HOMimic: Distilling Manipulation Trajectories from Human Videos via Multi-Stage Interaction Reasoning and Taxonomy-Aware Retargeting

Zicheng Guo*, Yitian Shi*†, Yu Hu*, Zhengqi Han, Di Wen, Rania Rayyes

***Abstract*— Transferring manipulation skills from human video demonstrations offers a scalable alternative to costly teleoperation and kinesthetic teaching. However, reliable transfer to robots with parallel-jaw grippers remains challenging due to the morphological and kinematic gap between human hands and robotic end-effectors. Existing grasp-transfer abstractions often rely on sparse hand-object interaction (HOI) cues, while object-centric pipelines typically depend on explicit geometric priors, limiting their robustness to unseen objects, occlusions, and long-horizon interactions. Therefore, we present HOMimic, a hand-centric framework that distills manipulation structure from human videos and transfers it into taxonomy-aware, executable parallel-jaw gripper trajectories. Starting from recovered 3D wrist motions and MANO hand poses, HOMimic introduces a coarse-to-fine Semantic Keyframe Recognition strategy that combines wrist-speed-based temporal proposals with vision language model (VLM)-based semantic verification to identify task-relevant contact segments. For cross-embodiment transfer, HOMimic employs an affordance-centric, taxonomy-aware grasp retargeting formulation that infers multi-modal parallel-jaw grasps from hand parameterizations and visual interaction cues, without requiring explicit object geometry.**

I. INTRODUCTION

Transferring manipulation from human videos to robot-executable motion has emerged as a promising approach to alleviating the cost and scalability bottlenecks of teleoperation and kinesthetic teaching. Due to their ease of collection and rich hand-object interaction (HOI) cues, human videos have increasingly been explored as a source of robot-executable trajectories [1], [2], [3], [4], [5], [6].

Despite this promise, extracting transferable manipulation structure from human videos remains challenging. For instance, the morphology and kinematic mismatch between the human hand and a parallel-jaw (PJ) end-effector makes direct hand-to-gripper transfer nontrivial. Existing grasp transfer abstractions are often overly coarse, relying on sparse fingertip cues or thumb-index pinching geometry [7], [8], [9], [10], [11]. Such approximations collapse diverse human grasp types and neglect whole-hand, contact-dependent grasp intent, limiting cross-embodiment generalization.

Besides, many approaches directly learn robot actions or motion policies from video dynamics [12], [13], [14], [15], [16], favoring observational imitation over explicit understanding of transferable contact phases and HOI patterns, which often limits generalization in diverse or long-horizon demonstrations. Although some works study keypoint extraction or temporal interaction localization [3], [17], for instance, YOTO [3] primarily compresses noisy hand motion for downstream policy learning, whereas EgoLoc [17] assumes simple interaction clips and is less reliable for multi-stage or repetitive manipulation. Moreover, many pipelines remain object-centric and rely on geometric priors and their pose estimation results [18], [19], [20], [21], limiting applicability under occlusion and unseen objects [22].

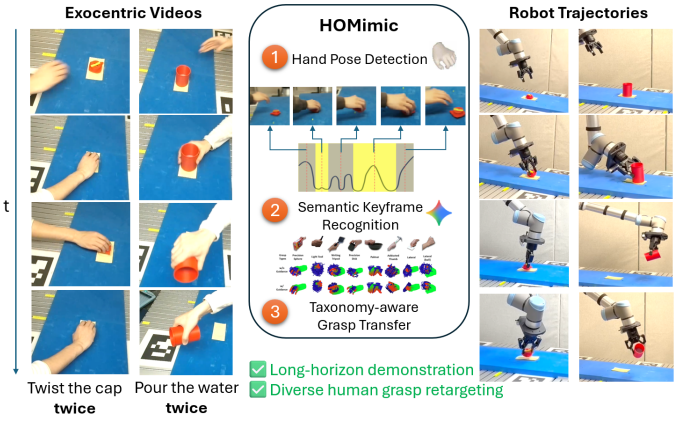

Fig. 1: HOMimic, a manipulation retargeting framework which may transfer human demonstrations with (i) diverse grasp types (ii) multi-stage in-/off-contact keyframes to executable robot trajectories

Therefore, we introduce HOMimic, a hand-centric framework for transferring manipulation from human videos to structured, taxonomy-aware and executable PJ trajectories. Starting from recovered hand pose and wrist motion by WiLoR [23], HOMimic identifies task-relevant contact segments leveraging the semantic understanding from vision-language models (VLMs). Then, we apply a taxonomy-aware retargeting module, instantiated with HOGraspFlow [24], to infer multi-modal taxonomy-aware PJ grasps from hand parametrizations and semantic cues. Finally, the transferred grasps on each keyframe are connected and smoothed, constructing executable PJ trajectories.

