

I’M HOI: Inertia-aware Monocular Capture of 3D Human-Object Interactions

Supplementary Material

In this supplementary, we commence by providing a comprehensive exposition on the implementation details of our method. Subsequently, we expound upon the calibration of the multi-modal system and the processing of IMU acceleration data during the data collection phase of IMHD². Finally, we present additional qualitative and quantitative results to further validate the efficacy of I’M-HOI, with a particular focus on the evaluation of each regularization term integrated within the category-specific interaction diffusion filter component.

A. Implementation Details

A.1. General Interaction Motion Inference

Network Architecture. We employ a pre-trained ResNet-34 model [17] $f^{\text{enc}} : \mathbb{R}^{H \times W \times 4} \mapsto \mathbb{R}^{8 \times 8 \times 512}$ to extract image features, where $H = W = 256$. Subsequently, we utilize 3 stacked deconvolution layers to construct a feature pyramid, with each layer $f_i^{\text{deconv}} : \mathbb{R}^{H_{\text{in}} \times W_{\text{in}} \times C_{\text{in}}} \mapsto \mathbb{R}^{2H_{\text{in}} \times 2W_{\text{in}} \times C_{\text{out}}}$ receiving input features with resolution $H_{\text{in}} = W_{\text{in}} = 8, 16, 32$ and channel $C_{\text{in}} = 512, 256, 256$ respectively, and producing a 256-channel feature map that is upsampled by a factor of 2. Following each deconvolution layer are Batch Normalization and ReLU activation layers. For every intermediate feature map, a specific regressor $f_i^{\text{reg}} : \mathbb{R}^{H_{\text{in}} \times W_{\text{in}} \times 256} \mapsto \mathbb{R}^{6+3}$ is tailored to embed it to \mathbb{R}^{2000} and concatenates it with $\hat{\mathbf{R}}_o^{(i-1)}, \hat{\mathbf{T}}_o^{(i-1)}$ to predict $\Delta \hat{\mathbf{R}}_o^{(i)}, \Delta \hat{\mathbf{T}}_o^{(i)}$. Each regressor comprises two hidden Linear layers with a dimension of 1024, as well as two output Linear layers that predict delta rotation and translation independently. Dropout layers with a probability of 0.5 are inserted between each pair of consecutive Linear layers.

Training. The proposed $f^{\text{enc}}, \{f_i^{\text{deconv}}\}_{i=1}^3, \{f_i^{\text{reg}}\}_{i=1}^3$ are trained end-to-end with the inverse kinematics layer, supervised by $\mathcal{L} = \mathcal{L}_{\text{kp3d}} + \lambda_{\text{j2d}} \mathcal{L}_{\text{j2d}} + \mathcal{L}_{\text{twist}} + \lambda_{\text{occ-sil}} \mathcal{L}_{\text{occ-sil}} + \lambda_{\text{area}} \mathcal{L}_{\text{area}}$. Particularly, the object-oriented mesh alignment feedback loss $\mathcal{L}_{\text{maf}} = \lambda_{\text{occ-sil}} \mathcal{L}_{\text{occ-sil}} + \lambda_{\text{area}} \mathcal{L}_{\text{area}}$ is added after 55 training epochs. The loss weights are: $\lambda_{\text{j2d}} = 1 \times 10^{-9}, \lambda_{\text{occ-sil}} = 1 \times 10^{-6}, \lambda_{\text{area}} = 2 \times 10^{-7}$. The model is trained for 190 epochs on 6 NVIDIA GeForce RTX 3090 GPUs. In each epoch, we randomly sample one from 8 images to train. The training batch size is set to 8.

Optimization. The optimization energy function defined as $\mathcal{E} = w_{\text{visual}} \mathcal{E}_{\text{visual}} + w_{\text{imu}} \mathcal{E}_{\text{imu}}$ is configured with: $w_{\text{visual}} = 20, w_{\text{imu}} = 1 \times 10^5$. We set the learning rate during optimization to 0.01 for 30-fps data (BEHAVE [4], InterCap [24] and CHAIRS [26]) and 5×10^{-4} for 60-fps data (IMHD² and HODome [102]).

