# FICTION: 4D Future Interaction Prediction from Video

Kumar Ashutosh, Georgios Pavlakos, Kristen Grauman
University of Texas at Austin

## Abstract

*Anticipating how a person will interact with objects in an environment is essential for activity understanding, but existing methods are limited to the 2D space of video frames—capturing physically ungrounded predictions of "what" and ignoring the "where" and "how". We introduce* FICTION *for 4D future interaction prediction from videos. Given an input video of a human activity, the goal is to predict which objects at what 3D locations the person will interact with in the next time period (e.g., cabinet, fridge), and how they will execute that interaction (e.g., poses for bending, reaching, pulling). Our novel model* FICTION *fuses the past video observation of the person's actions and their environment to predict both the "where" and "how" of future interactions. Through comprehensive experiments on a variety of activities and real-world environments in Ego-Exo4D, we show that our proposed approach outperforms prior autoregressive and (lifted) 2D video models substantially, with more than 30% relative gains.*

## 1. Introduction

Humans constantly move around their environment and interact with objects to accomplish various daily tasks. A person making a salad will first get a bowl from the cabinet, get lettuce from the fridge, and then dressings from the shelf. An assistive AI agent, by observing her intent, could help her in various ways—fetching a better dressing for her, or helping her get the bowl if she has trouble bending to the lower cabinet. Such an AI assistant needs to understand (a) the location of the person and various objects, (b) whether those objects will be interacted with in the future, and (c) how will the person interact with each object. Besides assistive robotics [17, 99], predicting future interactions could enable path planning and navigation [3, 4, 9, 88], imitation learning [15, 43, 87], AI coaching and AR assistants [6, 94], and view planning [44, 47].

Current methods attempt to solve object interaction anticipation as a 2D video problem—inferring 2D heatmaps on frames, naming the next action, or detecting the likely next active object [5, 8, 30, 33, 61–63, 78, 80]. However, this common 2D formulation overlooks the persistent 3D context of the environment and object layout, instead viewing the world as detached glimpses from the camera's point of view. Other work anticipates future 3D human poses while taking context from the scene [12, 41, 100], typically using autoregressive models [28, 35, 111]. However, these methods are not interaction-centric, i.e., they focus on motion generation without attempting to model object interactions. Both lines of work typically predict only a few seconds into the future, within the same action (e.g., walking, chopping).

Our key insight is that a person's movement and interaction are tightly linked to the activity they are doing and the objects in their environment. Therefore, it is crucial to address this task with 3D knowledge about the persistent surrounding environment, its objects, and the person's body poses. The activity and the objects in the surroundings impact both *where* will be the future interactions and *how* the person will interact in the environment. For example, if a person is making tea, it is likely that the person will interact with some water source to get water. Moreover, if the water dispenser is wall-mounted, the person will extend their arm to reach it. See Fig. 1.

To overcome the gap in today's models, we introduce a 4D formulation of the interaction anticipation problem, and we propose FICTION, a novel model to address it. As input, our model takes a video observation of a person doing an activity, together with a 3D scene representation containing the objects and the person.[1] The proposed model encodes the past observations into a multimodal representation with a transformer, then learns to decode back to all future 3D interaction locations and, for each location, a distribution of likely body poses that will be executed there. We hypothesize, and experimentally validate, that video and an explicit 3D scene context helps predict the interaction location (*where*) and the human body pose at the time of interaction (*how*). Furthermore, considering interactions in 4D facilitates longer-horizon anticipation—imagining how the

Project Page: https://vision.cs.utexas.edu/projects/FIction/

[1] The 3D scene representation can either be derived directly from the video using state-of-the-art methods [22, 59, 75, 96], or estimated from richer sensors when available [25].

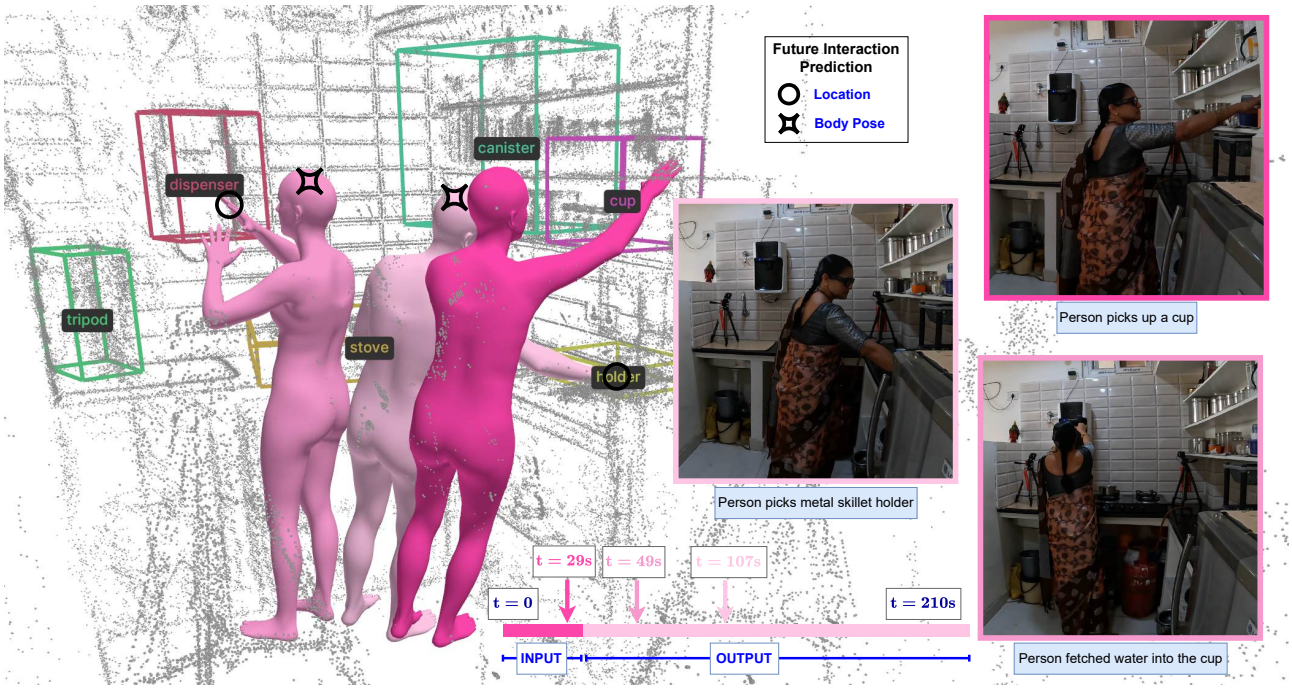


Figure 1. **Future interaction prediction.** When doing a procedure like *making milk tea*, a person moves around in their environment, interacting with different objects like *water dispenser*, *stove*, *skillet holder*, and *cups*. Each interaction has an associated body pose, e.g., using two hands when fetching water, extending the body to reach upper wall cabinets. Given an environment and the procedure till a time $t$, we predict *all* future object interactions (specifically, pooled over the next 3 mins) and the likely body poses during those object interactions. Best viewed in zoom. Only representative object bounding boxes shown for clarity.

activity in a given physical space will unfold over the course of minutes, not just seconds, as a person moves about to different regions of the 3D environment.

We leverage the recently introduced Ego-Exo4D [38] dataset to create the training and the testing dataset for future interaction location prediction (*where*) and future interaction body pose prediction (*how*), thus capturing both essential aspects. In extensive experiments, we show that our method outperforms various related state-of-the-art methods in interaction anticipation—autoregressive action/pose/hotspot prediction [5, 35, 63, 111], and video-based methods without explicit 3D environment conditioning [55, 114]. Specifically, our method is more than 30% better than the best performing baseline. Overall, this work offers an important step towards realizing the synergy between 3D environments and human action, which are often treated independently in the literature to date.

## 2. Related Work

**Learning about activity in videos.** Video datasets like HowTo100M [70], Kinetics [13], Ego4D [37] and Ego-Exo4D [38] contain videos of people doing a wide variety of activities and support various computer vision tasks including activity recognition [27, 34, 53, 54, 102], procedural planning [11, 14, 116], task graph learning [7, 23, 58, 117], action anticipation [1, 29–33, 51, 69, 81, 98], goal forecasting [90], and representation learning [5, 56, 71, 103]. Despite many promising results, there is a heavy emphasis on 2D video devoid of the underlying persistent 3D context. At best, models are expected to implicitly learn the spatial arrangement of objects in 3D. This fundamentally limits the spatial understanding of these models, particularly for long-horizon tasks like future interaction prediction. In this work, we propose to use explicit 3D environment context as an essential input alongside the video frames. We show this *early-fusion* enables the model to understand the spatial nature of the activity, resulting in a superior performance over video-based models without 3D context or whose 3D context is infused post-hoc.

**Predicting object interactions.** Humans interact with objects in various ways, e.g., lifting, dragging, sitting, sleeping upon. Human object interaction [19, 20, 74, 101, 107] focuses on understanding affordances and contact points for a given object. Likewise, hand-object interaction [36, 46, 73, 82, 109] predicts the hand pose when performing actions like grasping or holding for a given object. However, all the prior work assumes a known object (out of environment context) that is the target interaction. Moreover, the interaction is instantaneous, typically spanning a few seconds. In contrast, our setting entails 3D scenes with

multiple objects, and we predict both things—the location of the objects that will be interacted with, and the corresponding body poses over the course of multiple minutes.