In summary, our main contributions are threefold. First, we present an interaction-aware framework for cross-embodiment manipulation transfer from human videos to executable PJ gripper trajectories. Second, we introduce a coarse-to-fine contact segmentation strategy that combines motion cues with VLM-based semantic reasoning to identify transferable contact segments. Third, we incorporate an affordance-centric, taxonomy-aware grasp transfer formu-

Karlsruhe Institute of Technology
*Equal contribution; †Corresponding author.
This work is supported by the German Federal Ministry of Research, Technology, and Space (BMFTR) under the Robotics Institute Germany (RIG), the DFG SFB-1574-471687386 project, and the Ministry of Science, Research and Arts of the Federal State of Baden-Württemberg within the InnovationCampus Future Mobility.

lation that preserves whole-hand grasp intent and without relying on explicit object geometry.

II. Related Works

A. Transferring manipulation from human videos

Transferring robot manipulation from human videos has emerged as a scalable alternative to costly robot data collection. Prior works learn robot behaviors from human videos either by directly training policies from video demonstrations or by introducing structured intermediates such as hand poses, keypoints, motion tracks, or contact cues [3], [7], [17], [25], [26], [27], [28], [29], [30], [31], [32], [33], [34], [35]. While these methods show that human videos provide rich supervision for manipulation transferring, they mainly focus on policy learning or coarse motion abstraction, rather than explicitly identifying transferable contact segments and distilling contact-aware manipulation structure from videos.

B. Cross-embodiment transfer and grasp retargeting

Another related direction studies how to bridge the gap between human demonstrations and robot execution. Existing methods do so by retargeting human hand motion to robot actions, aligning human observations with robot views, or transforming human videos into robot-compatible training inputs [3], [25], [28], [29], [36], [2], [9], [37], [38], [39], [40], [41], [42], [43], [44], [45]. In parallel, task-oriented grasping and affordance-aware methods study grasp transfer and HOI-aware grasp generation [8], [46], [47], [48], [49], [50]. However, many grasp synthesis pipelines remain object-centric and rely on explicit geometric priors such as meshes, point clouds, contact geometry, or object pose estimation [18], [19], [20], [21], [51], [52], which can limit robustness in unconstrained human videos.

III. Methodology

We address the task of extracting manipulation skills from human videos and transferring them to a robot equipped with a PJ gripper. Unlike prior pipelines that rely on explicit geometric priors on target objects, our hand-centric approach is designed to achieve high-fidelity extraction and cross-embodiment retargeting of human-object interactions through an automated pipeline.

Figure 2 summarizes the overall pipeline of our approach, which consists of three stages. Given video demonstrations, **Hand Pose Estimation** (Sec. III-A) first recovers the 3D wrist poses in the robot base frame together with the MANO hand parameters [53] using WiLoR [23] and stereo triangulation [54]. **Semantic Keyframe Recognition** (Sec. III-B) then identifies task-relevant contact segments from the extracted wrist trajectory through a two-stage query scheme. Finally, **Cross-Embodiment Trajectory** (Sec. III-C) invokes *HOGraspFlow* [24] at the start frame of each contact segment to retarget the human grasp to a PJ gripper, propagates the resulting grasp along the localized wrist trajectory, and outputs a smooth executable gripper trajectory.

A. Hand Pose Estimation

Given a stereo video sequence $\mathcal{V} = \{(I_t^1, I_t^2)\}_{t=1}^T$, for each frame t, the **Hand Pose Estimation** module recovers the wrist pose $M_t = (\omega_t, q_t)$ and the MANO hand parameters $\mathcal{H}_t = (\theta_t, \beta_t)$. Here, $\omega_t \in \mathbb{R}^3$ denotes the wrist orientation in axis-angle form, $q_t \in \mathbb{R}^3$ the 3D wrist position in the robot base frame, $\theta_t \in \mathbb{R}^{48}$ the MANO pose parameters whose first 3 dimensions correspond to the global wrist orientation and remaining 45 dimensions to the 15 finger joint rotations, and $\beta_t \in \mathbb{R}^{10}$ the MANO shape parameters. Following Fig. 2, the estimator consists of three steps: monocular hand reconstruction, wrist localization, and rotation averaging for multi-view fusion.

In the first step, for each view $n \in \{1, 2\}$, WiLoR [23] is used to estimate the MANO parameters (θ^n, β^n) from the input image I_t^n. The corresponding 2D joint observations from the stereo pair are then triangulated by Direct Linear Transform (DLT) to recover the 3D hand joints, and the wrist position q_t is obtained as the reconstructed root joint.

After that, multi-view fusion is performed by rotation averaging. Let g_t^1 and g_t^2 denote the quaternion converted from the two wrist orientations q_t^1 and q_t^2. The fused wrist orientation is computed as:

$$\tilde{g}_t = \frac{g_t^1 + g_t^2}{||g_t^1 + g_t^2||} \tag{1}$$

Finally, combining the triangulated position and fused orientation yields the wrist pose $M_t = (\omega_t, q_t)$, and together with the MANO parameters $\mathcal{H}_t = (\theta_t, \beta_t)$, which form the complete hand representation for each frame.