A.2. Category-specific Motion Diffusion Filter

Network Architecture. We employ 4 transformer encoder-only layers, each equipped with 4 attention heads, to learn category-specific human-object interaction manifold. The model dimension $D_{\text{model}} = 1024$ and the key, value dimension $D_{\text{key}} = D_{\text{value}} = 512$. We take $N = 1000$ steps and sinusoidal positional encoding function during denoising phase. In contrast to the methodology outlined in [44], where the condition is exactly a part of the target motion, we leverage outcomes from the preceding stage alongside raw IMU measurements as conditions to model the transition from the predictive distribution to the authentic manifold.

Training. We initially warm up the diffusion model solely on our complete training dataset using simple objective function for 100 epochs. Subsequently, we proceed to train the model on category-specific data, incorporating specially designed regularization terms $\mathcal{L}_{\text{consist}}, \mathcal{L}_{\text{vel}}$ and \mathcal{L}_{imu} to implicitly model distinct interaction patterns. The regularization term weights are $\lambda_{\text{off}} = \lambda_{\text{vel}} = \lambda_{\text{consist}} = 1, \lambda_{\text{imu}} = 100$. More detailed, we apply \mathcal{L}_{off} and \mathcal{L}_{vel} for 35 epochs before adding $\mathcal{L}_{\text{consist}}$ and \mathcal{L}_{imu} . To enhance the generation results, we maintain an exponential moving average (EMA) version of the model throughout training, updating it every 10 epochs with a decay rate of 0.995. Additionally, we leverage Automatic Mixed Precision (AMP) to accelerate the training procedure. The model is trained for 55 epochs on a single NVIDIA GeForce RTX 3090 GPU, with the training batch size set to 128.

B. Data Preparation Details

B.1. System Calibration of IMHD²

Temporal Synchronization. In order to synchronize RGB data with IMU measurements, we instructed the performer to wear an additional IMU sensor on the ankle area and execute a takeoff motion at the onset of each interaction segment. By detecting the point at which the performer falls to the ground based on the gravitational acceleration mutation in the IMU signals, we automatically pinpointed this moment as the starting frame and manually annotated it within the RGB sequences.

Spatial Alignment. To mitigate spatial misalignment between camera and IMU, we conducted spatial alignment once per ten minutes. Specifically, in our multi-modal and multi-sensor system, there exists multiple coordinate frames, including $\{\mathcal{F}_{C_i}\}_{i=0}^{31}$ for cameras, \mathcal{F}_W for world and \mathcal{F}_I for inertia. Since the transformation $\mathcal{T}_{W \rightarrow C_i} \in SO(3)$ from


	Holdhandle Hit Holdhead Hit Lefthand Swing Midpart Rotate Pickup Putdown Righthand Swing Rub Throw Catch Twoends Rotate Twohands Swing		suitcase	Lefthand Carry Lefthand Push Lift Putdown Pickup Ride Play Righthand Carry Righthand Push Twohands Carry Twohands Push Twohands Pull		Ollie Kickflip Grind Manual Heelflip Pop Shove-it Nollie Varial Kickflip McTwist Darkslide				
	Left Biceps Left Lunges Left Triceps Right Biceps Right Lunges Right Triceps		kettlebell	Forward Swing Backward Swing Snatch Turkish Get-up Goblet Squat Windmill		tennis racket	Forehand Backhand Volley Overhead Smash Slice Drop Shot		pan	Hold Stir Shake Flip
	golf club		chair	Drive Putt Chip Pitch Sand Shot Fade Hook Draw Grip Slice		broom	Sit Lean Adjust Swivel Recline Rest Clean Lift Rock Kick		Sweep Push Pull Twist Store Tap Tilt Lift Grip Maintain	

Table 5. IMHD² collects 10 distinct objects along with a range of interaction motions associated with each object.

