WHOLE: World-Grounded Hand-Object Lifted from Egocentric Videos

Supplementary Material

In supplementary material, we provide further details on implementing network, full VLM prompt, and evaluation metrics (Sec. A). We also visualize more comparisons and results in supplementary videos.

A. Implementation Details

Network Architecture. We use a 4-layer transformer decoder with 4 attention heads for hand-to-object diffusion. The network is non-autoregressive, processing sequences jointly following [6], with 12.35M parameters. The diffusion variable x is a 73-dimensional vector comprising: a 9D object state, a 2D bimanual contact indicator, and bimanual hand representations (2×31D)—global orientation (3), translation (3), pose PCA (15), and shape parameters (10). All inputs are projected to a consistent latent dimension of 512. The network is trained for 1,000,000 iterations using AdamW at a learning rate of 2×10^{-4}. To reduce overfitting, we augment the object template by sampling a random canonical pose applying a random rotation and a small translation jitter—for each training window.

Training Loss. In addition to the DDPM loss, we introduce auxiliary objectives to enhance realism. (1) *Interaction Loss* ($\mathcal{L}_{\text{inter}}$) encourages realistic contact between the predicted hand-object motions and contact labels. It penalizes hand-object distances when contact is predicted and enforces near-rigid transport of contact points across consecutive contact frames [3, 4]. Specifically, for each hand joint, we find its nearest object point p^i, rotate it by the object's relative motion, and penalize deviation from its counterpart p^{i+1}, *i.e.* $\|\boldsymbol{R}^{i+1}(\boldsymbol{R}^i)^T p^i - p^{i+1}\|$, where $\boldsymbol{R}$ is the object rotation. (2) *Consistency Loss* ($\mathcal{L}_{\text{const}}$) promotes agreement among hand features before and after MANO forward kinematics, $\|\boldsymbol{J}_\psi - \text{MANO}(\boldsymbol{\Gamma}_\psi, \boldsymbol{\Lambda}_\psi, \boldsymbol{\Theta}_\psi)\|_2$. (3) *Temporal Smoothness* ($\mathcal{L}_{\text{smooth}}$) further penalizes large accelerations.

Running Time. On a single NVIDIA RTX 6000 Blackwell GPU, our model processes a 150-frame clip in an average of 59.34 seconds. The inference time is dominated by the guidance step (59.06s), with the diffusion step requiring only 0.28s. This represents an orders-of-magnitude speedup over prior works such as [1] (30 hours) and [8] (1 hour). The peak memory footprint is 14GB. VLM queries take 18.6s on average per image with GPT-5.

VLM Prompt. We prompt a VLM to label contact info, with additnoal in-context-learning exmaples. Full prompts are illustrated in Table in this appendix.

Evaluation Metrics. All metrics are computed on 150-frame clips, which correspond to the original sequence length in HOT3D-CLIP, in contrast to prior work [10], which typically evaluates on shorter 60–100 frame segments taken from the middle of the videos.

To compute W/WA-MPJPE, we align the predicted trajectory to the ground truth using an affine transformation (scale, rotation, translation) estimated from selected keypoints. WA-MPJPE uses all joints from all frames, while W-MPJPE uses only the joints from the first two frames. Although trajectory error could be computed without alignment given ground-truth cameras, we follow the standard alignment protocol used in prior work [5, 9, 10].

For ADD/ADD-S of objects, we align predictions using the ground-truth camera poses. For HOI ADD/ADD-S, we first globally align the hand trajectory (as in WA-MPJPE) and then evaluate object error in this aligned space. The usual AUC [2, 7] threshold of 0.1 is overly strict for egocentric HOI due to severe occlusion, truncation, and out-of-view frames, leading to saturated low scores. We therefore use a more permissive threshold of 0.3 to obtain a more informative evaluation.