B. Semantic Keyframe Recognition

Given the frame-wise wrist poses M and the corresponding hand orientation parameters θ, a coarse wrist trajectory will be constructed to localize the contact segments C. These segments correspond to task-relevant HOI phases, which serve as a structured basis for subsequent grasp retargeting and trajectory transfer.

To improve robustness under and viewpoint variations, we introduce a **Two-Stage Query** scheme that combines motion analysis with visual semantic reasoning: the first stage (i.e. **Keyframe Detection**) generates candidate intervals from the wrist trajectory, and the second stage (i.e. **In-Contact Detection**) refines them using hand-object spatial cues.

Based on consecutive wrist positions, a speed profile v_t is constructed from M_t to characterize the motion magnitude of the wrist: $v_t = \|q_t - q_{t-1}\|$.

As illustrated in Fig. 2 (right), the **Keyframe Detection** provides the speed profile and an analysis prompt to the VLM[1], to detect low-speed regions and produce valley intervals Z_k. Then, the valid trajectory is partitioned into a sequence of temporal segments J_k, consisting of both the candidate valley intervals Z_k (e.g. the yellow interval in the speed profile of Fig. 2) and the non-valley intervals between them (e.g. the grey interval in the speed profile of Fig. 2), so

[1]We use Gemini 2.5 Flash [55] for prompting in this work

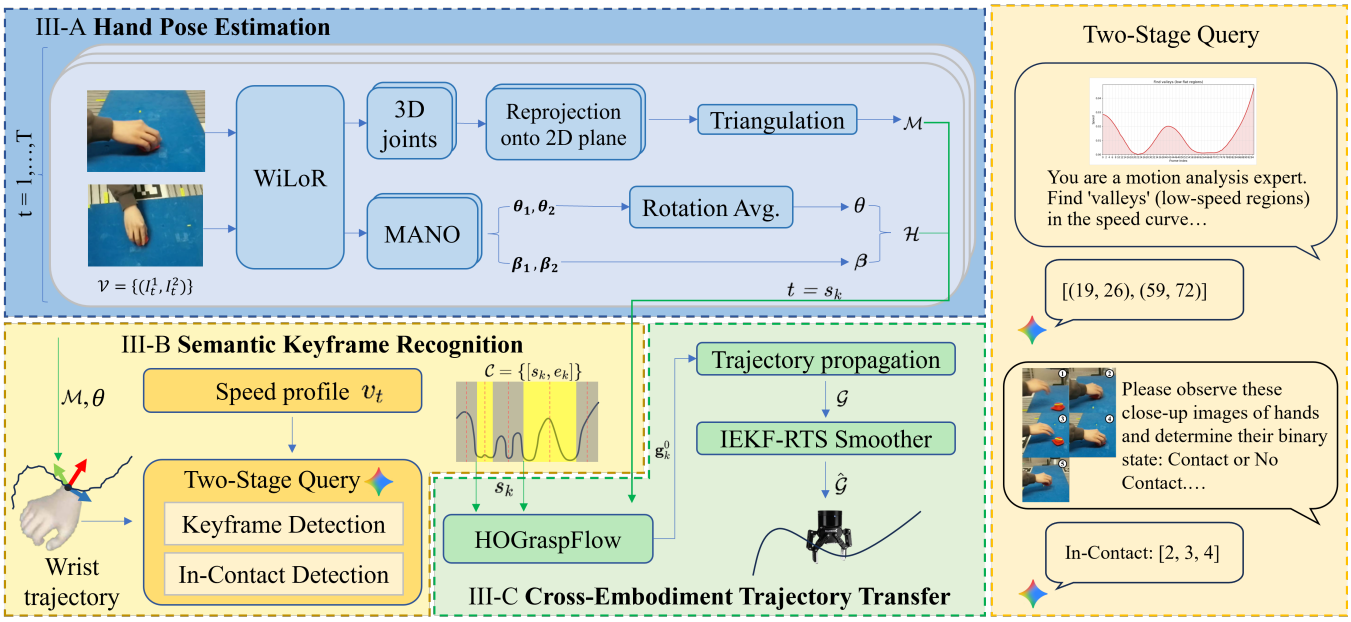

Fig. 2: Pipeline for HOMimic.

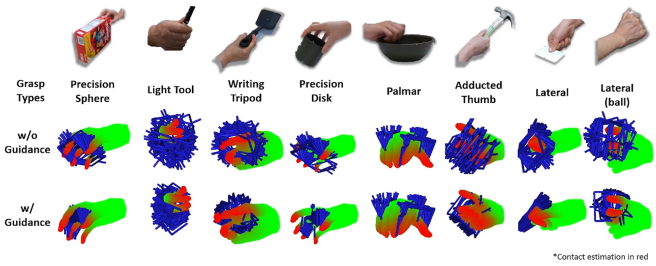

Fig. 3: *HOGraspFlow* [24], a flow matching-based grasp retargeting framework that aims to transfer diverse human grasp types to multi-modal parallel-jaw grasps.

that the entire trajectory is covered by alternating candidate and non-candidate segments.