\mathcal{F}_W to \mathcal{F}_{C_i} is easy to obtain through off-the-shelf multi-camera calibration toolbox, our goal is to calibrate the transformation $\mathcal{T}_{I \rightarrow W} \in \mathcal{SO}(3)$ from \mathcal{F}_I to \mathcal{F}_W .

In our implementation, we capture the global orientation $\{\mathbf{R}_t^W \in \mathcal{SO}(3)\}_{t=0}^{T-1}$ of the performer who circles around in \mathcal{F}_W by [1]. The inertial rotation measurement $\{\mathbf{R}_t^I \in \mathcal{SO}(3)\}_{t=0}^{T-1}$ in \mathcal{F}_I is simultaneously recorded by an IMU sensor positioned at the waist area. Suppose the IMU sensor is relatively fixed to the performer, we can construct the following equation:

$$\mathcal{T}_{I \rightarrow W} \mathbf{R}_t^I (\mathcal{T}_{I \rightarrow W} \mathbf{R}_{t+s}^I)^{-1} = \mathbf{R}_t^W (\mathbf{R}_{t+s}^W)^{-1}, \quad (14)$$

where $s = 5$ is the stride. Let $\mathbf{B}_t = \mathbf{R}_t^I (\mathbf{R}_{t+s}^I)^{-1}$ and $\mathbf{A}_t = -\mathbf{R}_t^W (\mathbf{R}_{t+s}^W)^{-1}$, we can reformulate Equation 14 as:

$$\mathbf{A}_t \mathcal{T}_{I \rightarrow W} + \mathcal{T}_{I \rightarrow W} \mathbf{B}_t = \mathbf{0}, \quad (15)$$

which is a Sylvester equation. To solve this equation, both analytical [56] and iterative optimization methods [34] can be used.

B.2. Acceleration Data Processing

Normalization on Real Data. Given the assumption that all objects are rigid and possess uniform rotational inertia,

with their centroids equivalent to their geometry centers, practical constraints arise when attempting to mount the IMU sensor precisely onto these centers, which may lie within the object. Consequently, extraneous linear acceleration may arise even from pure rotational motion, introducing undesirable noise. To eliminate such disturbances, we initially fix a mounting point for each object and manually measure the directional offset \vec{r} from the center to that point using mesh processing software [9]. By leveraging recorded angular velocity $\vec{\omega}$, the additional linear velocity stemming from rotation is $\vec{v} = \vec{\omega} \times \vec{r}$, and we can calculate the excess linear acceleration: $\delta \mathbf{a}_t = \frac{\mathbf{v}_t - \mathbf{v}_{t-\Delta t}}{\Delta t}$. Finally, the normalized acceleration data can be attained by subtracting $\delta \mathbf{a}_t$ from the raw measurements.

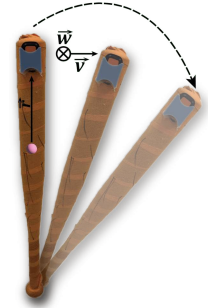


Figure 8. Illustration of why extra linear acceleration occurs.

Simulation on Synthetic Data. Furthermore, we simulate synthetic IMU data based on ground-truth object motion annotations of [4, 24, 26, 102]. In particular, to derive inertial

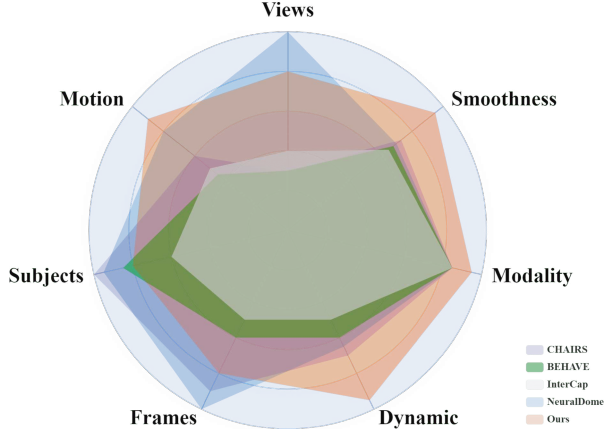


Figure 9. Attributes comparison between different datasets.