Hotspot anticipation [8, 61, 62, 78, 80] is a related problem of localizing interaction points as heatmaps in 2D video frames. However, this prior work has no persistent 3D spatial understanding. In contrast, we leverage 3D spatial information, extractable from the same egocentric video, to predict interactions registered in the underlying 3D space containing multiple objects over a long duration. Unlike prior work that infers the affordance map of a static 3D environment [77, 79], our work anticipates the actions a person will perform next given a partial video of their activity.

**Human pose from videos.** Extracting human body pose from images or video is an active area of research [24, 35, 64, 83, 92, 108]. Typical output formats are SMPL [64] or body skeletons [57]. Some work attempts to predict human poses in the future—using motion priors [2, 68, 89, 110, 112, 113] and 3D scene context [12, 16, 65, 67, 100, 105, 115] as additional information. These methods can generate plausible motion patterns for short-term actions—especially periodic ones like walking, running, or jumping—but they do not tackle anticipation of longer-term behaviors, namely actions extending beyond the current one that entail interacting with different future objects in different future steps of the activity.

Recent work [16, 41, 42, 100, 115] shows how humans would interact with an object in the scene, e.g., chair, sofa, staircase. The datasets used for this task are mostly simulation-based [12, 85] or small-scale [40, 41]. Most importantly, these methods generate pose sequences for a fixed (singleton) object, whereas ours predicts future poses conditioned on past activity in a dynamic scene and in the presence of multiple potential interacting objects. Our idea can be seen as "interaction-centric pose anticipation", a new and important dimension of this problem space.

# 3. Approach

We first formally describe the problem (Sec. 3.1). Next, we describe the model design for our task (Sec. 3.2), the dataset preparation strategy (Sec. 3.3), and implementation details (Sec. 3.4).

## 3.1. Future interaction prediction

Given an observation of a person performing an activity, up to a time instant, we want to anticipate *all* subsequent interactions, up to a future time (3 minutes in our experiments). We capture both aspects of an interaction—its location (*where*), and its body poses (*how*). We define interaction as making contact with an object of interest to use it, e.g., *lifting a cup*, *opening a fridge*, as opposed to non-contact actions, e.g., *monitoring the oven*, *observing the bike tire*.

**Future interaction location prediction:** Firstly, given an observation video $\mathcal{V}$ till time $\tau_o$, and the 3D locations, encoded as $\mathcal{P}$ (obtained from $\mathcal{V}$ or otherwise), we aim to learn a future object interaction function $\mathcal{F}_o$ that predicts all future object interactions, till time $\tau_f$. Following [33], we allow for a small anticipation time $\tau_a$ buffer immediately following $\tau_o$. The output should be a set of points $\mathbf{x} \in \mathbb{R}^3$ such that the person interacts with the object at the 3D point $\mathbf{x}_{\tau_k}$ at the timestamp $\tau_k$, for all such interaction timepoints before $\tau_f$:

$$\mathcal{F}_o\left(\mathcal{V}_{0:\tau_o}, \mathcal{P}\right) = \left\{\mathbf{x}_{\tau_k} \mid \mathcal{I}(\mathbf{x}_{\tau_k}) \in \mathcal{O}\right\}, \tag{1}$$

where $\tau_f > \tau_k > \tau_o + \tau_a$, and the interaction function $\mathcal{I}(\mathbf{x}_{\tau_k}) = o \in \mathcal{O}$ if the person is interacting with an object $o$ at location $\mathbf{x}$ at time $\tau_k$; $\phi$ otherwise. Here, $\mathcal{O}$ is the set of all objects.

**Future interaction pose prediction:** Next, at each interaction point, we want to predict the body pose. However, there are various correct possible future body poses for a given interaction point and an object. For example, a person can reach out to the faucet to *turn it on*, or to *clean it*, both having different body poses. Thus, we learn a function to return the distribution of the likely body poses, rather than a deterministic body pose:

$$\mathcal{F}_p\left(\mathcal{V}_{0:\tau_o}, \mathcal{P}, \mathbf{x}_{\tau_k}\right) = \mathbb{P}\left(\theta, t\right), \tag{2}$$

where we learn a distribution over the SMPL [64] body pose parameters $\theta$ and the location $t$, for any given query location $\mathbf{x}_{\tau_k}$. We omit the SMPL shape parameter $\beta$, since it does not change during interactions. Likely body poses are obtained by sampling from $\mathcal{F}_p$.

## 3.2. Learning future interactions

Next we introduce the training framework for learning the future object interaction $\mathcal{F}_o$ and future pose distribution $\mathcal{F}_p$ functions. The common input to both these functions is the video observation till time $\tau_o$. We use the following components for observation input: the egocentric video of the actor $V$, the body pose of the actor $P$, and the object and the actor placement in the 3D scene. We obtain a common modality-agnostic representation for all the inputs and feed it into a multimodal transformer. The transformer encoder output is then used to predict the future interaction locations. For future pose distribution prediction, we additionally provide a location as a query to a CVAE [93] based encoder-decoder. The architecture is shown in Fig. 2 and detailed below.

**Representing the 3D scene as voxels.** We discretize the 3D space as a $N \times N \times N$ voxel grid. The boundaries are chosen as the maximum of the actor movement and the object placements. We choose a discretized, explicit voxel representation for ease of training and interpretability, as opposed to alternatives based on implicit NeRF [72]-like functions.

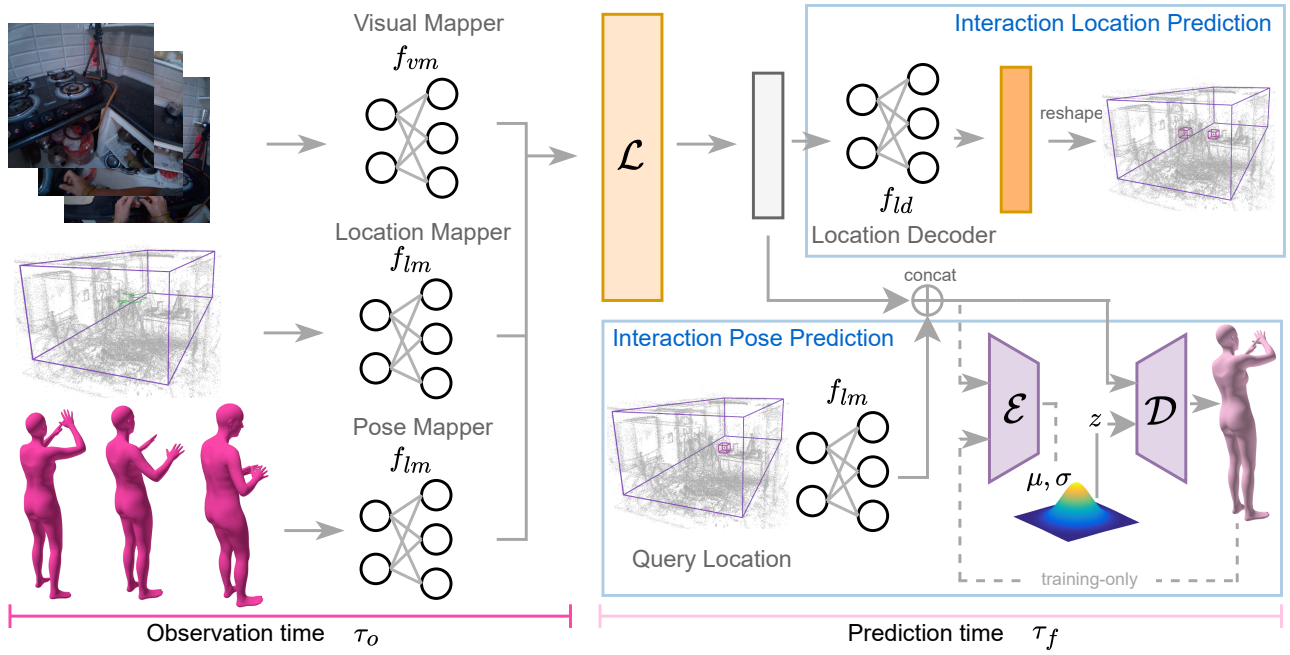


Figure 2. **Overview of the FICTION approach.** The past observation information (video, pose, and environment) is encoded into a multimodal representation (left). The multimodal encoder $\mathcal{L}$ encodes the past observation, and is used to predict the interaction locations using a decoder (top right). We use the past observation encoding, along with the query location, to train a CVAE encoder decoder to generate a location-specific pose distribution conditioned on the past activity (bottom right).

**Encoding egocentric actor observation.** Given an egocentric observation till time $\tau_o$, we use an off-the-shelf feature extractor $f_V$ to obtain video representation $f_V(V)$. The visual feature extractor is a large pretrained encoder network that offers good semantics about the activity. We use a visual mapper $f_{vm}$ to output a modality-agnostic representation $\bar{\mathbf{v}} = f_{vm}(f_V(V))$ that can be fed to a transformer. The visual mapper [52, 60] is used to convert visual features to a representation that can be directly used with a transformer model.