For **In-Contact Detection**, a representative keyframe is extracted from the temporal midpoint of each segment J_k. These keyframes are first cropped by the WiLoR's hand detection results and arranged into a numbered grid image. Together with a predefined text prompt, they're given to VLM to determine whether genuine contact is present in each segment. Based on the VLM output, every J_k is assigned a semantic label of either contact or non-contact, which directly determines the gripper state. The final contact segments are then obtained by selecting the segments labeled as contact and merging adjacent ones when necessary, yielding $C_k = [s_k, e_k]$, where s_k and e_k denote the start and end frame of the k-th contact segment, respectively.

C. Cross-Embodiment Trajectory Transfer

This stage converts the task-relevant segments into an executable PJ gripper trajectory. Given the contact segments $C = \{[s_k, e_k]\}$ identified in the previous stage, a grasp pose g_k^0 at the first frame s_k will be initialized of each segment using *HOGraspFlow* (Sec. III-C.a), and then propagate it along the corresponding wrist trajectory obtain the gripper motion $\mathcal{G}$, which is further refined using an $SE(3)$-aware Iterative Extended Kalman Filter with a Rauch–Tung–Striebel (IEKF-RTS) smoother [56] to yield the execution-ready output $\hat{\mathcal{G}}$ (Sec. III-C.b),

1) HOGraspFlow for Grasp Transfer: *HOGraspFlow* [24] is an affordance-centric grasp retargeting module that converts a single HOI RGB frame into multi-modal executable PJ grasp poses, which do not require explicit object geometry priors and infers grasp intent directly from hand reconstruction and visual interaction cues. The example outcomes of *HOGraspFlow* are shown in Fig. 3.

For the k-th contact segment, we use the start frame s_k together with the corresponding **Hand Pose Estimation** outputs M_{s_k} and $\mathcal{H}_{s_k}$ as the input to *HOGraspFlow*. Concretely, the cropped hand-object image is processed by DINOv2 [57] to extract a visual feature z_j, while the estimated hand pose and shape parameters are encoded into a hand feature h_j. These two modalities are fused by self-attention into a single HOI-aware descriptor y_j, which serves as a compact intermediate representation of the local HOI semantics: $y_j = \mathrm{SelfAttn}([h_j, z_j]) \in \mathbb{R}^D$.

The fused descriptor is further regularized by two complementary branches. First, a lightweight contact decoder predicts per-vertex contact probabilities $\hat{c}_j$ over the MANO mesh. Second, a grasp taxonomy classifier predicts a probability distribution π_j over the 33 grasp types defined by the GRASP taxonomy. A learnable codebook is then used to construct a soft taxonomy-aware prior: $\hat{\gamma}_j = \sum_{k=1}^{K} \pi_{j,k}\gamma_k$, where $\hat{\gamma}_j$ encodes the intended grasp category in a continuous form, allowing the model to preserve multiple valid retargeted solutions while still constraining the overall structure

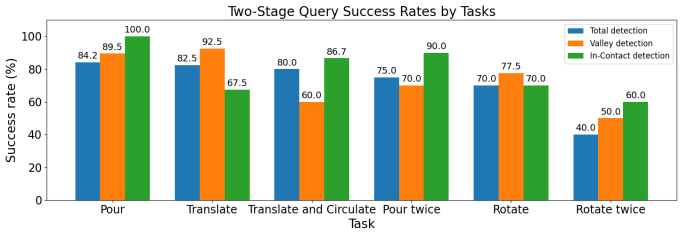

Fig. 4: Success rate of keyframe and contact recognition.

of the pose distribution.

Conditioned on the HOI-aware descriptor y_j and the taxonomy-aware prior $\hat{\gamma}_j$, *HOGraspFlow* generates a set of PJ poses:

$$g_j := (\rho_j, \phi_j) \sim p(g \mid y_j, \hat{\gamma}_j), g \in SE(3) \tag{2}$$

by leveraging flow matching on the $SE(3)$ manifold [21]:

$$\rho_{t'} = \rho_0 + t' R_{\phi_0} \Delta\rho, \quad \phi_{t'} = \phi_0 \cdot \exp_{SE(3)}(t' \Delta\phi), \tag{3}$$

where $\rho_j \in \mathbb{R}^3$ is the gripper translation and $\phi_j \in S^3$ is the unit quaternion representing orientation. $t' \in [0, 1]$ is the discrete time step for iterative sampling on the estimated flow geodesics.

Finally, given the grasp predicted at the contact onset s_k denoted by g_k^0, it is transformed to the world frame through the corresponding hand poses M_t in the world frame, and used to initialize the subsequent trajectory transfer.