acceleration data, we follow [47, 67, 93] to calculate the second-order difference of object translation:

$$\mathbf{a}_t = \frac{\mathbf{T}_{o,t-n} + \mathbf{T}_{o,t+n} - 2\mathbf{T}_{o,t}}{(n\tau)^2}, \quad (16)$$

where $n = 4$ is the smoothing factor to enhance the approximation to actual acceleration, and $\tau = \frac{1}{\text{fps}}$ represents the time interval between consecutive frames.

B.3. Dataset Statistics

We present a comprehensive overview of the contents of IMHD² in Table 5. It reveals that, for each object, we curated a wide array of interaction patterns involving different human body segments. Complementary to existing datasets characterized by numerous participants, extensive recording frames and super dense views, Figure 9 illustrates that IMHD² offers a more challenging, diverse and quality collection of motion data focusing on object-oriented interactions. Specifically, measured through metrics such as average motion velocity and jitter, IMHD² encompasses more dynamic interaction motions with better smoothness. Moreover, IMU data is concurrently collected alongside RGB images, serving not only to align with ground-truth annotations, but also as network input to enhance accuracy and efficiency in motion capture.

C. More Experiments

C.1. More Results

In preceding sections, we have demonstrated the robustness of I^m-HOI under severe occlusions. Expanding on this, we now present sequential capture results of I^m-HOI to showcase spatial-temporal coherence. Figure 10 illustrates that our approach captures accurate and consistent human-object spatial arrangements within a temporal context, which

Regularization terms	CD (per-frame)		CD (10s)	
	simpl	object	simpl	object
w/o \mathcal{L}_{off}	7.07	7.31	7.59	9.17
w/o \mathcal{L}_{vel}	7.06	7.62	7.62	10.66
w/o $\mathcal{L}_{\text{consist}}$	6.29	6.98	6.30	9.01
w/o \mathcal{L}_{imu}	7.10	7.61	7.87	10.94
Ours	6.50	6.93	5.36	8.53

Table 6. Quantitative evaluations on regularization terms.

validates that our proposed network learns reasonable interaction distributions and recognizes continuous interaction behaviors from input data featuring a hybrid modality.

C.2. More Comparisons

We also present additional qualitative comparisons of sequential capture results with baselines in Figure 11 and Figure 12. It can be observed that even within an extremely short time interval (approximately 0.07 seconds), the image-based baselines [85, 104] exhibit jittery object tracking results, focusing on static interactions while disregarding temporal information. Conversely, the video-based method [86] yields temporally consistent but erroneous predictions without inertial measurements, particularly evident in tracking object rotational motions. In stark contrast, I^m-HOI makes use of both visual cues and IMU signals, cooperating with the design of object-oriented mesh alignment feedback and category-specific interaction prior. This combination contributes significantly to achieving consistent and correct results.

C.3. Ablation on Regularization Terms

To further evaluate the effectiveness of the regularization terms in training of interaction diffusion filter, we conduct a comparative analysis of the full model against downgraded versions that exclude individual terms. As reported in Table 6, the inclusion of \mathcal{L}_{off} restricts objects to a more specific and precise region. $\mathcal{L}_{\text{consist}}$ enforces predicted joint rotations to align with detected 3D joints after forward kinematics, which prevents overfitting to pseudo ground-truth annotations. Both \mathcal{L}_{vel} and \mathcal{L}_{imu} contribute to improve performance in the temporal domain. However, only applying \mathcal{L}_{vel} may lead to oversmooth results due to the loss of physical dynamics. Incorporating second-order supervision \mathcal{L}_{imu} is verified beneficial, not only for smooth results but also for capturing physically plausible interaction motions.