**Encoding body pose observation.** Next, we encode the observed SMPL [64] body pose parameter $\theta$ of the pose $P$. These are real-valued numbers, and we use a pose mapper $f_{pm}$ to convert the pose parameters into a representation $\bar{\mathbf{p}} = f_{pm}(\theta)$. The location $t$ is encoded along with the object bounding boxes, explained next, for uniformity.

**Encoding object bounding boxes.** We represent the 3D scene $S$ as voxels, as described above. We assign an index $i$ at a voxel location containing object $o$ if $o = \mathcal{O}[i]$, where $\mathcal{O}$ is the object taxonomy [39]. We use a reserved index to denote the actor location $t$ in this scene $S$. Next, we use a location mapper $f_{lm}$ to encode the object representations and actor locations as $\bar{\mathbf{o}} = f_{lm}(S)$, same as above.

**Multimodal transformer encoder.** Next, we take the multimodal representations $\bar{\mathbf{v}}, \bar{\mathbf{p}}$ and $\bar{\mathbf{o}}$ and use a transformer with self-attention to learn an output representation, $\bar{\mathbf{r}} = \mathcal{L}(\bar{\mathbf{v}} \mid \bar{\mathbf{p}} \mid \bar{\mathbf{o}})$. We choose the first output of the transformer as the output representation, following standard practice [10, 106]. The output representation $\bar{\mathbf{r}}$ encapsulates the observation information till time $\tau_o$.

**Decoding future interaction location.** The final stage of learning $\mathcal{F}_o$ involves finding *all* the future interaction locations. We use the same $N \times N \times N$ voxel representation at the output. A voxel is marked 1 when the corresponding location has a future interaction; 0 otherwise. (We detail how to extract these voxel maps from unlabeled video in Sec. 3.3.) To achieve this, we use a simple location decoder linear layer $f_{ld}$ that maps $\bar{\mathbf{r}}$ to a vector having $N^3$ dimensions. The output is then reshaped to $N \times N \times N$ grid to obtain the predicted future interaction location grid $\hat{L}$.

**CVAE for future pose distribution.** In addition to the output representation $\bar{\mathbf{r}}$, the pose distribution function $\mathcal{F}_p$ requires an additional input for any query location $x \in \mathbb{R}^3$ that denotes the location at which the pose distribution is desired. We use the same voxel representation, and the location mapper $f_{lm}$ to encode the reference location as $\bar{\mathbf{x}}$.

Following the standard CVAE [93] architecture, the CVAE encoder $\mathcal{E}$ takes as input the multimodal output representation $\bar{\mathbf{r}}$, the location embedding $\bar{\mathbf{x}}$, and the desired pose output $P \sim \mathbb{P}(\theta, t)$ and outputs the latent distribution parameters $\mu, \sigma$, i.e. $\mu, \sigma = \mathcal{E}(\bar{\mathbf{r}}, \bar{\mathbf{x}}, P)$. Here, $P$ is a sample from the pose distribution. Next, the decoder $\mathcal{D}$ tries to

reconstruct the sampled pose $P$ using the inputs $\bar{\mathbf{r}}, \bar{\mathbf{x}}$ and a sample $z$ such that $z \sim \mathcal{N}(\mu, \sigma)$, i.e. $\hat{P} = \mathcal{D}(z, \bar{\mathbf{r}}, \bar{\mathbf{x}})$. At inference, we can sample multiple $z \sim \mathcal{N}(0, 1)$ and use that to predict output pose samples, $\hat{P} = \mathcal{D}(z, \bar{\mathbf{r}}, \bar{\mathbf{x}})$. We use $\beta$ from the past observation, alongside the predicted $\theta, t$, to reconstruct the SMPL body mesh/3D joints.

**Training objectives.** We now describe the training objective to learn the functions $\mathcal{F}_o$ and $\mathcal{F}_p$. As described in Sec. 3.2, we represent the future location as a binary voxel: 1 if it contains a future interaction, and 0 otherwise. Consequently, we learn $\hat{L}$ as the output of the function $\mathcal{F}_o$. Similarly, we represent the ground truth output locations $\{\mathbf{x}_{\tau_k}\}$ (Eq. 1) as a voxel grid $L$. Thus, we use the standard binary cross entropy (BCE) loss for the future interaction location prediction task.

Next, for the future pose distribution prediction task, we use a combination of reconstruction loss and the KL divergence between the predicted distribution parameter and the standard normal. For the reconstruction loss, we use the MSE loss between the predicted SMPL parameters $\hat{P} = (\hat{\theta}, \beta, \hat{t})$ and the ground truth parameters $P = (\theta, \beta, t)$. Moreover, following [35], we also convert the SMPL parameters to 3D body joints, $J = \text{SMPL}(P)$, and compute the joint $L_1$ error. Finally, the KL divergence error is computed between the predicted parameters $(\mu, \sigma)$ and $(0, 1)$. Overall, the training objective is to minimize

$$w_S \|P - \hat{P}\|_2 + w_J \|J - \hat{J}\|_1 + KL\left(\mathcal{N}(\mu, \sigma), \mathcal{N}(0, 1)\right),$$

where the weights $w_S$ and $w_J$ control the loss contributions.

### 3.3. Future interaction dataset

As introduced above, we use the egocentric video $V$ and body pose $P$ as the modalities in this task, and we also use the object interaction locations in the 3D scene. While no existing dataset provides exactly this prepared data, Ego-Exo4D [38] offers the necessary raw data and serves as an excellent large-scale, diverse testbed for our work (see Sec. 3.4 for scope and statistics). Ego-Exo4D data is recorded with Aria glasses [25] having dedicated SLAM cameras, thus enabling superior 3D registration, an advantage over using RGB frames for 3D information extraction [86, 91], as done in [76, 97]. We find the object placements and human poses as follows.

**3D object bounding boxes.** Object segmentation in 3D [50, 95] and 2D [49, 118] is an active area of research. Ego-Exo4D's SLAM cameras provide a mapping between the video pixels to the 3D locations. Therefore, we perform object segmentation in 2D video frames using Detic [118] (that uses the object taxonomy from LVIS [39]) and use the above mapping to convert segmented pixels to object point clouds. Next, we use DBSCAN [26] for density-based clustering to find the count of an object and make the bounding boxes tight. Lastly, we find an oriented bounding box [48] for every object in the scene. See Supp. for more implementation details and visualizations.

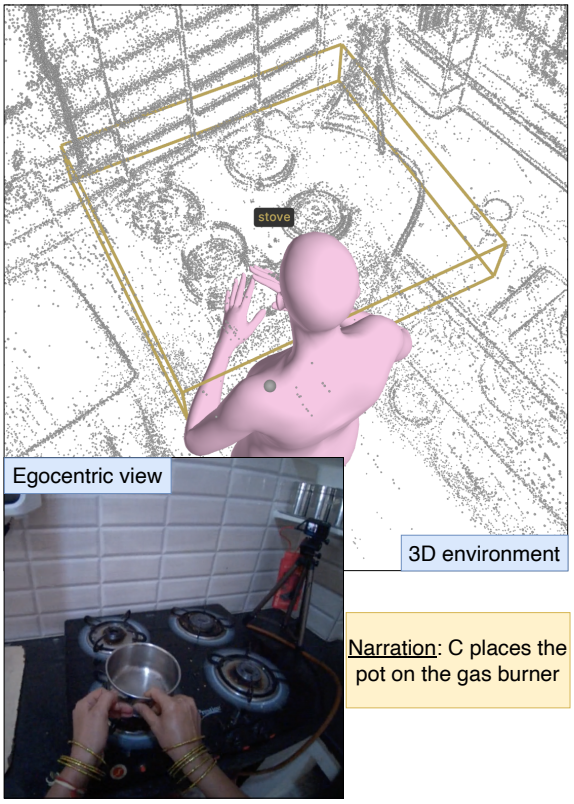


Figure 3. **An interaction instance.** We mark a timestamp as an interaction when the hands are within the 3D bounding box of an object referenced in the narration. An LLM is used to match the narration with the objects in the detector's vocabulary, e.g., *stove* and *gas burner*.

**Extracting body pose.** We use the state-of-the-art WHAM [92] model to obtain body pose and shape from videos. The method first tracks the person in the video [104], followed by a regression head that predicts the SMPL parameters $(\theta, \beta, t)$. Even though the Ego-Exo4D dataset assumes one actor per video, there are additional people present in the video that are tracked by WHAM; see Supp. for examples, and how we disambiguate the actor. Also, though the dataset contains multi-view videos, we use only the view having the maximum joint visibility for extracting SMPL. Lastly, we extract all the body poses in the local coordinate system, and use the egocentric parameters to place the person in the 3D global coordinate frame.