2) Trajectory Propagation and Smoothing: The initialized grasp g_k^0 at frame s_k should be propagated over the full contact segment $[s_k, e_k]$. Following the rigid-coupling assumption after contact establishment, the relative transformation between the wrist poses and the retargeted gripper grasp is kept constant within the same segment. Let $T_w(s_k)$ denote the wrist pose at the segment onset. The wrist-relative grasp transform is computed as:

$$g_k(t) = T_w(t)\, T_w(s_k)^{-1} g_k^0, \qquad t \in [s_k, e_k]. \tag{4}$$

Applying this propagation to each contact segment and concatenating the resulting segment-wise trajectory yields the full gripper trajectory $\mathcal{G}$.

Since the propagated trajectory $\mathcal{G}$ still contains high-frequency noise from **Hand Pose Estimation** and temporal propagation, we further smooth it using the IEKF-RTS Smoother [56] on the $SE(3)$ manifold. The final PJ gripper trajectory $\hat{\mathcal{G}}$ preserves the task-relevant interaction pattern of the human video while adapted to the target embodiment. The examples of trajectory outcomes are illustrated in Fig. 5.

IV. Experiments

We evaluate the proposed **Semantic Keyframe Recognition** module by explicitly analyzing its two-stage query design: (i) *Keyframe Detection*, which identifies candidate low-velocity valley intervals from the wrist-speed profile, and (ii) *In-Contact Detection*, which determines whether each candidate segment corresponds to genuine hand-object contact. (iii) Finally, we report the end-to-end *total detection* success rate of the complete two-stage pipeline.

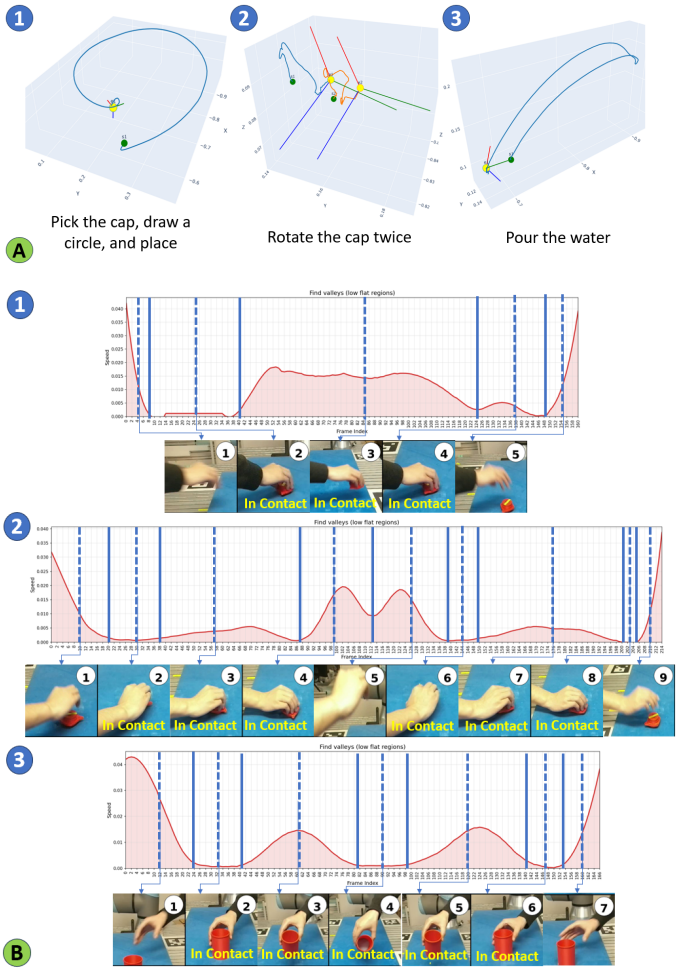

Fig. 5: Visualization of outcomes from HOMimic results: (A) Smoothed robot trajectories and corresponding task descriptions (B) For each task, we visualize the speed profiles, recognized keyframes, and contact recognition results from Gemini 2.5 flash.

As summarized in Fig. 4, the experiments are conducted on 6 representative manipulation tasks that cover both single-stage and multi-stage interaction patterns with varying degrees of occlusion and rotational complexity. The results show that the first stage achieves strong valley localization performance on structurally regular motions, while the second stage yields high in-contact recognition accuracy on tasks with clearer sustained contact. These indicate that the proposed coarse-to-fine formulation effectively decouples motion-based temporal proposal generation from semantic contact verification, leading to robust recognition performance even under substantial object occlusion (e.g. the small red cap) and repeated interaction.

V. Conclusions

We presented HOMimic, a hand-centric framework for transferring manipulation from human videos to executable parallel-jaw gripper trajectories. Specifically, HOMimic combines 3D hand pose recovery, VLM-based two-stage seman-

tic keyframe recognition for contact segment identification, and taxonomy-aware grasp retargeting without relying on explicit object geometry. Experimental results indicate robust recognition performance across diverse tasks, including occluded and multi-stage interactions.

