Supplementary Material

1. Motion-Related Gaps and Egomotion Homography

In this section, we provide a detailed analysis for filling the motion-related gaps in hand-object interaction (HOI) prediction mentioned in Sec. I of the main text with the egomotion homography. To narrow the view gap between the last observation and the other observations, homography works as a bridge to connect the pixel positions $\mathbf{p}_0$, $\mathbf{p}_t \in \mathbb{R}^2$ of one 3D hand waypoint/contact point on I_t ($t \leq 0$) and I_0, which can be represented by $\mathbf{p}_0 = M_t \mathbf{p}_t$. We let the denoising network be aware of the egomotion features E_t encoded from M_t and enable it to capture the above-mentioned transformation when predicting future hand trajectories and contact points on the last observed image as a canvas.

For the 2D-3D gaps, we first discover the relationship between 2D pixel movements and 3D hand movements. For a 3D point that moves from $\mathbf{P}_t \in \mathbb{R}^3$ in the camera coordinate system at timestamp t ($t \leq 0$) to $\mathbf{P}_0 \in \mathbb{R}^3$ in the camera coordinate system at timestamp $t = 0$, we first project them to the image plane by $\mathbf{p}_t = K\mathbf{P}_t$ and $\mathbf{p}_0 = K\mathbf{P}_0$, where K is the intrinsic parameters. Then we transform $\mathbf{p}_t$ to the last canvas image by $\mathbf{p}'_t = M_t \mathbf{p}_t$. The 2D pixel movement on the last image can be formulated as:

$$\mathbf{p}_0 - \mathbf{p}'_t = K\mathbf{P}_0 - M_t\mathbf{p}_t = K\mathbf{P}_0 - M_t K\mathbf{P}_t.$$

Therefore, the 3D action ($\mathbf{P}_t \rightarrow \mathbf{P}_0$) uniquely corresponds to the 2D pixel movement ($\mathbf{p}_t \rightarrow \mathbf{p}_0$) once K and M_t are both determined. Since K is a constant for each video clip, only M_t changing along the time axis determines the spatial relationship between observations. Therefore, we enable our proposed model aware of egomotion by encoding M_t to a feature vector absorbed by multi-head cross attention of Motion-Aware Denoising Transformer as mentioned in Sec. III-B of the main text, narrowing the gap between 2D pixel movement and 3D actions. Note that we do not utilize SE(3) here due to scale-agnostic estimation with only 2D images as input.

2. Iterative Non-Autoregressive Paradigm vs. Autoregressive Paradigm

Our proposed Diff-IP2D is an iterative non-autoregressive (iter-NAR) model, showing better HOI prediction performance compared to the state-of-the-art methods [6], [8] with the autoregressive (AR) paradigm. AR models reason about the next HOI state only according to the previous steps (Fig. I(a) in the main text), leading to the forward-only constraint. They overlook the backward constraint which we think is also important for HOI prediction. We provide an example in Fig. 2 to further explain the significance of the backward constraint. The human hand generally picks up a cup (Fig. 2(a)) by its handle because side-gripping by the handle is more stable and allows for a faster target approach than other ways. It is more likely for a hand to approach the cup from the side (red arrow in Fig. 2(b)) than from the top (green arrow in Fig. 2(b)) in the near future. Consequently, the final state of the future HOI can be approximately determined, which dictates the hand movement toward the cup, thereby establishing potential backward constraints on *spatial causality*. Therefore, we argue that HOI prediction should be modeled as the non-autoregressive process considering the bidirectional constraints within the holistic sequence, rather than the autoregressive process with only forward constraints on *temporal causality*.

