with perception annotations using diffusion models. In *NeurIPS*, 2023. 3

[105] Zhen Wu, Jiaman Li, and C. Karen Liu. Human-object interaction from human-level instructions. In *arXiv:2406.17840*, 2024. 2, 3

[106] Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, Li Yi, Angel X. Chang, Leonidas J. Guibas, and Hao Su. Sapien: A simulated part-based interactive environment. In *CVPR*, 2020. 2, 6

[107] Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. Interdiff: Generating 3d human-object interactions with physics-informed diffusion. In *ICCV*, 2023. 2, 3, 7

[108] Sirui Xu, Ziyin Wang, Yu-Xiong Wang, and Liang-Yan Gui. Interdreamer: Zero-shot text to 3d dynamic human-object interaction. In *NeurIPS*, 2024. 2, 3, 7

[109] Sirui Xu, Hung Yu Ling, Yu-Xiong Wang, and Liang-Yan Gui. Intermimic: Towards universal whole-body control for physics-based human-object interactions. In *CVPR*, 2025. 2

[110] Yuhang Yang, Wei Zhai, Hongchen Luo, Yang Cao, Jiebo Luo, and Zheng-Jun Zha. Grounding 3d object affordance from 2d interactions in images. In *ICCV*, 2023. 1, 2

[111] Yuhang Yang, Wei Zhai, Hongchen Luo, Yang Cao, and Zheng-Jun Zha. Lemon: Learning 3d human-object interaction relation from 2d images. In *CVPR*, 2024. 1

[112] Yufei Ye, Xueting Li, Abhinav Gupta, Shalini De Mello, Stan Birchfield, Jiaming Song, Shubham Tulsiani, and Sifei Liu. Affordance diffusion: Synthesizing hand-object interactions. In *CVPR*, 2023. 3

[113] Hongwei Yi, Justus Thies, Michael J. Black, Xue Bin Peng, and Davis Rempe. Generating human interaction motions in scenes with text control. In *ECCV*, 2024. 2, 3, 7

[114] Lap-Fai Yu, Noah Duncan, and Sai-Kit Yeung. Fill and transfer: A simple physics-based approach for containability reasoning. In *ICCV*, 2015. 1

[115] Jiyao Zhang, Mingdong Wu, and Hao Dong. Generative category-level object pose estimation via diffusion models. In *NeurIPS*, 2023. 6

[116] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 3

[117] Xiaohan Zhang, Bharat Lal Bhatnagar, Sebastian Starke, Vladimir Guzov, and Gerard Pons-Moll. Couch: Towards controllable human-chair interactions. In *ECCV*, 2022. 2

[118] Yuxuan Zhang, Wenzheng Chen, Huan Ling, Jun Gao, Yinan Zhang, Antonio Torralba, and Sanja Fidler. Image gans meet differentiable rendering for inverse graphics and interpretable 3d neural rendering. In *ICLR*, 2021. 3

[119] Yuxuan Zhang, Huan Ling, Jun Gao, Kangxue Yin, Jean-Francois Lafleche, Adela Barriuso, Antonio Torralba, and Sanja Fidler. Datasetgan: Efficient labeled data factory with minimal human effort. In *CVPR*, 2021. 3

[120] Kaifeng Zhao, Shaofei Wang, Yan Zhang, Thabo Beeler, and Siyu Tang. Compositional human-scene interaction synthesis with semantic control. In *ECCV*, 2022. 2

[121] Chenyang Zhu, Kai Li, Yue Ma, Chunming He, and Li Xiu. Multibooth: Towards generating all your concepts in an image from text. In *arXiv:2404.14239*, 2024. 3

[122] Yuke Zhu, Alireza Fathi, and Li Fei-Fei. Reasoning about object affordances in a knowledge base representation. In *ECCV*, 2014. 1

[123] Yuke Zhu, Ce Zhang, Christopher Ré, and Li Fei-Fei. Building a large-scale multimodal knowledge base system for answering visual queries. In *arXiv:1507.05670*, 2015.

[124] Yixin Zhu, Chenfanfu Jiang, Yibiao Zhao, Demetri Terzopoulos, and Song-Chun Zhu. Inferring forces and learning human utilities from videos. In *CVPR*, 2016. 1