REFERENCES

[1] H. Chen *et al.*, "Vidbot: Learning generalizable 3d actions from in-the-wild 2d human videos for zero-shot robotic manipulation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 27 661–27 672.

[2] S. Kareer *et al.*, "Egomimic: Scaling imitation learning via egocentric video," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 13 226–13 233.

[3] H. Zhou *et al.*, "You only teach once: Learn one-shot bimanual robotic manipulation from video demonstrations," *arXiv preprint arXiv:2501.14208*, 2025.

[4] K. Shaw, S. Bahl, and D. Pathak, "Videodex: Learning dexterity from internet videos," in *Conference on Robot Learning*. PMLR, 2023, pp. 654–665.

[5] Q. Zeng *et al.*, "Activeumi: Robotic manipulation with active perception from robot-free human demonstrations," *arXiv preprint arXiv:2510.01607*, 2025.

[6] H. Xiong *et al.*, "Robotube: Learning household manipulation from human videos with simulated twin environments," in *6th Annual Conference on Robot Learning*, 2022.

[7] G. Papagiannis *et al.*, "R+ x: Retrieval and execution from everyday human videos," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 8284–8290.

[8] R. Wang *et al.*, "Gat-grasp: Gesture-driven affordance transfer for task-aware robotic grasping," in *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2025, pp. 1076–1083.

[9] M. Lepert, J. Fang, and J. Bohg, "Phantom: Training robots without robots using only human videos," *arXiv preprint arXiv:2503.00779*, 2025.

[10] T. Feix *et al.*, "The grasp taxonomy of human grasp types," *IEEE Transactions on human-machine systems*, vol. 46, no. 1, pp. 66–77, 2015.

[11] J. Bohg *et al.*, "Data-driven grasp synthesis—a survey," *IEEE Transactions on robotics*, vol. 30, no. 2, pp. 289–309, 2013.

[12] M. J. Kim, J. Wu, and C. Finn, "Giving robots a hand: Learning generalizable manipulation with eye-in-hand human video demonstrations," *arXiv preprint arXiv:2307.05959*, 2023.

[13] H. Bharadhwaj *et al.*, "Zero-shot robot manipulation from passive human videos," *arXiv preprint arXiv:2302.02011*, 2023.

[14] ——, "Towards generalizable zero-shot manipulation via translating human interaction plans," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 6904–6911.

[15] ——, "Gen2act: Human video generation in novel scenarios enables generalizable robot manipulation," *arXiv preprint arXiv:2409.16283*, 2024.

[16] J. Kerr *et al.*, "Robot see robot do: Imitating articulated object manipulation with monocular 4d reconstruction," *arXiv preprint arXiv:2409.18121*, 2024.

[17] J. Ma *et al.*, "Egoloc: A generalizable solution for temporal interaction localization in egocentric videos," *arXiv preprint arXiv:2508.12349*, 2025.

[18] H.-S. Fang *et al.*, "Graspnet-1billion: A large-scale benchmark for general object grasping," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 11 441–11 450.

[19] M. Sundermeyer *et al.*, "Contact-graspnet: Efficient 6-dof grasp generation in cluttered scenes," in *2021 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2021, pp. 13 438–13 444.

[20] J. Urain *et al.*, "Se (3)-diffusionfields: Learning smooth cost functions for joint grasp and motion optimization through diffusion," in *2023 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2023, pp. 5923–5930.

[21] B. Lim *et al.*, "Equigraspflow: SE(3)-equivariant 6-dof grasp pose generative flows," in *8th Annual Conference on Robot Learning*, 2024.

[22] Z.-H. Yin, S. Yang, and P. Abbeel, "Object-centric 3d motion field for robot learning from human videos," *arXiv preprint arXiv:2506.04227*, 2025.

[23] R. A. Potamias *et al.*, "Wilor: End-to-end 3d hand localization and reconstruction in-the-wild," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 12 242–12 254.

[24] Y. Shi *et al.*, "Hograspflow: Taxonomy-aware hand-object retargeting for multi-modal se(3) grasp generation," *arXiv preprint arXiv:2509.16871*, 2026.

[25] T. Zhang *et al.*, "Easymimic: A low-cost framework for robot imitation learning from human videos," *arXiv preprint arXiv:2602.11464*, 2026.

[26] J. Shi *et al.*, "Zeromimic: Distilling robotic manipulation skills from web videos," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 16 939–16 947.

[27] H. Yuan *et al.*, "Demograsp: Universal dexterous grasping from a single demonstration," *arXiv preprint arXiv:2509.22149*, 2025.

[28] J. Ren *et al.*, "Motion tracks: A unified representation for human-robot transfer in few-shot imitation learning," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 8802–8810.

[29] S. Haldar and L. Pinto, "Point policy: Unifying observations and actions with key points for robot manipulation," *arXiv preprint arXiv:2502.20391*, 2025.