We also provide an illustration comparison between iter-NAR parallel generation and AR generation in Fig. 1. Our proposed iter-NAR paradigm predicts future HOI states in parallel considering bidirectional constraints encompassing both forward and backward constraints within the holistic interaction sequence. It also shifts the limited iterations along the time axis to the sufficient iterations in the diffusion denoising direction (also shown in Fig. IV of the main text). Following the derivation of the previous work DiffuSeq [11] which is used for text generation, here we further mathematically prove that our proposed Diff-IP2D prediction process can be regarded as an iter-NAR process. We first introduce a series of intermediate HOI states $\{\mathbf{F}^y_s\}_{s=0}^S$ decoded from $\{\mathbf{y}_s\}_{s=0}^S$, where $\mathbf{y}_s$ denotes the future part of $\mathbf{z}_s$ and $\mathbf{y}_S \sim \mathcal{N}(0, \mathbf{I})$. $\mathbf{F}^x$ represents the past latent HOI features $F^{\mathrm{R}}_{\mathrm{seq}}$ or $F^{\mathrm{L}}_{\mathrm{seq}}$ from Side-Oriented Fusion Module. $\mathbf{M}$ denotes the egomotion guidance M_{seq} here and will be extended by other perception information in our future work. Therefore, the inference process of our proposed diffusion-based approach can be formulated as follows:

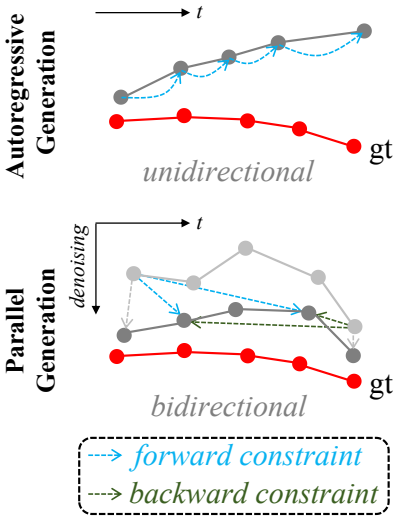


Fig. 1: Autoregressive generation vs. parallel generation.

$$
\begin{aligned}
&p_{\text{Diff-IP2D}}(\mathbf{F}^y|\mathbf{F}^x) \\
&= \sum_{\mathbf{F}^y_S,\ldots,\mathbf{F}^y_1} \int_{\mathbf{y}_S,\ldots,\mathbf{y}_0} p(\mathbf{F}^y|\mathbf{y}_0,\mathbf{F}^x) \prod_{s=S,\ldots,1} p(\mathbf{y}_{s-1}|\mathbf{F}^y_s) p(\mathbf{F}^y_s|\mathbf{y}_s,\mathbf{F}^x,\mathbf{M}) \\
&= \sum_{\mathbf{F}^y_S,\ldots,\mathbf{F}^y_1} \int_{\mathbf{y}_S,\ldots,\mathbf{y}_0} p(\mathbf{F}^y_S|\mathbf{y}_S,\mathbf{F}^x) \prod_{s=S-1,\ldots,0} p(\mathbf{F}^y_s|\mathbf{y}_s,\mathbf{F}^x,\mathbf{M}) p(\mathbf{y}_s|\mathbf{F}^y_{s+1}) \\
&= \sum_{\mathbf{F}^y_S,\ldots,\mathbf{F}^y_1} p(\mathbf{F}^y_S|\mathbf{y}_S,\mathbf{F}^x) \prod_{s=S-1,\ldots,0} \int_{\mathbf{y}_s} p(\mathbf{F}^y_s|\mathbf{y}_s,\mathbf{F}^x,\mathbf{M}) p(\mathbf{y}_s|\mathbf{F}^y_{s+1}).
\end{aligned}
$$

