GraspAuto: Real-Time Single-Tap Grasp Authoring for VR/AR Experiences

QINGXUAN YAN, Cornell Tech, Cornell University, USA
XIANG CHANG, The Hong Kong University of Science and Technology (Guangzhou), China

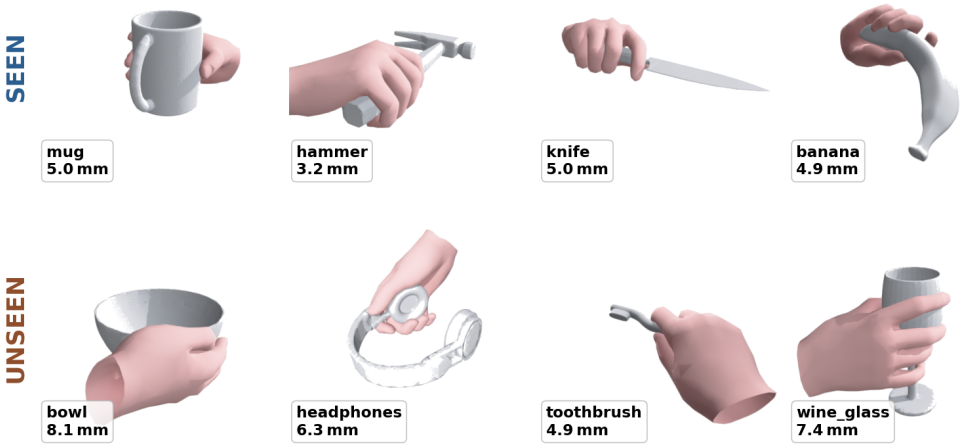

Fig. 1. **GraspAuto** synthesizes XR Hand joint transforms at the user's tap point. Top: four objects from the 21-class **TRUE SEEN** subset; bottom: four objects from the 4-class **CP-UNSEEN** split (bowl, headphones, toothbrush, wine_glass). Per-object label: best-of-1 mode-cover error of the deployable single-checkpoint model (10.64 mm TRUE SEEN average, 60 ms on Quest 3).

GraspAuto is a method for generating a realistic VR hand grasp from a single user tap on a 3D object, producing a controllable pose that conforms to the indicated contact point without motion-capture recording or per-joint tuning. Inference fits within a standalone VR headset's latency budget, so the same system serves content creators as an offline grasp-authoring tool and VR applications as a live hand-synthesis module at runtime. The key idea is a geometric conditioning token, a grip-sphere specifying where to grasp, how wide to open, and from which direction to approach, easy to elicit from a VR controller and dataset-agnostic enough for joint training across grasp corpora. On ContactPose, a single GraspAuto checkpoint transfers to never-seen object classes roughly twice as reliably as a heatmap-conditioned baseline with the same backbone and training data.

Additional Key Words and Phrases: hand-object interaction, AR/VR, hand grasp synthesis, conditional flow matching

1 INTRODUCTION

Convincing hand-object interaction is central to a believable VR experience, yet authoring hand grasps for 3D objects remains slow: it requires either in-headset motion-capture recording or the manual tuning of dozens of hand-joint parameters. A single object often needs several distinct grasp poses for different approach directions and user intents. These poses must be pre-made, and modern VR increasingly ships runtime-generated 3D content for which no grasp was ever authored, which breaks immersion the moment a user reaches for the new geometry.

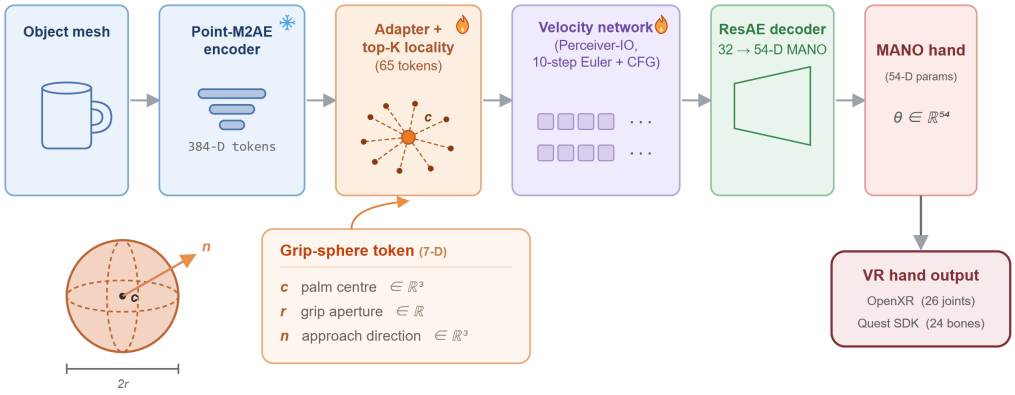

Fig. 2. GraspAuto pipeline. A frozen Point-M2AE encodes the object mesh into 64 patch tokens; the adapter fuses them with the 7-D grip-sphere (palm centre c, aperture r, approach direction n) through top-K tap-centered locality gating. A Perceiver-IO velocity network samples in a 32-D latent space via conditional flow matching, which the ResAE decodes to MANO parameters and a VR hand pose in OpenXR or Quest SDK format.