[30] S. Bahl *et al.*, "Affordances from human videos as a versatile representation for robotics," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 13 778–13 790.

[31] A. Patel *et al.*, "Learning to imitate object interactions from internet videos," *arXiv preprint arXiv:2211.13225*, 2022.

[32] H. Bharadhwaj *et al.*, "Track2act: Predicting point tracks from internet videos enables generalizable robot manipulation," in *European Conference on Computer Vision*. Springer, 2024, pp. 306–324.

[33] H. G. Singh *et al.*, "Hand-object interaction pretraining from videos," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 3352–3360.

[34] H. Chen *et al.*, "Tool-as-interface: Learning robot policies from observing human tool use," *arXiv preprint arXiv:2504.04612*, 2025.

[35] Y. Zhu *et al.*, "Vision-based manipulation from single human video with open-world object graphs," *arXiv preprint arXiv:2405.20321*, 2024.

[36] H. Freeman, C. H. Kim, and G. Kantor, "Warped: Wrist-aligned rendering for robot policy learning from egocentric human demonstrations," *arXiv preprint arXiv:2604.10809*, 2026.

[37] M. Lepert, J. Fang, and J. Bohg, "Masquerade: Learning from in-the-wild human videos using data-editing," *arXiv preprint arXiv:2508.09976*, 2025.

[38] L. Y. Chen *et al.*, "Mirage: Cross-embodiment zero-shot policy transfer with cross-painting," *arXiv preprint arXiv:2402.19249*, 2024.

[39] ——, "Rovi-aug: Robot and viewpoint augmentation for cross-embodiment robot learning," *arXiv preprint arXiv:2409.03403*, 2024.

[40] L. Heng *et al.*, "Rwor: Generating robot demonstrations from human hand collection for policy learning without robot," in *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2025, pp. 13 544–13 551.

[41] Z. Qian *et al.*, "Wristworld: Generating wrist-views via 4d world models for robotic manipulation," *arXiv preprint arXiv:2510.07313*, 2025.

[42] H. Ding *et al.*, "Imagination at inference: Synthesizing in-hand views for robust visuomotor policy inference," *arXiv preprint arXiv:2509.15717*, 2025.

[43] V. Liu *et al.*, "Egozero: Robot learning from smart glasses," *arXiv preprint arXiv:2505.20290*, 2025.

[44] Y. Xu *et al.*, "Egodemogen: Novel egocentric demonstration generation enables viewpoint-robust manipulation," *arXiv preprint arXiv:2509.22578*, 2025.

[45] S. Tian *et al.*, "View-invariant policy learning via zero-shot novel view synthesis," *arXiv preprint arXiv:2409.03685*, 2024.

[46] Z. Xiao, R. Wang, and X. Chen, "Robopca: Pose-centered affordance learning from human demonstrations for robot manipulation," *arXiv preprint arXiv:2603.07691*, 2026.

[47] M. Kokic, D. Kragic, and J. Bohg, "Learning task-oriented grasping from human activity datasets," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3352–3359, 2020.

[48] W. Dong *et al.*, "Rtagrasp: Learning task-oriented grasping from human videos via retrieval, transfer, and alignment," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 1–7.

[49] T. Fukumi, "Trading strategy adipted optimization of european call option," *arXiv preprint math/0503444*, 2005.

[50] D. Huang *et al.*, "Hgdiffuser: efficient task-oriented grasp generation via human-guided grasp diffusion models," in *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2025, pp. 19 538–19 545.
[51] N. Khargonkar *et al.*, "Neuralgrasps: Learning implicit representations for grasps of multiple robotic hands," in *Conference on robot learning*. PMLR, 2023, pp. 516–526.
[52] M. Attarian *et al.*, "Geometry matching for multi-embodiment grasping," in *Conference on Robot Learning*. PMLR, 2023, pp. 1242–1256.
[53] J. Romero, D. Tzionas, and M. J. Black, "Embodied hands: Modeling and capturing hands and bodies together," *arXiv preprint arXiv:2201.02610*, 2022.
[54] H. C. Longuet-Higgins, "A computer algorithm for reconstructing a scene from two projections," *Nature*, vol. 293, no. 5828, pp. 133–135, 1981.
[55] G. Team *et al.*, "Gemini: a family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2023.
[56] S. Särkkä, "Unscented rauch–tung–striebel smoother," *IEEE transactions on automatic control*, vol. 53, no. 3, pp. 845–849, 2008.
[57] M. Oquab *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

VI. Appendix

We visualize the entire pipeline of HOGraspFlow [24] in Fig 6 and the attended feature y_j in Fig 7. The hardware setups is visualized in Fig 8.