**Finding interaction instances.** After obtaining the body pose and object bounding boxes, the final stage of the data preparation involves finding the interaction timestamps. One potential approach is to simply find the intersection points of the body pose hands with the object bounding boxes. However, this would be susceptible to minor errors in the estimated 3D object bounding boxes and body pose

parameters. Hence, for a good quality dataset, we propose a geometric video-language approach. We take the natural language "narrations" describing each action in Ego-Exo4D and their accompanying timestamps. Then we use Llama-3.1-8B [21] to classify all the narrations into either a touch-based interaction or a non-touch interaction, and to match the object mentioned in the narration to the object detection vocabulary [39]—a task well-suited for a large language model. For example, *"the person picks up a metal skillet holder"* shows an interaction with a skillet, whereas *"the person watches the skillet"* does not. See Supp. for the prompt details. Using these outputs, a final interaction instance is when either of the person's hands is within the 3D object bounding box *and* the LLM marks a timestamp as a touch-based interaction. See Fig. 3.

In summary, we use videos and state-of-the-art methods to extract objects, body poses, and interaction instances, all registered in the same coordinate space. While our implementation takes advantage of the high quality camera calibration offered in Ego-Exo4D, as visual SLAM continues to improve [22, 59, 75, 96], such an approach will generalize increasingly better to lower quality data as well. We emphasize that ours is the first work to curate a dataset for future interaction prediction in 4D videos. We are sharing our data to facilitate benchmarking by other researchers.

### 3.4. Implementation details

**Dataset and statistics.** Ego-Exo4D [38] contains 5,035 takes covering physical and procedural scenarios. Physical scenarios like soccer and basketball are typically performed outdoors and/or contain limited object interaction. Thus, we focus on the procedural scenarios—*cooking*, *bike repair*, and *health*, which are comprised of 196,363 total instances of interactions (e.g., *pick up cup*, *tighten chain*, *twist open bottle*, etc.), and span 92 unique environments. Each take has, on average, 130 interactions, with 18 distinct object interactions. The average length of a take is 5.75 minutes. Following our task definition (see Sec. 3.1), we set $\tau_o = 30s, \tau_a = 5s, \tau_f = 180s$.

This dataset creation strategy results in 87,373 training (cooking: 62,495, bike repair: 13,589, health: 11,289), 10,563 validation (cooking: 7,837, bike repair: 1,478, health: 1,248), and 10,124 test episodes (cooking: 7,197, bike repair: 771, health: 2,156). We apply the same take-level train/test split as in [38]. The environment encountered at test time is never observed during training. Across cooking, bike repair, and health, each episode has an average of 10, 20 and 6 future interactions and 3.5 different body poses per interaction location, respectively.

**Network architecture and parameters.** $f_V$ is the EgoVLPv2 [84] visual encoder. We use pre-extracted features for a faster I/O. All the mappers $f_{vm}$, $f_{pm}$, and $f_{lm}$ are linear layers, with the same output dimension, i.e., 4096. The input dimension to $f_{vm}$ is 4096, consistent with the output of $f_V$. Similarly, $f_{pm}$ has 207 ($23 \times 3 \times 3$) input dimension representing the SMPL pose $\theta$, and $f_{lm}$ represents a location in the $N \times N \times N$ grid, i.e., $3N$ input dimension. We use $N = 16$ to convert the 3D space into voxels. This choice reasonably captures the objects in the environment, while being experimentally feasible (i.e., fits the GPU memory). The method can be directly applied to any $N$, with the model parameters scaling as $N^3$, same as other voxel-based methods [55, 114]. All modalities are processed at 1 feature per second. We train the distinct scenarios separately (cooking, health, bike repair). We train all the models on eight Quadro RTX 6000 23GB GPUs. We set a learning rate of $5 \times 10^{-5}$ for 3 epochs for predicting interaction location and $5 \times 10^{-6}$ for 6 epochs for predicting interaction pose. All runs take a maximum of 10 hours.

## 4. Experiments

We first describe the baselines, ablations, and metrics. Then, we discuss our experimental results and comparisons, along with some output visualizations.

**Baselines.** We compare to a total of 6 state-of-the-art models developed for related tasks. While these models are all relevant, none of them exactly performs 4D future interaction prediction, so we appropriately modify the existing models to create strong baselines organized into two families: autoregressive and video-to-3D. The family of autoregressive baselines (**HierVL** [5] for long-term anticipation, **OCT** [63] for hotspot prediction, **4D-Humans** [35] and **T2M-GPT** [111] for pose generation/prediction) are transformer-based models that use the observed context to generate the next token for a given modality (e.g., action label, next frame heatmap, or future body pose). For fair comparison to our model, we use the camera parameters in Ego-Exo4D to lift the 2D models' predictions of interaction points into the 3D environment. The video-to-3D models (**VoxFormer** [55], **OccFormer** [114], **Video-to-pose** CVAE [93]) directly predict the 3D location or pose as output from video, hence implicitly learning the 3D semantics of the scene. VoxFormer [55] and OccFormer [114] are originally trained to predict 3D scene occupancy from images, so we adapt to video. We train all the video-to-3D prediction models on our data and explore fine-tuning T2M-GPT-FT [111] on our dataset as well (**T2M-GPT-FT**). Please see the Supp. for all implementation details.

**Ablations.** In addition to the baselines, we also compare against various ablations of the architecture design, where we remove each component of our method's input in turn to discern their impact: the observed egocentric video (**FICTION w/o video**), the person's body poses (**FICTION w/o pose**), and the environment context, i.e., object layout (**FICTION-w/o env**). See Supp. for ablations with hyperparameters.

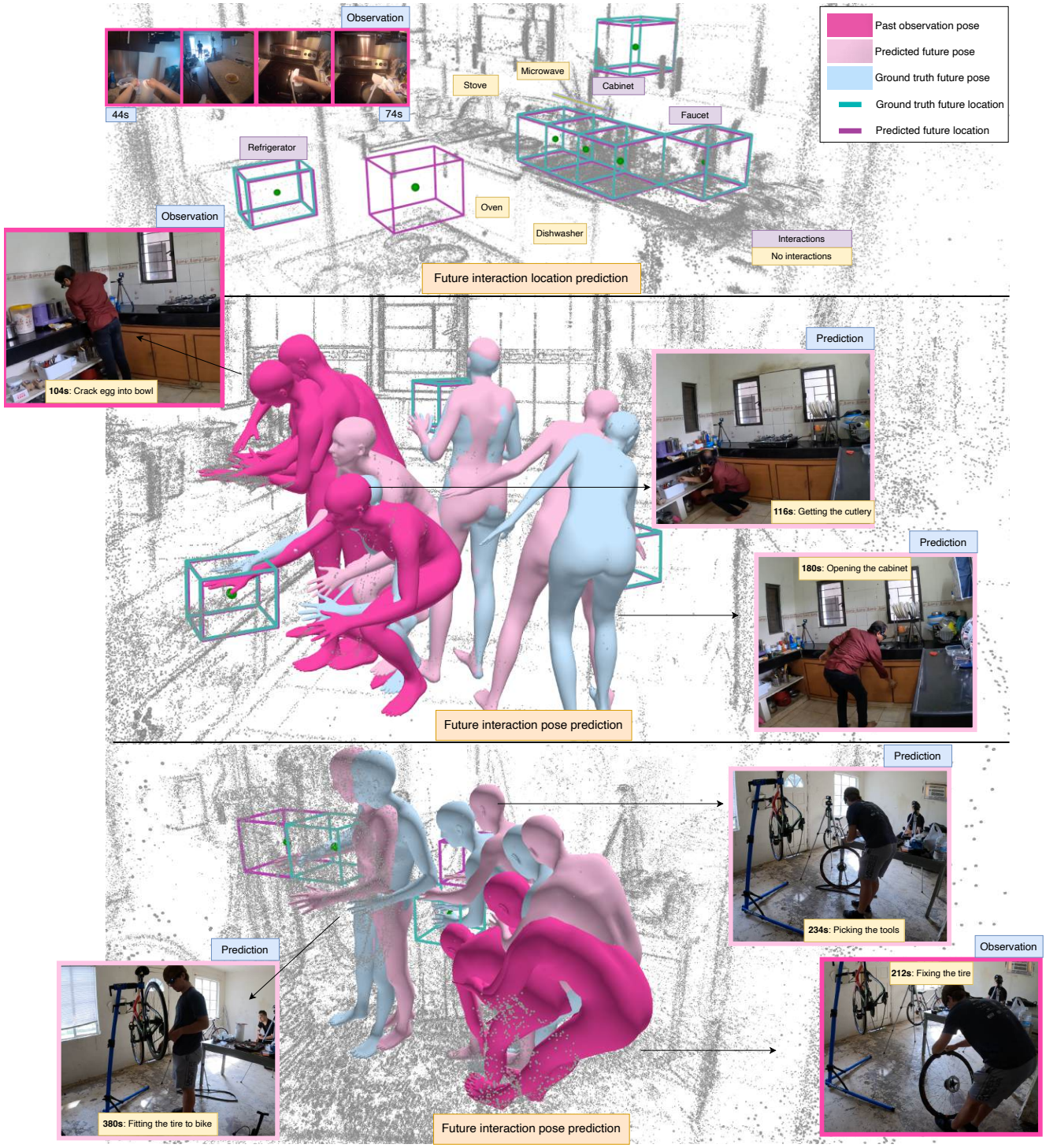


Figure 4. **Visualization of example results.** (Top) Interaction location prediction in a cooking take. Based on the observed input ego, the model is able to correctly predict future interaction locations— *refrigerator, faucet, cabinet*. (Middle, bottom): Interaction sequences in a cooking and a bike repair take. Given the past observation information, our model is able to accurately predict future interactions. If the observed sequence shows the person fixing the tire, the model predicts that the tire will be fixed onto the bike at a later stage, at a different spatial location (bottom, left most pose visualization). Best viewed in zoom. Exo view shown for visualization only.