Then we marginalize over $\mathbf{y}$ and obtain the initial iterative non-autoregressive form of our proposed approach:

$$
\begin{aligned}
&p_{\text{Diff-IP2D}}(\mathbf{F}^y|\mathbf{F}^x) \\
&= \sum_{\mathbf{F}^y_S,\ldots,\mathbf{F}^y_1} p(\mathbf{F}^y_S|\mathbf{y}_S,\mathbf{F}^x) \prod_{t=S-1,\ldots,0} p(\mathbf{F}^y_s|\mathbf{F}^y_{s+1},\mathbf{F}^x,\mathbf{M}) \\
&\equiv \sum_{\mathbf{F}^y_1,\ldots,\mathbf{F}^y_{K-1}} p(\mathbf{F}^y_1|\mathbf{F}^x) \prod_{k=1,\ldots,K-1} p(\mathbf{F}^y_{k+1}|\mathbf{F}^y_k,\mathbf{F}^x,\mathbf{M}),
\end{aligned}
$$

where we align the variable s, which denotes the diffusion steps, with the commonly used iteration variable k in typical iterative formulas. Here what we pursue using the denoising diffusion model is to recover implicit features of future HOI states instead of directly decoding the final explicit hand waypoints or contact points. Therefore, we can regard the iterative process (latents→ explicit HOI→ latents) inherent in the above-mentioned equation as an equivariant mapping (latents→ latents). The above equation can be further transformed to the ultimate iter-NAR form of our proposed Diff-IP2D:

$$
\begin{aligned}
&p_{\text{Diff-IP2D}}(\mathbf{y}|\mathbf{F}^x) \\
&= \sum_{\mathbf{y}_1,\ldots,\mathbf{y}_{K-1}} p(\mathbf{y}_1|\mathbf{F}^x) \prod_{k=1,\ldots,K-1} p(\mathbf{y}_{k+1}|\mathbf{y}_k,\mathbf{F}^x,\mathbf{M}) \\
&= \sum_{\mathbf{y}_1,\ldots,\mathbf{y}_{K-1}} \prod_{i=1,\ldots,N_{\mathrm{f}}} p(\mathbf{y}_{1,i}|\mathbf{F}^x) \prod_{k=1,\ldots,K-1} \prod_{i=1,\ldots,N_{\mathrm{f}}} p(\mathbf{y}_{k+1,i}|\mathbf{y}_{k,1:N_{\mathrm{f}}},\mathbf{F}^x,\mathbf{M}).
\end{aligned}
$$

3. Motivation of the Regularization Loss

We propose a regularization term $\mathcal{L}_{\mathrm{reg}}$ in Eq. (4) of the main text for better model optimization. Here we provide more details about the motivation of the regularization loss. In the training process, the tokenizer embeds RGB information and hand-object locations to latent features for the following denoising diffusion process. Here we describe the detailed function of the tokenizer: For each input image, we first exploit a pretrained Temporal Segment Network [27] and extract hand and object RoIAlign [28] features given the detected bounding boxes from [16]. Specifically, the center coordinates of the detected bounding boxes are encoded into the hand and object intermediate features, meaning that the latent features transformed from them encompass the spatial information of hands and objects within each image. After being corrupted to noisy features in the forward process and being denoised in the reverse process, the reconstructed latents are further transformed into locations of future hands and contact points by the predictors, including Hand Trajectory Head and Object Affordance Head. As can be seen, the latents are generated from input HOI states by the tokenizer before the forward process, and are further converted to output HOI states by the predictor after the reverse process.

(a) next-active object (b) possible approaching (c) prediction of Diff-IP2D (d) prediction of USST

Fig. 2: An example to clarify our motivation to propose an iter-NAR paradigm considering bidirectional constraints. The hand waypoints from ground-truth labels and HOI prediction approaches are connected by blue and white dashed lines respectively. Note that we reverse the RGB values of each image to display the arm's positions more clearly. There is a lack of backward constraints in AR-based USST [8], leading to a shorter predicted trajectory (almost curled up into a point) and larger accumulated displacement errors. In contrast, our Diff-IP2D with iter-NAR paradigm is potentially guided by final HOI states, and thus predicts more accurate hand trajectories following both spatial causality and temporal causality.