Existing solutions fall short. Physics-based optimization [3] takes seconds per sample. Motion-capture methods [2] need whole-body tracking unavailable on stand-alone headsets. Recent learning-based methods either lack explicit user-intent conditioning [4, 9] or require user-clicked semantic contact maps [5] incompatible with a single runtime tap, and none are reported with measured on-device VR latency.

We target this gap with a single-tap generator that turns a tap-indicated palm target into 26 XR Hand joint transforms inside the 60 ms Quest 3 budget. **Contributions.** (1) **GraspAuto**, a method that generates a realistic hand grasp on any 3D object from a single user tap, without gesture recording or per-joint tuning. (2) A practical deployment: a 28.9 MB fp32 Unity Sentis ONNX bundle runs in 60 ms per tap on Quest 3, serving both offline grasp authoring and live on-device runtime synthesis.

2 METHOD

Given an object and a grip-sphere gesture, GraspAuto emits a 54-D MANO [10] parameter vector (6-D rotation + 3-D translation + 45-D pose) that a forward kinematics pass turns into 26 XR Hand transforms.

(i) Object encoding. At scene load we sample the object mesh to 2048 points and run a frozen Point-M2AE encoder [6] once, producing 64 patch tokens (384 D) with 3-D centres; cached per object.

(ii) Grip-sphere intent token. Prior generators encode user intent as dense contact heatmaps [3, 5] or heatmap-statistic tokens tied to a dataset-specific contact head, and both break under cross-dataset training (Supp. S7). We use a 7-D **grip-sphere**: palm centre (3-D, mean of five MANO palm anchors), grip aperture (1-D, mean centre-to-fingertip distance), and approach direction (3-D, palm normal towards fingertips). The sphere is deterministic from MANO at training time and directly specifiable from a VR controller at inference. A top-K_{patch} locality gate keeps the 8 patches closest to the sphere centre (zeroing the rest), trading $\sim$1 mm of TRUE SEEN cost for $\sim$1 mm of UNSEEN gain.

(iii) Latent flow-matching generator. A 4-block residual MANO autoencoder maps 54-D MANO to a 32-D latent; only the 1.08 M-parameter decoder is shipped for inference (**0.19 mm** reconstruction error).

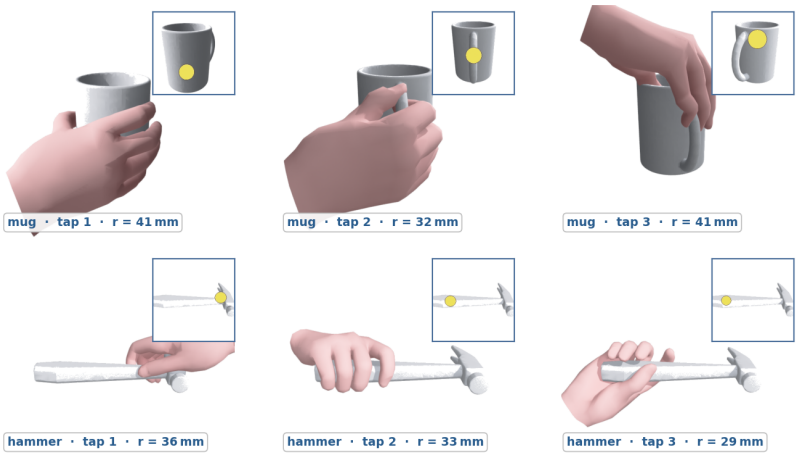

Fig. 3. **Tap-conditioned controllability.** Three tap locations on a mug (top) and a hammer (bottom); the inset mini-map marks each tap in yellow. All six grasps come from a single GraspAuto checkpoint (bo-64 selector-picked); r is the grip aperture in millimetres.

A 6.0 M Perceiver-IO velocity network cross-attends the 65-token bundle and is trained with conditional flow matching [7]; inference runs 10 Euler steps with CFG scale 1.5. A frozen K-means codebook adds a light natural-mode commitment loss (Supp. S3).

(iv) Ensemble pool and deployable ranker. We train eight diverse sphere-conditioned checkpoints (Supp. S2), each sampling B=64 grasps per tap. The 512-candidate pool is GT-ranked for the oracle (5.14 mm median TRUE SEEN). For deployment we use a **medoid consensus** ranker (candidate with smallest mean distance to the rest), GT-free and ∼2 s per tap, reaching 9.74/12.71 mm SEEN/UNSEEN. GT-target distillation [8] did not beat the best single checkpoint and is deferred.

3 RESULTS

We report **mode-cover (MC)** error [1], which rewards matching *any* recorded GT grasp for the object:

$$\mathrm{MC}(\hat{v}, \mathcal{G}_o) \;=\; \min_{g \in \mathcal{G}_o} \frac{1}{778} \sum_{i=1}^{778} \|\hat{v}_i - v_i^{(g)}\|_2 \quad \text{(mm)},$$

with $\hat{v}$ the generated MANO mesh in the object frame and bo-N the rank-1 pick over N candidates. Val participants are disjoint from training (standard ContactPose practice). We split the val set by object class: **TRUE SEEN** uses 21 classes (N=179) and **CP-UNSEEN** the remaining 4 (bowl, headphones, toothbrush, wine_glass; 37 samples). Both subsets contain classes seen in training from different participants (Supp. S6).