**Future interaction location prediction**

| | Cooking | | Bike Repair | | Health | |
|---|---|---|---|---|---|---|
| | PR | Ch ↓ | PR | Ch ↓ | PR | Ch ↓ |
| HierVL [5] | 11.2 | 11.0 | 10.6 | 11.3 | 7.8 | 51.3 |
| OCT [63] | 16.9 | 9.0 | 13.3 | 9.4 | 11.2 | 44.8 |
| OccFormer [114] | 13.6 | 9.8 | 14.0 | 10.0 | 10.3 | 46.5 |
| VoxFormer [55] | 15.1 | 9.5 | 14.1 | 10.5 | 9.9 | 46.5 |
| FICTION | **21.0** | **7.4** | **18.7** | **7.2** | **12.7** | **41.7** |
| w/o video | 20.0 | 7.6 | **18.7** | 9.2 | 12.0 | 43.7 |
| w/o pose | 19.1 | 7.7 | 18.3 | 7.6 | 11.7 | 43.4 |
| w/o env | 9.9 | 11.4 | 6.0 | 13.2 | 4.7 | 56.0 |

**Future interaction pose prediction**

| | Cooking | | Bike Repair | | Health | |
|---|---|---|---|---|---|---|
| | M ↓ | PA ↓ | M ↓ | PA ↓ | M ↓ | PA ↓ |
| 4D-Humans [35] | 473 | 65 | 492 | 125 | 307 | 72 |
| T2M-GPT [111] | 397 | 65 | 470 | 117 | 276 | 70 |
| T2M-GPT-FT [111] | 264 | 61 | 410 | 99 | 221 | 66 |
| Video-to-pose | 267 | 60 | 402 | 97 | 226 | 66 |
| FICTION | **229** | **56** | **372** | **91** | **172** | **62** |
| w/o video | 234 | 58 | 375 | 92 | 175 | 68 |
| w/o pose | 236 | 60 | 382 | 95 | 176 | 66 |
| w/o env | 336 | 63 | 475 | 93 | 280 | 67 |

Table 1. **Results of future interaction prediction.** We show the results for both future interaction location prediction (left) and future interaction pose prediction (right) for three scenarios—cooking, bike repair, and health. We outperform all prior work on both the tasks in all scenarios. See text for details. (PR: precision-recall area under curve, Ch: chamfer distance, M: MPJPE, PA: PA-MPJPE.)

**Metrics.** For future interaction location prediction, we use Chamfer distance and PR-AUC (precision-recall area under curve). Chamfer distance is lower for a better method and reported in voxel units, whereas PR-AUC ranges in $[0, 1]$, higher the better (shown out of 100 for clarity). For future interaction pose prediction, we report Mean Per Joint Position Error (MPJPE) in world-coordinates, and Procrustes Aligned MPJPE (PA-MPJPE), following prior work [35]. Both metrics measure the per-joint error in mm, lower the better. Recall that we predict a distribution of the pose parameters for a given object interaction location. At inference, we sample $N = 5$ poses, and report the error of the closest pose w.r.t. the ground truth, same for any baseline that outputs multiple poses such as T2M-GPT.

**Results.** Tab. 1 (left) shows the results for future interaction location prediction. Our method significantly outperforms all prior work on all scenarios for both the PR-AUC and Chamfer metrics, with relative gains more than 32% (absolute 4.6%). In particular, autoregressive models (HierVL [5], OCT [63]) are good at predicting the next few interaction locations, but the errors in the prediction accumulates and the location often diverges. Similarly, video-to-3D future prediction models (Occformer [114], VoxFormer [55]) also fail to learn the future location due to their lack of an explicit activity and 3D reasoning.

Comparing with ablations, providing both the video and pose sequence improves the performance over a single modality. The performance drops significantly if the location context of the environment is not provided—showcasing its importance. Note that other autoregressive methods are given location information as *late-fusion*. The performance in health (COVID-19 test) is lower than other scenarios due to fewer interactions with objects in the chosen off-the-shelf object vocabulary [39], meaning limited context is passed to the model.

Tab. 1 (right) shows the results for future interaction pose prediction. Our method outperforms all baselines, by up to 49 mm. Same as above, the autoregressive methods (4D-humans [35] and T2M-GPT [111]) diverge when required to generate very long (3 min) pose sequences. Fine-tuning T2M-GPT marginally improves the performance, with the autoregressive nature of sequential pose generation being the limiting factor. Similarly, the vanilla video-to-pose CVAE is unable to capture the future poses, due to the missing environment context. We observe the same trend as above when comparing with the ablations. Note that in case of 'w/o pose', the model is still strong due to the other modalities—environment context, including the actor's location, and the video observation.

Finally, Fig. 4 shows some qualitative results. Our method is able to predict the correct interaction locations across different parts of the environment. Given the past observation of cleaning a pan, the model predicts fetching veggies from the refrigerator and cabinet, and using the faucet (top). The model also predicts that microwave, stove, among other things, will not be interacted with in the next 3 minutes. We also see how our method can identify the current activity and predict the future pose, e.g., if the person is fixing a bike tire, the person will likely interact with the bike to put the tires back on (see Fig. 4, middle and bottom). For a qualitative comparison between our method and baselines, please see Supp. Fig. 5.

Overall, these strong results suggest the effectiveness of our proposed model for future interaction prediction, and our dataset and evaluation paradigm establish a valuable benchmark for continued work.

## 5. Conclusion

We propose FICTION, a novel method to predict future interactions. We use video observations, along with the 3D scene context, to anticipate the location of the interaction and the person's body pose during that future interaction. We design a multimodal architecture to capture the activity intent and the person's location in the 3D scene. Our performance improvement over state-of-the-art work addressing related problems showcases the effectiveness of our approach and the novelty of the task itself. In the future, we plan to explore generalizations for predicting the movements of dynamic objects and a streaming variant that could repeatedly revise its future predictions.

## References


[1] Yazan Abu Farha, Alexander Richard, and Juergen Gall. When will you do what?-anticipating temporal occurrences of activities. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5343–5352, 2018. 2

[2] Emre Aksan, Manuel Kaufmann, Peng Cao, and Otmar Hilliges. A spatio-temporal transformer for 3d human motion prediction. In *2021 International Conference on 3D Vision (3DV)*, pages 565–574. IEEE, 2021. 3

[3] Ziad Al-Halah, Santhosh Kumar Ramakrishnan, and Kristen Grauman. Zero experience required: Plug & play modular transfer learning for semantic visual navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17031–17041, 2022. 1

[4] Peter Anderson, Angel Chang, Devendra Singh Chaplot, Alexey Dosovitskiy, Saurabh Gupta, Vladlen Koltun, Jana Kosecka, Jitendra Malik, Roozbeh Mottaghi, Manolis Savva, et al. On evaluation of embodied navigation agents. *arXiv preprint arXiv:1807.06757*, 2018. 1

[5] Kumar Ashutosh, Rohit Girdhar, Lorenzo Torresani, and Kristen Grauman. Hiervl: Learning hierarchical video-language embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23066–23078, 2023. 1, 2, 6, 8

[6] Kumar Ashutosh, Tushar Nagarajan, Georgios Pavlakos, Kris Kitani, and Kristen Grauman. ExpertAF: Expert actionable feedback from video. *arXiv preprint arXiv:2408.00672*, 2024. 1

[7] Kumar Ashutosh, Santhosh Kumar Ramakrishnan, Triantafyllos Afouras, and Kristen Grauman. Video-mined task graphs for keystep recognition in instructional videos. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[8] Shikhar Bahl, Russell Mendonca, Lili Chen, Unnat Jain, and Deepak Pathak. Affordances from human videos as a versatile representation for robotics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13778–13790, 2023. 1, 3

[9] Dhruv Batra, Aaron Gokaslan, Aniruddha Kembhavi, Oleksandr Maksymets, Roozbeh Mottaghi, Manolis Savva, Alexander Toshev, and Erik Wijmans. Objectnav revisited: On evaluation of embodied agents navigating to objects. *arXiv preprint arXiv:2006.13171*, 2020. 1

[10] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, page 4, 2021. 4

[11] Jing Bi, Jiebo Luo, and Chenliang Xu. Procedure planning in instructional videos via contextual modeling and model-based policy learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15611–15620, 2021. 2

[12] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qi-Zhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 387–404. Springer, 2020. 1, 3