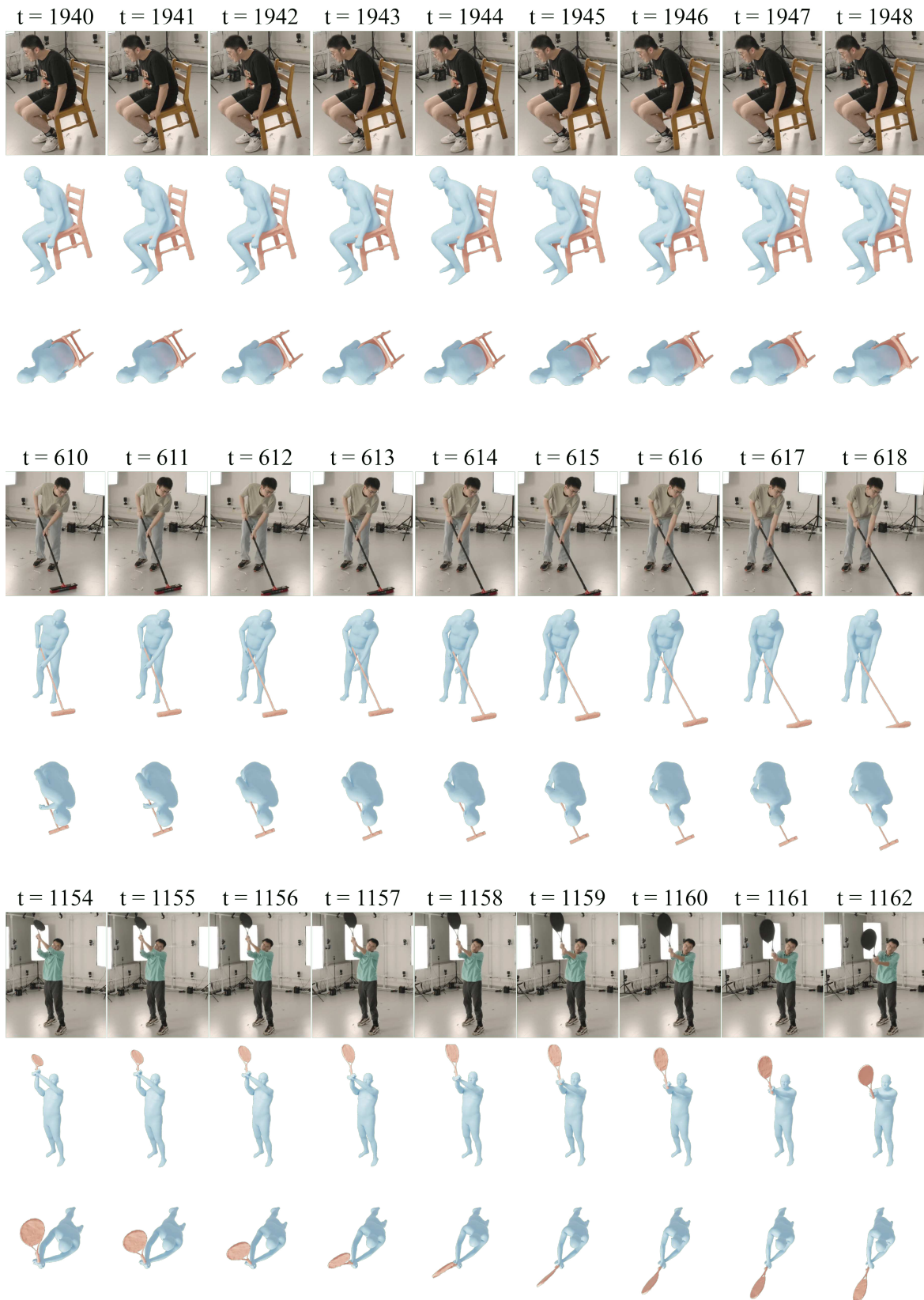


Figure 10. Additional qualitative results of Γ m-HOI on IMHD². We present sequential RGB images, captured motion from camera view and top-view visualizations.