TABLE 1: Joint evaluation results in the ablation study on egomotion guidance

approach	EK55			EK100		
	SIM* ↑	AUC-J* ↑	NSS* ↑	SIM* ↑	AUC-J* ↑	NSS* ↑
Diff-IP2D*	0.216	0.718	0.842	0.198	0.712	0.778
Diff-IP2D	**0.222**	**0.730**	**0.888**	**0.204**	**0.727**	**0.844**
improvement	2.8%	1.7%	5.5%	3.0%	2.1%	8.5%

Diff-IP2D*: Diff-IP2D w/o egomotion guidance

This is why we regarded latents before and after the denoising diffusion process as representing the same "profile" of the input HOI sequence. They both inherently encompass HOI state information in the same interaction duration, and the training process can be further regarded as the predictor distilling HOI state knowledge from the tokenizer. Therefore, we build a closer gradient connection between the tokenizer and the predictor by introducing the regularization term into the training process to enhance the knowledge distillation. Tab. III in the main text presents the improvement in HOI prediction from our proposed regularization strategy.

4. More Details about Datasets and Diff-IP2D Training Configurations

The training sets of EK55 [35] and EK100 [36] contain 8523 and 24148 video clips respectively. Their test sets consist of 1894 and 3513 samples for hand trajectory evaluation, and 241 and 401 samples for object affordance evaluation. In contrast to Epic-Kitchens, the EG dataset [37] offers a smaller data volume, including 1880 training samples, 442 evaluation hand trajectories, and 69 evaluation interaction hotspots. All the training sets are automatically generated following Liu et al. [6]. Note that we exclusively use the test part of the EG dataset to assess generalization ability in the experiments of Sec. IV-B and Sec. IV-C since it contains insufficient training samples for reasonable convergence.

For training Diff-IP2D, we use AdamW optimizer with a learning rate 2e-4. The total loss function is depicted below. The loss weights are initially set as $\lambda_{\text{VLB}} = 1$, $\lambda_{\text{traj}} = 1$, $\lambda_{\text{aff}} = 0.1$, and $\lambda_{\text{reg}} = 0.2$. All the networks in Diff-IP2D are trained for 30 epochs with a batch size of 8 on 2 A100 GPUs.

$$\mathcal{L}_{\text{total}} = \lambda_{\text{VLB}}(\mathcal{L}^{\text{R}}_{\text{VLB}} + \mathcal{L}^{\text{L}}_{\text{VLB}}) + \lambda_{\text{traj}}(\mathcal{L}^{\text{R}}_{\text{traj}} + \mathcal{L}^{\text{L}}_{\text{traj}}) + \lambda_{\text{aff}}\mathcal{L}_{\text{aff}} + \lambda_{\text{reg}}(\mathcal{L}^{\text{R}}_{\text{reg}} + \mathcal{L}^{\text{L}}_{\text{reg}}).$$

5. Additional Experimental Results

A. Joint Evaluation on the Effect of Egomotion Guidance

We present the supplementary evaluation results in the ablation study on egomotion guidance. Our proposed joint evaluation protocol is applied here to show the positive effect of egomotion guidance for denoising diffusion. As can be seen in Tab. 1, the use of the egomotion features enhances the joint prediction performance of Diff-IP2D on both EK55 and EK100. EK100 has a larger data volume which contains much more human motion patterns than EK55, leading to larger improvement on SIM*, AUC-J*, and NSS* by 3.0%, 2.1%, and 8.5% respectively.

TABLE 2: Comparison of performance on hand trajectory prediction on EgoPAT3D-DT

metric	OCT [6]	USST [8]	Diff-IP2D† (ours)
ADE (seen)	0.108	0.082	**0.076**
FDE (seen)	0.122	0.118	**0.112**
ADE (unseen)	0.091	0.060	**0.055**
FDE (unseen)	0.147	0.087	**0.083**

Final displacement errors of baselines [6], [8] are re-evaluated according to the erratum from Bao *et al.* [8] in their open-source repository.