[13] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 2

[14] Chien-Yi Chang, De-An Huang, Danfei Xu, Ehsan Adeli, Li Fei-Fei, and Juan Carlos Niebles. Procedure planning in instructional videos. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI*, pages 334–350. Springer, 2020. 2

[15] Xuxin Cheng, Yandong Ji, Junming Chen, Ruihan Yang, Ge Yang, and Xiaolong Wang. Expressive whole-body control for humanoid robots. *arXiv preprint arXiv:2402.16796*, 2024. 1

[16] Enric Corona, Albert Pumarola, Guillem Alenya, and Francesc Moreno-Noguer. Context-aware human motion prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6992–7001, 2020. 3

[17] Serhan Coşar, Manuel Fernandez-Carmona, Roxana Agrigoroaie, Jordi Pages, François Ferland, Feng Zhao, Shigang Yue, Nicola Bellotto, and Adriana Tapus. Enrichme: Perception and interaction of an assistive robot for the elderly at home. *International Journal of Social Robotics*, 12:779–805, 2020. 1

[18] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision: collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, 130(1):33–55, 2022. 1

[19] Shengheng Deng, Xun Xu, Chaozheng Wu, Ke Chen, and Kui Jia. 3d affordancenet: A benchmark for visual object affordance understanding. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1778–1787, 2021. 2

[20] Christian Diller and Angela Dai. Cg-hoi: Contact-guided 3d human-object interaction generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19888–19901, 2024. 2

[21] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 6, 1, 3

[22] Hugh Durrant-Whyte and Tim Bailey. Simultaneous localization and mapping: part i. *IEEE robotics & automation magazine*, 13(2):99–110, 2006. 1, 6

[23] Nikita Dvornik, Isma Hadji, Hai Pham, Dhaivat Bhatt, Brais Martinez, Afsaneh Fazly, and Allan D Jepson. Flow graph to video grounding for weakly-supervised multi-step localization. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV*, pages 319–335. Springer, 2022. 2

[24] Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, Yao Feng, and Michael J Black. Tokenhmr: Advancing human mesh re-

covery with a tokenized pose representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1323–1333, 2024. 3

[25] Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brighid Meredith, et al. Project aria: A new tool for egocentric multi-modal ai research. *arXiv preprint arXiv:2308.13561*, 2023. 1, 5

[26] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, pages 226–231, 1996. 5, 1

[27] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 2

[28] Yao Feng, Jing Lin, Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, and Michael J. Black. Chatpose: Chatting about 3d human pose, 2024. 1

[29] Antonino Furnari and Giovanni Maria Farinella. Rolling-unrolling lstms for action anticipation from first-person video. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):4021–4036, 2020. 2

[30] Antonino Furnari, Sebastiano Battiato, Kristen Grauman, and Giovanni Maria Farinella. Next-active-object prediction from egocentric videos. *Journal of Visual Communication and Image Representation*, 49:401–411, 2017. 1

[31] Jiyang Gao, Zhenheng Yang, and Ram Nevatia. Red: Reinforced encoder-decoder networks for action anticipation. *arXiv preprint arXiv:1707.04818*, 2017.

[32] Jiyang Gao, Zhenheng Yang, and Ram Nevatia. Red: Reinforced encoder-decoder networks for action anticipation. *arXiv preprint arXiv:1707.04818*, 2017.

[33] Rohit Girdhar and Kristen Grauman. Anticipative video transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13505–13515, 2021. 1, 2, 3

[34] Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens van der Maaten, Armand Joulin, and Ishan Misra. Omnivore: A single model for many visual modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16102–16112, 2022. 2

[35] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4d: Reconstructing and tracking humans with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14783–14794, 2023. 1, 2, 3, 5, 6, 8

[36] Patrick Grady, Chengcheng Tang, Christopher D Twigg, Minh Vo, Samarth Brahmbhatt, and Charles C Kemp. Contactopt: Optimizing contact to improve grasps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1471–1481, 2021. 2

[37] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 2, 1

[38] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *CVPR*, 2024. 2, 5, 6, 3

[39] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019. 4, 5, 6, 8, 1

[40] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2282–2292, 2019. 3

[41] Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael J Black. Stochastic scene-aware motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11374–11384, 2021. 1, 3

[42] Mohamed Hassan, Yunrong Guo, Tingwu Wang, Michael Black, Sanja Fidler, and Xue Bin Peng. Synthesizing physical character-scene interactions. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–9, 2023. 3

[43] Tairan He, Zhengyi Luo, Wenli Xiao, Chong Zhang, Kris Kitani, Changliu Liu, and Guanya Shi. Learning human-to-humanoid real-time whole-body teleoperation. *arXiv preprint arXiv:2403.04436*, 2024. 1

[44] Siyuan Huang, Zan Wang, Puhao Li, Baoxiong Jia, Tengyu Liu, Yixin Zhu, Wei Liang, and Song-Chun Zhu. Diffusion-based generation, optimization, and planning in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16750–16761, 2023. 1

[45] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2014. 2

[46] Hanwen Jiang, Shaowei Liu, Jiashun Wang, and Xiaolong Wang. Hand-object contact consistency reasoning for human grasps generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11107–11116, 2021. 2

[47] Haoyi Jiang, Tianheng Cheng, Naiyu Gao, Haoyang Zhang, Tianwei Lin, Wenyu Liu, and Xinggang Wang. Symphonize 3d semantic scene completion with contextual instance queries, 2023. 1

[48] Yingying Jiang, Xiangyu Zhu, Xiaobing Wang, Shuli Yang, Wei Li, Hua Wang, Pei Fu, and Zhenbo Luo. R2cnn: Rotational region cnn for orientation robust scene text detection. *arXiv preprint arXiv:1706.09579*, 2017. 5

[49] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer

Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 5

[50] Theodora Kontogianni, Ekin Celikkan, Siyu Tang, and Konrad Schindler. Interactive object segmentation in 3d point clouds. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2891–2897. IEEE, 2023. 5

[51] Tian Lan, Tsung-Chuan Chen, and Silvio Savarese. A hierarchical representation for future action prediction. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part III 13*, pages 689–704. Springer, 2014. 2

[52] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890*, 2023. 4

[53] Kunchang Li, Yali Wang, Junhao Zhang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unifying convolution and self-attention for visual recognition. *arXiv preprint arXiv:2201.09450*, 2022. 2

[54] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4804–4814, 2022. 2

[55] Yiming Li, Zhiding Yu, Christopher Choy, Chaowei Xiao, Jose M. Alvarez, Sanja Fidler, Chen Feng, and Anima Anandkumar. Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9087–9098, 2023. 2, 6, 8

[56] Kevin Qinghong Lin, Alex Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Zhongcong Xu, Difei Gao, Rongcheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. In *NeurIPS*, 2022. 2

[57] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 3

[58] Xudong Lin, Fabio Petroni, Gedas Bertasius, Marcus Rohrbach, Shih-Fu Chang, and Lorenzo Torresani. Learning to recognize procedural activities with distant supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13853–13863, 2022. 2

[59] Lahav Lipson, Zachary Teed, and Jia Deng. Deep patch visual slam. In *European Conference on Computer Vision*, pages 424–440. Springer, 2025. 1, 6

[60] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 4

[61] Miao Liu, Siyu Tang, Yin Li, and James M Rehg. Forecasting human-object interaction: joint prediction of motor attention and actions in first person video. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 704–721. Springer, 2020. 1, 3

[62] Shaowei Liu, Subarna Tripathi, Somdeb Majumdar, and Xiaolong Wang. Joint hand motion and interaction hotspots prediction from egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3282–3292, 2022. 3

[63] Shaowei Liu, Subarna Tripathi, Somdeb Majumdar, and Xiaolong Wang. Joint hand motion and interaction hotspots prediction from egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3282–3292, 2022. 1, 2, 6, 8

[64] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: a skinned multi-person linear model. *ACM Trans. Graph.*, 34(6), 2015. 3, 4

[65] Zhenyu Lou, Qiongjie Cui, Haofan Wang, Xu Tang, and Hong Zhou. Multimodal sense-informed forecasting of 3d human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2144–2154, 2024. 3

[66] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5442–5451, 2019. 2

[67] Wei Mao, Richard I Hartley, Mathieu Salzmann, et al. Contact-aware human motion forecasting. *Advances in Neural Information Processing Systems*, 35:7356–7367, 2022. 3

[68] Julieta Martinez, Michael J Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2891–2900, 2017. 3

[69] Esteve Valls Mascaro, Hyemin Ahn, and Dongheui Lee. Intention-conditioned long-term human egocentric action forecasting@ ego4d challenge 2022. *arXiv preprint arXiv:2207.12080*, 2022. 2

[70] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2630–2640, 2019. 2

[71] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9879–9889, 2020. 2

[72] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view syn-

thesis. *Communications of the ACM*, 65(1):99–106, 2021. 3

[73] Andrew T Miller and Peter K Allen. Graspit! a versatile simulator for robotic grasping. *IEEE Robotics & Automation Magazine*, 11(4):110–122, 2004. 2

[74] Kaichun Mo, Leonidas J Guibas, Mustafa Mukadam, Abhinav Gupta, and Shubham Tulsiani. Where2act: From pixels to actions for articulated 3d objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6813–6823, 2021. 2