System Instruction

```
You are a precise visual classifier for hand-object contact detection in
cluttered scenes.

CRITICAL CONSTRAINTS:
1. Each hand (left/right) can be in contact with AT MOST ONE object at a time.
2. "In contact" means direct physical touch: grasping, holding, pressing, or
   any visible contact.
3. If a hand is not clearly touching any object, you must mark all objects as 0
   for that hand.
```

User Prompt Template

```
Analyze this image for hand{object contact (actual touching, not just reaching).
VISUAL GUIDANCE:
The image has been annotated with colored masks:
- GREEN dot = Left hand
- RED dot = Right hand
- Other COLORED masks = Candidate objects (each object has a unique color)
CANDIDATE OBJECTS (in order):
1. obj1
2. obj2
3. obj3
...
STRICT DEFINITION OF CONTACT:
For this task, contact means clear physical touching in this frame only.
Contact (label = 1) requires BOTH:
1. Mask intersection:
   - The hand mask and the object mask share some pixels or directly overlap
     at the boundary (no visible gap).
2. Touching region:
   - The overlap is at a plausible touching area (finger tips, fingers, palm,
     side of hand) on the visible surface of the object.
NO Contact (label = 0) in all of these cases:
- The hand is reaching toward, hovering above, or very close to an object with
  a visible gap between masks.
- The hand is aligned in depth (e.g., above or behind the object) but the masks
  do not intersect.
- The hand is in a pose that suggests future contact, but there is no current
  touching in this single frame.
- There is only a tiny, ambiguous intersection (1{2 pixels) that could be noise
  or occlusion. In such uncertain cases, choose 0 (no contact).
IMPORTANT:
- **Reaching or hovering is NOT contact.**
- **If you are unsure whether contact is happening, choose 0 (no contact).**
CONSTRAINTS (VALIDATION CHECK):
- Each hand can touch AT MOST ONE object.
  - Sum of left across all objects must be <= 1.
  - Sum of right across all objects must be <= 1.
- If a hand is not clearly touching any object, it should have 0 for all objects.
OUTPUT FORMAT:
Return only a JSON object in this exact format (no extra text):
{
  "obj1": {"left": 0, "right": 1},
  "obj2": {"left": 0, "right": 0},
  "obj3": {"left": 1, "right": 0}
```

```
}
Where:
- 1 = the specified hand is clearly touching that object in this frame.
```

References

[1] Zicong Fan, Maria Parelli, Maria Eleni Kadoglou, Xu Chen, Muhammed Kocabas, Michael J Black, and Otmar Hilliges. Hold: Category-agnostic 3d reconstruction of interacting hands and objects from video. In *CVPR*, 2024. 1

[2] Tomáš Hodaň, Martin Sundermeyer, Bertram Drost, Yann Labbé, Eric Brachmann, Frank Michel, Carsten Rother, and Jiří Matas. Bop challenge 2020 on 6d object localization. In *ECCV*, 2020. 1

[3] Jiaman Li, Jiajun Wu, and C Karen Liu. Object motion guided human motion synthesis. *TOG*, 2023. 1

[4] Jiaman Li, Alexander Clegg, Roozbeh Mottaghi, Jiajun Wu, Xavier Puig, and C Karen Liu. Controllable human-object interaction synthesis. In *ECCV*, 2024. 1

[5] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J. Black. WHAM: Reconstructing world-grounded humans with accurate 3D motion. In *CVPR*, 2024. 1

[6] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *ICLR*, 2023. 1

[7] Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. Foundationpose: Unified 6d pose estimation and tracking of novel objects. In *CVPR*, 2024. 1

[8] Yufei Ye, Abhinav Gupta, Kris Kitani, and Shubham Tulsiani. G-HOP: Generative hand-object prior for interaction reconstruction and grasp synthesis. In *CVPR*, 2024. 1

[9] Paper Authors Your. Slahmr: Simultaneous localization and human mesh recovery. In *CVPR*, 2023. 1

[10] Jinglei Zhang, Jiankang Deng, Chao Ma, and Rolandos Alexandros Potamias. Hawor: World-space hand motion reconstruction from egocentric videos. *CVPR*, 2025. 1