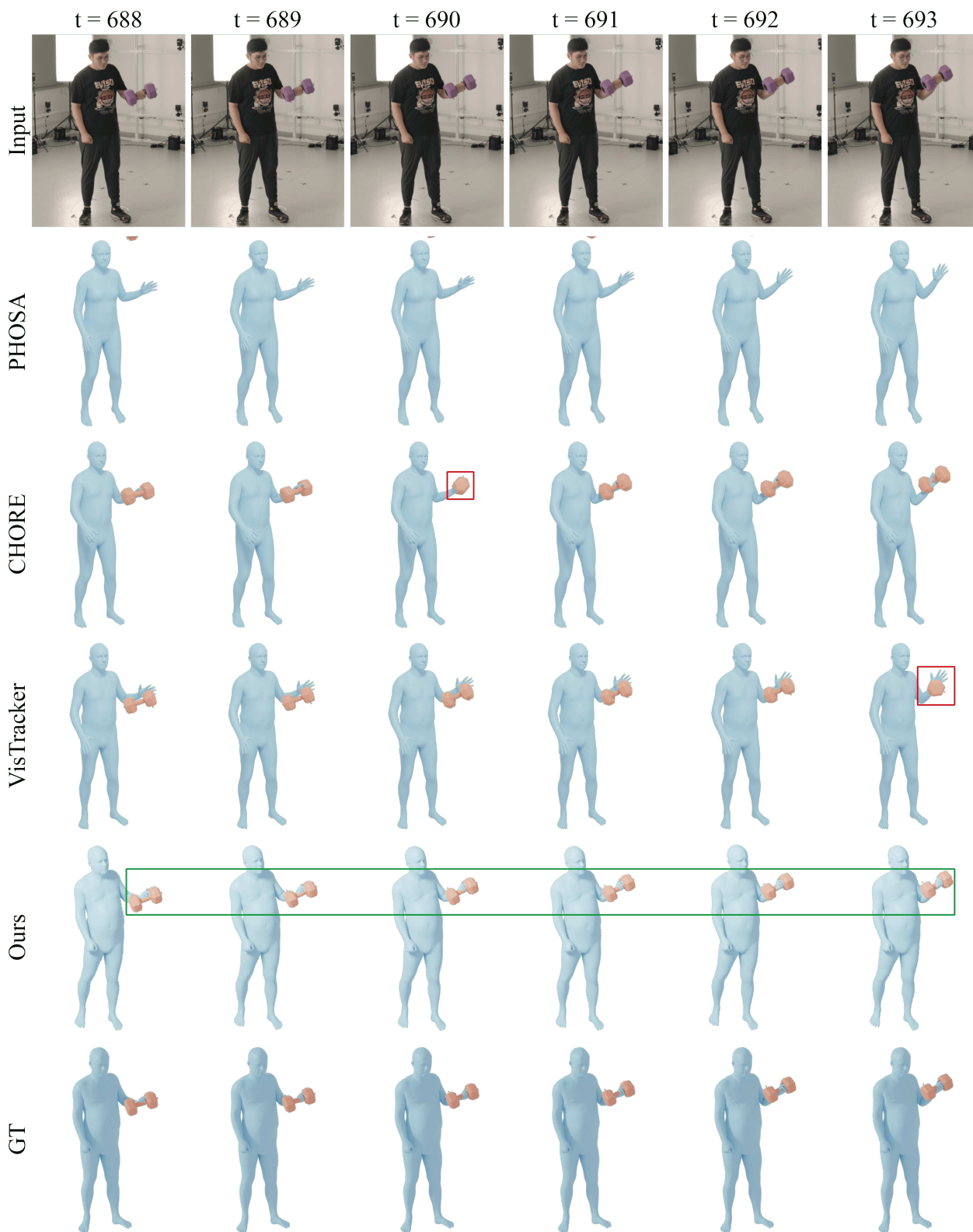


Figure 11. Additional qualitative comparisons. I^m-HOI outperforms baselines on sequential data.