B. Ablation Study on Observation Time

We use the EK55 dataset to demonstrate the effect of observation time on HOI prediction performance. We present the change of hand trajectory prediction errors with different input sequence lengths $\{2, 4, 6, 8, 10\}$, corresponding to the observation time $\{0.5\,\text{s}, 1.0\,\text{s}, 1.5\,\text{s}, 2.0\,\text{s}, 2.5\,\text{s}\}$. We first use Diff-IP2D trained with 10 observation frames to implement zero-shot prediction with different sequence lengths. Fig. 3(a) illustrates that the prediction performance drops significantly when the number of observation frames decreases. In contrast, once our proposed model is trained from scratch with the predefined observation time, it generates plausible prediction results as Fig. 3(b) shows. Especially when the number of observation frames decreases to 4, our method still outperforms the baseline which is trained from scratch with 10 observation frames. This demonstrates the strong generation ability of our diffusion-based approach with limited conditions.

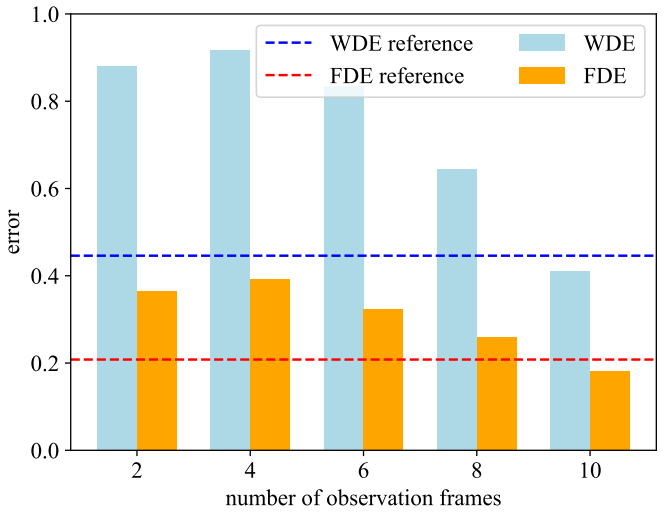
C. Additional Visualization of Object Affordance Prediction on Epic-Kitchens

We additionally illustrate the predicted contact points with average distances to the ground-truth points on frames of Epic-Kitchens. As Fig. 4 shows, our proposed method still outperforms the second-best baseline considering the center of 10 predicted candidates.

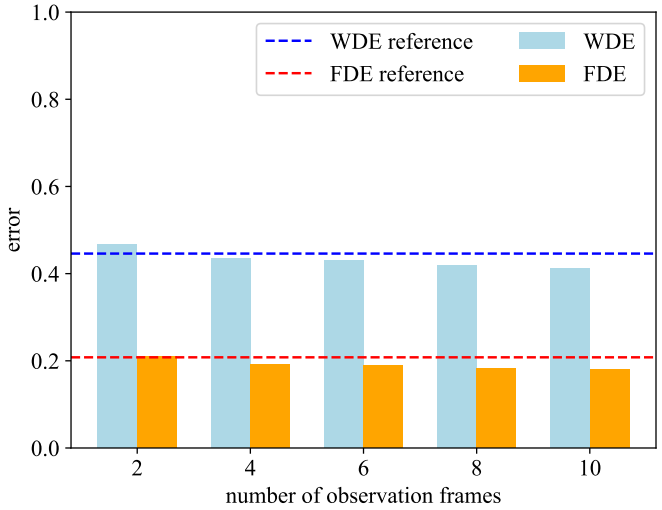
We also provide two cases in which our Diff-IP2D predicts object affordances away from ground truth but more reasonable than the counterparts of the baseline. As Fig. 5 shows, our proposed Diff-IP2D focuses more on "meaningful" parts of objects such as handles even though its prediction has a similar distance away from ground-truth contact points compared to the baseline.

D. Evaluation on EgoPAT3D-DT

We further conduct an additional experiment on a new public dataset EgoPAT3D-DT [8], [38] to compare the performance of our proposed Diff-



(a) Zero-shot prediction



(b) Prediction by models trained from scratch

Fig. 3: Ablation study on observation time. The reference line represents the performance of the second-best baseline trained from scratch using 10 observation frames.