[75] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: a versatile and accurate monocular slam system. *IEEE transactions on robotics*, 31(5):1147–1163, 2015. 1, 6

[76] Lorenzo Mur-Labadia, Jose J Guerrero, and Ruben Martinez-Cantin. Multi-label affordance mapping from egocentric vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5238–5249, 2023. 5

[77] Lorenzo Mur-Labadia, Jose J Guerrero, and Ruben Martinez-Cantin. Multi-label affordance mapping from egocentric vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5238–5249, 2023. 3

[78] Lorenzo Mur-Labadia, Ruben Martinez-Cantin, Josechu Guerrero, Giovanni Maria Farinella, and Antonino Furnari. Aff-ttention! affordances and attention models for short-term object interaction anticipation. *arXiv preprint arXiv:2406.01194*, 2024. 1, 3

[79] T. Nagarajan and K. Grauman. Learning affordance landscapes for interaction exploration in 3d environments. In *Proceedings of the Advances on Neural Information Processing Systems (NeurIPS)*, 2020. 3

[80] Tushar Nagarajan, Christoph Feichtenhofer, and Kristen Grauman. Grounded human-object interaction hotspots from video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 1, 3

[81] Lukas Neumann, Andrew Zisserman, and Andrea Vedaldi. Future event prediction: If and when. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 2

[82] Andreas ten Pas and Robert Platt. Using geometry to detect grasps in 3d point clouds. *arXiv preprint arXiv:1501.03100*, 2015. 2

[83] 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 *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 3

[84] Shraman Pramanick, Yale Song, Sayan Nag, Kevin Qinghong Lin, Hardik Shah, Mike Zheng Shou, Rama Chellappa, and Pengchuan Zhang. Egovlpv2: Egocentric video-language pre-training with fusion in the backbone. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5285–5297, 2023. 6

[85] Xavier Puig, Eric Undersander, Andrew Szot, Mikael Dallaire Cote, Tsung-Yen Yang, Ruslan Partsey, Ruta Desai, Alexander William Clegg, Michal Hlavac, So Yeon Min, et al. Habitat 3.0: A co-habitat for humans, avatars and robots. *arXiv preprint arXiv:2310.13724*, 2023. 3

[86] Tong Qin, Peiliang Li, and Shaojie Shen. Vins-mono: A robust and versatile monocular visual-inertial state estimator. *IEEE transactions on robotics*, 34(4):1004–1020, 2018. 5

[87] I Radosavovic, X Wang, L Pinto, and J Malik. State-only imitation learning for dexterous manipulation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2021. 1

[88] Santhosh Kumar Ramakrishnan, Devendra Singh Chaplot, Ziad Al-Halah, Jitendra Malik, and Kristen Grauman. Poni: Potential functions for objectgoal navigation with interaction-free learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18890–18900, 2022. 1

[89] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J Guibas. Humor: 3d human motion model for robust pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11488–11499, 2021. 3

[90] Nicholas Rhinehart and Kris M Kitani. First-person activity forecasting with online inverse reinforcement learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3696–3705, 2017. 2

[91] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 5

[92] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J Black. Wham: Reconstructing world-grounded humans with accurate 3d motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2070–2080, 2024. 3, 5

[93] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28, 2015. 3, 4, 6, 2

[94] Bilge Soran, Ali Farhadi, and Linda Shapiro. Generating notifications for missing actions: Don't forget to turn the lights off! In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4669–4677, 2015. 1

[95] Lyne Tchapmi, Christopher Choy, Iro Armeni, JunYoung Gwak, and Silvio Savarese. Segcloud: Semantic segmentation of 3d point clouds. In *2017 international conference on 3D vision (3DV)*, pages 537–547. IEEE, 2017. 5

[96] Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *Advances in neural information processing systems*, 34:16558–16569, 2021. 1, 6

[97] Vadim Tschernezki, Ahmad Darkhalil, Zhifan Zhu, David Fouhey, Iro Larina, Diane Larlus, Dima Damen, and Andrea Vedaldi. EPIC Fields: Marrying 3D Geometry and Video Understanding. In *Proceedings of the Neural Information Processing Systems (NeurIPS)*, 2023. 5

[98] Joost Van Amersfoort, Anitha Kannan, Marc'Aurelio Ranzato, Arthur Szlam, Du Tran, and Soumith Chintala. Transformation-based models of video sequences. *arXiv preprint arXiv:1701.08435*, 2017. 2

[99] Chaoqun Wang, Jiyu Cheng, Wenzheng Chi, Tingfang Yan, and Max Q.-H. Meng. Semantic-aware informative path planning for efficient object search using mobile robot. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 51(8):5230–5243, 2021. 1

[100] Jiashun Wang, Huazhe Xu, Jingwei Xu, Sifei Liu, and Xiaolong Wang. Synthesizing long-term 3d human motion and interaction in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9401–9411, 2021. 1, 3

[101] Yian Wang, Ruihai Wu, Kaichun Mo, Jiaqi Ke, Qingnan Fan, Leonidas J Guibas, and Hao Dong. Adaafford: Learning to adapt manipulation affordance for 3d articulated objects via few-shot interactions. In *European conference on computer vision*, pages 90–107. Springer, 2022. 2

[102] Chao-Yuan Wu, Yanghao Li, Karttikeya Mangalam, Haoqi Fan, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Memvit: Memory-augmented multiscale vision transformer for efficient long-term video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13587–13597, 2022. 2

[103] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*, 2021. 2

[104] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in Neural Information Processing Systems*, 35:38571–38584, 2022. 5

[105] Haitao Yan, Qiongjie Cui, Jiexin Xie, and Shijie Guo. Forecasting of 3d whole-body human poses with grasping objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1726–1736, 2024. 3

[106] Jianwei Yang, Yonatan Bisk, and Jianfeng Gao. Taco: Token-aware cascade contrastive learning for video-text alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11562–11572, 2021. 4

[107] Yuhang Yang, Wei Zhai, Hongchen Luo, Yang Cao, and Zheng-Jun Zha. Lemon: Learning 3d human-object interaction relation from 2d images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16284–16295, 2024. 2

[108] Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Decoupling human and camera motion from videos in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21222–21232, 2023. 3

[109] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22479–22489, 2023. 2

[110] Ye Yuan and Kris Kitani. Dlow: Diversifying latent flows for diverse human motion prediction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 346–364. Springer, 2020. 3

[111] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14730–14740, 2023. 1, 2, 6, 8

[112] Jason Y Zhang, Panna Felsen, Angjoo Kanazawa, and Jitendra Malik. Predicting 3d human dynamics from video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7114–7123, 2019. 3

[113] Yan Zhang, Michael J Black, and Siyu Tang. We are more than our joints: Predicting how 3d bodies move. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3372–3382, 2021. 3

[114] Yunpeng Zhang, Zheng Zhu, and Dalong Du. Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9433–9443, 2023. 2, 6, 8

[115] Yang Zheng, Yanchao Yang, Kaichun Mo, Jiaman Li, Tao Yu, Yebin Liu, C Karen Liu, and Leonidas J Guibas. Gimo: Gaze-informed human motion prediction in context. In *European Conference on Computer Vision*, pages 676–694. Springer, 2022. 3

[116] Yiwu Zhong, Licheng Yu, Yang Bai, Shangwen Li, Xueting Yan, and Yin Li. Learning procedure-aware video representation from instructional videos and their narrations. *arXiv preprint arXiv:2303.17839*, 2023. 2

[117] Honglu Zhou, Roberto Martín-Martín, Mubbasir Kapadia, Silvio Savarese, and Juan Carlos Niebles. Procedure-aware pretraining for instructional video understanding. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2

[118] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *European Conference on Computer Vision*, pages 350–368. Springer, 2022. 5, 1, 3

# FICTION: 4D Future Interaction Prediction from Video

## Supplementary Material

## A. List of supplementary materials

We attach a supplementary video containing an overview of the paper, including dataset and result visualization. We will also release the interaction dataset and code.

## B. Future interaction dataset

This section contains additional details about the future interaction dataset, discussed in Sec. 3.3.

**3D object bounding boxes.** We use Detic [118], along with the object taxonomy from LVIS [39], to find the mapping between the pixels in the video frames and the object labels. Since the inference from this method is fast, we perform segmentation at the original frame rate, i.e., 30 frames per second. Note that each pixel on the SLAM camera has an associated 3D location, which we use to map the object labels to a point in the 3D space. We perform object segmentation on SLAM frames directly because they have a direct mapping from 2D to 3D.

The DBSCAN [26] algorithm mentioned in Sec. 3.3 is useful in tightening the bounding boxes. For example, if there are two chairs in the scene, attempting to create a bounding box directly results in the box containing everything between the two chairs. Thus, we use DBSCAN to find the approximate bounding boxes. We set $50cm$ as the threshold for distinct clusters, and require 100 points at least to register as a unique object. This choice can correctly capture most of the objects seen in the chosen scenarios.