Sphere vs heatmap. Swapping the 11-D heatmap token for the 7-D grip-sphere (Table 1) cuts bo-1 TRUE SEEN from 12.48 to 10.64 mm and CP-UNSEEN from 27.9 to 13.7 mm (51% reduction); the 8-member pool reaches 5.14/6.79 mm median SEEN/UNSEEN at 512 candidates.

Deployment. The shipped adapter, velocity network, and AE decoder are 7.2 M parameters (28.9 MB fp32, ∼14.6 MB fp16) with per-object Point-M2AE features precomputed offline (95 KB each); a full on-device mode that also runs the encoder adds 61 MB fp32 (Supp. S4). Code: https://github.com/crazycatseven/GraspAuto.

Table 1. Main results on ContactPose (mode-cover error, mm; bo-1 with CFG 1.5 except for pool rows). Row groups, top to bottom: prior-architecture baselines, single sphere-conditioned checkpoints, and the 8-member pool at B=64 (512 candidates). Oracle rows are upper bounds; the consensus medoid row is the deployable pool pick (no GT).

Config	**TRUE SEEN**	**CP-UNSEEN**	**<10 mm**	**Latency**
codebook retrieval (ours)	62.3	–	0%	5 s
heatmap-palm distill student	12.48	27.9	36%	**60 ms**
sphere, CP-only	12.36	15.92	45%	**60 ms**
sphere, CP + OakInk mix	11.70	14.86	55%	**60 ms**
sphere, CP + OakInk filtered	**10.64**	**13.70**	**59%**	**60 ms**
8-member pool, GT oracle (mean)	**6.13**	**7.58**	**89%**	30 s
median	**5.14**	**6.79**	–	–
consensus medoid (deployable, no GT)	**9.74**	**12.71**	**67%**	2 s

4 DISCUSSION

The grip-sphere unlocks cross-dataset training that prior heatmap-statistic tokens could not support. A frozen-checkpoint control (Supp. S7) shows that geometric replacements of the non-centroid palm-heatmap dimensions perform *worse* than zeroing them: those dimensions carried dataset-specific learned semantics, not transferable geometry. With sphere conditioning, multi-dataset mixing becomes monotonic as data quality rises (12.36 → 11.70 → 10.64 mm bo-1 SEEN). **Limitations:** the real-time bo-1 at 10.64 mm trails the 5.14 mm pool median, and closing that gap via distillation is open; no VR user study yet (Supp. S6 proxies).

REFERENCES

[1] Samarth Brahmbhatt, Chengcheng Tang, Christopher D. Twigg, Charles C. Kemp, and James Hays. 2020. ContactPose: A Dataset of Grasps with Object Contact and Hand Pose. In *European Conference on Computer Vision (ECCV)*.

[2] Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. 2020. GRAB: A Dataset of Whole-Body Human Grasping of Objects. In *European Conference on Computer Vision (ECCV)*.

[3] Shaowei Liu, Yang Zhou, Jimei Yang, Saurabh Gupta, and Shenlong Wang. 2023. ContactGen: Generative Contact Modeling for Grasp Generation. In *IEEE International Conference on Computer Vision (ICCV)*.

[4] Xiaofei Wu, Tao Liu, Caoji Li, Yuexin Ma, Yujiao Shi, and Xuming He. 2024. FastGrasp: Efficient Grasp Synthesis with Diffusion. *arXiv preprint arXiv:2411.14786* (2024).

[5] Peishan Hou, Xuefeng Lin, Xuemiao Wang, Lei Xiang, Yi Zhang, and Xiaojuan Qi. 2024. ClickDiff: Click to Induce Semantic Contact Map for Controllable Grasp Generation with Diffusion Models. In *ACM International Conference on Multimedia (ACM MM)*.

[6] Renrui Zhang, Ziyu Guo, Peng Gao, Rongyao Fang, Bin Zhao, Dong Wang, Yu Qiao, and Hongsheng Li. 2022. Point-M2AE: Multi-scale Masked Autoencoders for Hierarchical Point Cloud Pre-training. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[7] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. 2023. Flow Matching for Generative Modeling. In *International Conference on Learning Representations (ICLR)*.

[8] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the Knowledge in a Neural Network. *arXiv preprint arXiv:1503.02531* (2015).

[9] Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael J. Black, Krikamol Muandet, and Siyu Tang. 2020. Grasping Field: Learning Implicit Representations for Human Grasps. In *International Conference on 3D Vision (3DV)*.

[10] Javier Romero, Dimitrios Tzionas, and Michael J. Black. 2017. Embodied Hands: Modeling and Capturing Hands and Bodies Together. *ACM Transactions on Graphics (SIGGRAPH Asia)* 36, 6 (2017).