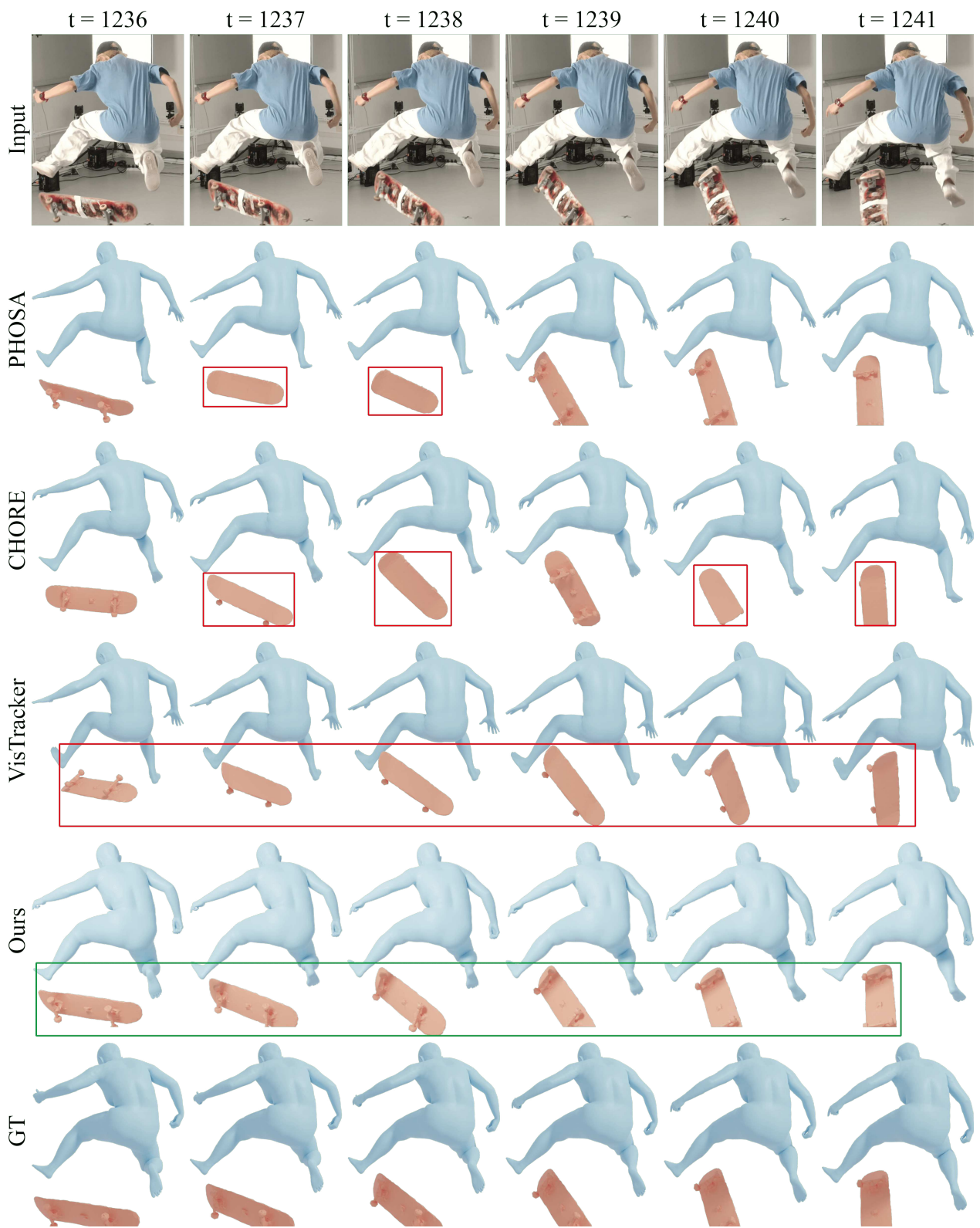


Figure 12. Additional qualitative comparisons. I^m-HOI outperforms baselines on sequential data.