**Extracting body poses.** As mentioned in Sec. 3.3, the dataset has only one actor per video. However, there are other people present in some views. They are either bystanders or data collection volunteers. The dataset does not provide a full coverage of the annotation of the main actor in the videos. Thus, we use our heuristics to use the multi-view and disambiguate the main actor. Furthermore, the dataset contains multi-view videos showing the same person. We use the following two observations to extract the human pose. Firstly, the camera rig in the capture setup ensures the actor will have the largest area in all the video frames. Secondly, the similarity of the poses of the same person from all views will be higher than with people in the background. We use these observations to find the actor and then choose the *best* view—having the maximum joint visibility—to obtain the extracted body pose. An alternative is to focus on maximum hand visibility. However, we do not over-emphaisze on the hands. Furthermore, comparing to manually annotated 3D poses in Ego-Exo4D (available for only a subset of the data), the MPJPE error is 82mm when we use max joints and 115mm when we use maximum hand visibility.

**Finding interaction instances.** We use the following prompt for Llama 3.1-8B [21]:

> **System:** You are a helpful AI assistant. Match the narrations with the object labels that is provided.
> **User:** You are given narrations labeled by human annotators for a video. You are also given a set of object labels as per an object detection vocabulary. Find all instances of object interaction where the person would touch an object and map it to all the synonyms or similar words in the vocabulary. Sentences like 'C looks at the fridge' has no object interaction. Objects like cup, glass can be grouped together. Here are the object labels that you have to use: {labels}. Answer in this format:
> 1. {rewrite first narration} - answer: (object1, object2)
> 2. {rewrite second narration} - answer: NO INTERACTION
> 3. {rewrite third narration} - answer: NO MATCHING OBJECTS. Use 'NO INTERACTION' and 'NO MATCHING OBJECTS' in cases with no interaction and matching objects, respectively. Here are the numbered narrations:
> {narrations}

## C. Details of baseline implementation

We introduce the baselines in Sec. 4. None of the baselines are directly applicable for 4D interaction prediction. Thus, we appropriately modify related models to create strong baselines for comparison. The model and task-specific adaptations are listed below:

- **HierVL [5]** is a recent method in Ego4D [37] long-term anticipation (LTA) benchmark with publicly available codebase. This LTA version generates future action labels (nouns and verbs). We use the output noun and locate the same in the 3D space, and mark all voxels for the predicted object as future interaction locations. Since HierVL is initially pretrained on Ego4D, we do not need to finetune the dataset since the egocentric videos are from a similar distribution. We do, however, finetune the last layer to match the output class dimension to the objects detected in Ego-Exo4D scenes.
- **OCT [63]** is a recent work in joint hand motion and interaction hotspot prediction from EPIC-Kitchens-100 [18]. We use this method to predict future interaction hotspot

for the next 3 minutes. We then use the camera parameters in Ego-Exo4D to map the 2D interaction points into the 3D environment. Since, this model is also trained on egocentric videos, and just requires images as input, we do not retrain this method on Ego-Exo4D.

- **OccFormer [114]** and **VoxFormer [55]** are methods originally designed for occupancy map prediction. We replace the image encoder in these networks with the video encoder $f_V$, used in our method.
- **4D-Humans [35]** and **T2M-GPT [111]** are recent works with autoregressive pose prediction capabilities. 4D-Humans extracts body pose from images and videos. We use the pose prediction module that predicts the next pose given the current body pose. We use this transformer autoregressively to generate multiple possible poses in the future. Similarly, T2M-GPT converts the body pose into a VQ-VAE based tokens and then predicts the pose tokens. The model is originally designed to generate pose based on the text condition; we modify the model to input prior pose tokens. Since our focus is not on *when* a pose is happening but rather *where*, we generously choose the prediction as the closest pose to the ground truth interaction location, out of all the generations. Both the methods are trained on large-scale pose datasets [45, 66]. Regardless, we finetune T2M-GPT (called **T2M-GPT-FT**) on our dataset to investigate the role of the training data. We choose to finetune the latter model due to a better performance and the stable nature of VQ-VAE codebooks for pose token generation.
- **Video-to-pose** CVAE [93] model takes as input the video of the person and generates a future pose distribution. We use the same video encoder $f_V$ but do not provide any additional 3D context and expect the model to learn the 3D semantics implicitly. We train this method on our dataset. At inference, we choose the pose closest to the ground truth location.

**Qualitative comparison with baselines.** Fig. 5 compares our output with baselines. We see that our method is able to predict the interaction location and pose better than both the baselines. Autoregressive methods cannot predict long-term change in location and pose, while video-to-3D additionally misses the correct environment context.

## D. Additional ablations

We also experiment with different choices of hyperparameters. We only report numbers on training performed on the cooking scenario. The numbers are reported for the validation split, distinct from the testing split, mentioned in Sec. 4. We only report PR-AUC and MPJPE for location prediction and pose prediction, respectively.

**Effect of the observation time $\tau_o$.** Table 2 shows the results. We see that past video observation is crucial for providing the activity context. Thus, not providing any location gives the worst performance. The performance with 30 seconds of past observation is at par with other past observation durations. Therefore, we choose $\tau_o = 30$ so that the model has enough context for interaction prediction.

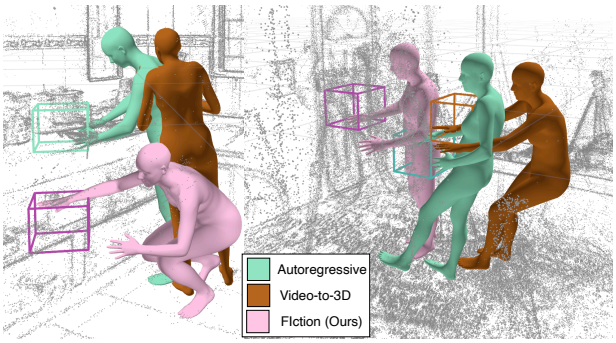


Figure 5. Comparison of our method with baselines and a cooking (left) and a bike-repair (right) take.

| Location prediction | | | | Pose prediction | | | |
|---|---|---|---|---|---|---|---|
| $0s$ | $30s$ | $60s$ | $120s$ | $0s$ | $30s$ | $60s$ | $120s$ |
| 16.0 | **21.2** | 21.0 | **21.2** | 225 | 215 | 213 | **212** |

Table 2. Effect of $\tau_o$ on the performance.

**Effect of the future time $\tau_f$.** Table 3 shows the results. We see an expected trend that the task becomes more difficult as $\tau_f$ increases. However, at a very high $\tau_f$, the interaction location prediction becomes an easier task since the person has navigated to a large part of the environment, thus making majority of the locations as ground truth. Thus, we choose $\tau_f = 180s$ as a challenging version of the future interaction location prediction.

| Location prediction | | | | Pose prediction | | | |
|---|---|---|---|---|---|---|---|
| $60s$ | $120s$ | $180s$ | $600s$ | $60s$ | $120s$ | $180s$ | $600s$ |
| **22.6** | 21.4 | 21.2 | 21.4 | **207** | 212 | 215 | 220 |

Table 3. Effect of $\tau_f$ on the performance.

**Effect of the learning rate.** Table 4 shows the results. We see that the model performs the best with a learning rate of $5 \times 10^{-5}$ for interaction location prediction, and $5 \times 10^{-6}$ for future pose prediction. This same parameter is chosen for all testing, as mentioned in Sec. 3.4.

| Location prediction | | | Pose prediction | | |
|---|---|---|---|---|---|
| $5.10^{-6}$ | $5.10^{-5}$ | $5.10^{-4}$ | $5.10^{-6}$ | $5.10^{-5}$ | $5.10^{-4}$ |
| 20.6 | **21.2** | 19.6 | **215** | 220 | 226 |

Table 4. Effect of learning rate on the performance.

**Effect of the encoder model size.** We use a simple transformer encoder $\mathcal{L}$ for encoding the environment context (Sec. 3.2). We experiment with varying number of transformer layers. We experiment with $2, 4$ and $6$ layers.

Table 5 shows the results. We observe that the number increases with the number of layers. This suggests that the performance can be further improved, with a larger transformer size. We do not experiment beyond 6 due to hardware constraints.

| Location prediction | | | Pose prediction | | |
|---|---|---|---|---|---|
| 2 | 4 | 6 | 2 | 4 | 6 |
| 20.2 | 20.9 | **21.2** | 228 | 222 | **215** |

Table 5. Effect of the model size on the performance. We vary the number of transformer layers.

## E. Limitations

As discussed in Sec. 3, our current method assumed one actor per video. The model design cannot explicitly handle multi-person scenarios. We will handle multi-person scenarios in the future. Nevertheless, the single-actor problem is still challenging with scope for improvement. We also assume a static point cloud when creating the dataset, while in practice, the object location can change with time. It is possible to use 3D information only from the last time segment for improving the spatial input to the model, we do not consider this case for the ease of the I/O. Note that this simplification does not affect the curated dataset quality, since we use narrations from Ego-Exo4D [38] as an additional signal. Finally, we use state-of-the-art methods Detic [118], WHAM [92] and Llama 3.1 [21] for creating the dataset, which are prone to errors. Any future improvement in these domains will further strengthen our dataset quality and the resulting trained model.