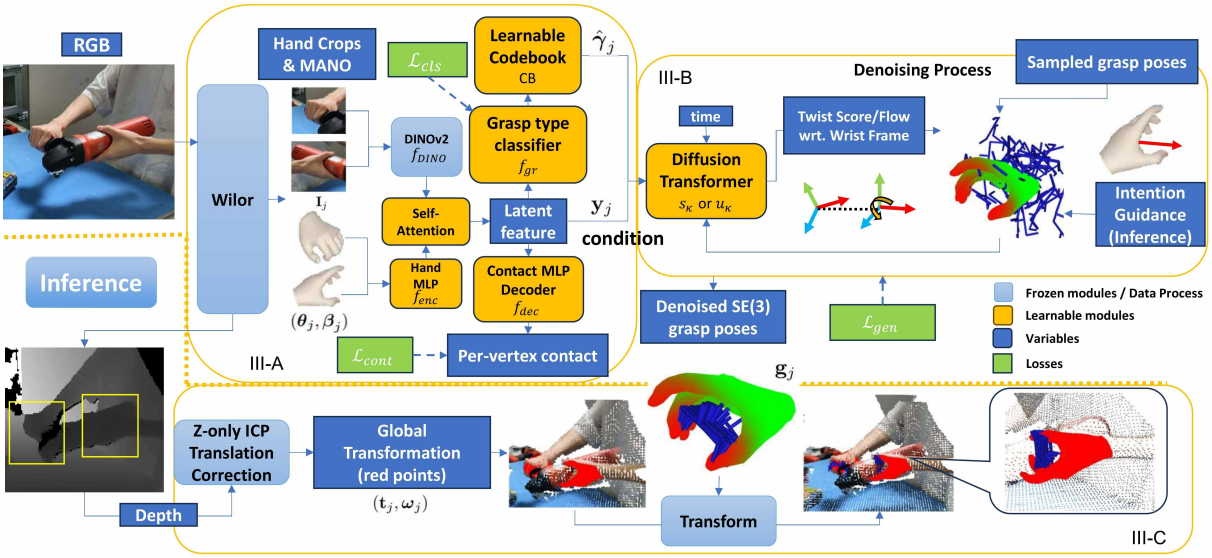

Fig. 6: HOGraspFlow pipeline

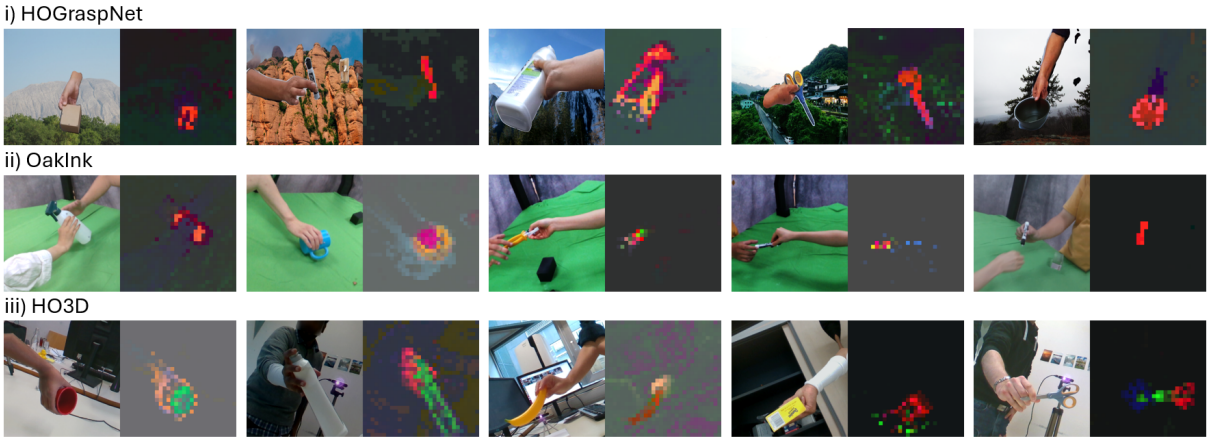

Fig. 7: Visualizations of the feature maps from the HOGraspflow's self-attention outcomes (y_j) on 3 HOI datasets (in PCA), including: HOGraspNet, OakInk and HO3D. Though HOGraspflow is a pure image-based grasp retargeting framework, it understands the coarse geometric information by focusing on the HOI pixels without being explicitly trained on object/hand segmentation.

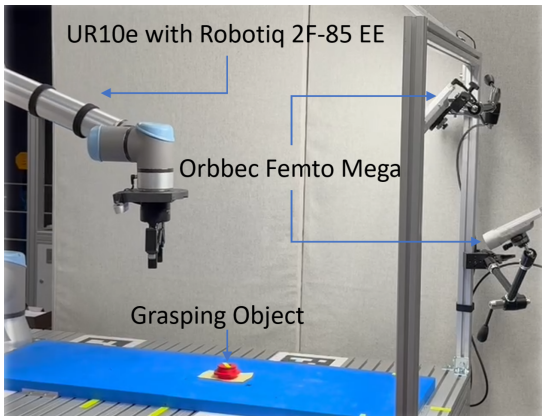

Fig. 8: Experiment setups