IP2D† and two state-of-the-art baselines, OCT [6] and USST [8]. There is no affordance annotation in EgoPAT3D-DT and thus we only report the results of hand trajectory prediction. Following the previous work [8], we use the fixed ratio 60% to split the past and future sequences at 30 FPS. EgoPAT3D-DT encompasses both seen scenes and unseen scenes, where the unseen scenes are only used for testing. We obtain 6356 training sequences, 846 validation sequences, and 1605 test sequences. As can be seen in Tab. 2, our Diff-IP2D† conducted on the iter-NAR paradigm with temporal enhancement outperforms the AR baselines on hand trajectory prediction on the EgoPAT3D-DT dataset. The better performance of our proposed approach on the unseen test scenes also demonstrates its solid generalization ability.

6. Supplementary Technical Details

How the GT future hand trajectories are obtained for training? How good are they? We follow the GT labels of future hand trajectories from Liu et al. [6]. They use a hand-object detector [16] to extract hand bounding boxes for each future image. Each bounding box center is projected to the last observation frame (canvas image) using estimated homography. The homography matrix between the future image and the canvas image is obtained by multiplying sequential homography matrices. The projected hand locations in the canvas image plane constitute future hand waypoints for training and testing our model. The GT annotations are high quality because: (1) the hand-object detector achieves around 90% IOU on egocentric datasets [16], (2) low-quality GT hand trajectories are manually removed by Liu et al. [6], and (3) we have rechecked the quality of GT hand trajectories in this work.

Fig. 4: Visualization of object affordance prediction grounded on frames of Epic-Kitchens. The ground-truth contact points are represented by red dots. The contact points predicted by our Diff-IP2D with average distances to the ground-truth points are represented by white dots. The counterparts predicted by OCT [6] are represented by blue dots.

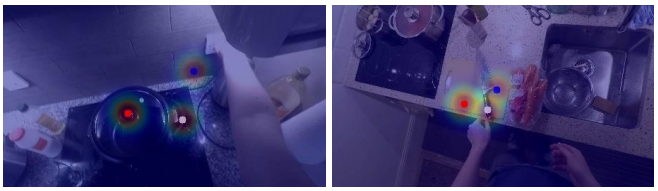


Fig. 5: Two additional explanatory cases.

How does the model handle the situation when only one hand is visible in the frames? The above-mentioned hand-object detector identifies the visibility of each hand, providing a side-aware valid mask for our work. During training and testing, we pad zero values to the output features of SOFM at the timestamps when the hand is invisible according to the valid mask. Besides, the mask is also used by MADT while computing self- and cross-attention. Moreover, Hermite spline interpolation is used to fill the missing GT waypoints caused by invisible hands. If one hand, e.g., the left hand, is absent throughout the entire video clip, Diff-IP2D focuses diffusion denoising on the visible side for higher efficiency, as the latent features from our SOFM are side-aware. In these cases, the invisible side cannot be used for both supervision and error calculation.

How are the ground truth 10 contact points obtained? How good are they? In the training process, we use GT contact points from Liu et al. [6], who exploit skin segmentation and fingertip detection to determine fingertip locations within hand-object bounding boxes. Each training video clip has one GT future contact point per valid side, averaged from detected fingertip locations. Diff-IP2D outputs one possible contact point per valid side after each forward process. For testing, we also follow Liu et al.'s [6] evaluation pipeline and use their high-quality GT object hotspot annotations from Amazon Mechanical Turk and rechecked by us. Each video clip has 1-5 GT points as the "contact center" annotated by workers in the canvas frame to generate GT hotspots, manually avoiding the occlusion problem. To generate object affordance predictions, we perform 10 inference samples per valid side and select the predicted contact point closest to the predicted hand trajectories. These 10 selected points represent sampled estimates of possible next-active object locations, which are used to calculate the object hotspot as affordance.