GraspAuto: Supplementary Material

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

S1. SCALING AND RANKING ABLATION

Mean mode-cover error (mm) on ContactPose val. TRUE SEEN is the 21-class subset (N=179); UNSEEN is the remaining 4-class subset (37 samples; bowl, headphones, toothbrush, wine_glass). Both subsets use participants disjoint from training; UNSEEN is a per-class reporting split, not a class-level holdout. Oracle rows are upper bounds (GT-ranked, not deployable); consensus and bo-1 are deployable.

Key patterns: OakInk contact-filter gives the largest single-checkpoint gain (-1.72 mm SEEN); pool oracle saturates past 6 members on mean but keeps improving on median; consensus-medoid closes 57% of the random-vs-oracle gap without any GT.

S2. ENSEMBLE MEMBER DETAILS

Members differ along five axes: seed, OakInk mix ratio, OakInk subset (all / contact-filtered / real-MoCap), naturalness weight, and sphere-radius reduction. Oracle picks are most spread across the OakInk-subset axis.

S3. DISTILLATION TRAINING PROTOCOL (ATTEMPTED, DEFERRED)

Status: the sphere-era ensemble-to-student distillation did not improve on the best single checkpoint in our experiments; we deferred it and report the protocol here for completeness. The oracle-pool ceiling (5.14 mm median SEEN) sits ∼5.5 mm below the best single checkpoint (10.64 mm bo-1), and collapsing that ceiling into a real-time student remains open.

Teacher targets are computed offline: for each ContactPose training sample we run all eight members (B=64 candidates per member) and keep the GT-ranked best by mode-cover error against the per-object reference grasps. The student is initialized from our best single checkpoint and trained with a CFM MSE loss in the 32-D AE latent space:

$$\mathcal{L}_{\text{distill}} = \mathbb{E}_{t,z_0} \| v_\phi(z_t, t \mid \mathbf{c}) - (z^\star - z_0) \|^2$$

where z_0 is the noise latent, z_t is the linear interpolation at time t, and $z^\star$ is the AE-encoded teacher pick. Hyperparameters:

Optimizer	AdamW, weight decay 1×10^{-5}
Learning rate	5×10^{-5} (constant)
Batch size	16
Epochs	300
Flow steps (inference)	10-step Euler
CFG dropout	0.1 (training); CFG scale 1.5 (inference)
Auxiliary losses	none (no naturalness commitment; $z^\star$ is near-natural by construction)

Table 1. Sphere ensemble: how pool size and ranker trade off quality vs. latency on ContactPose. Oracle uses GT ranking (upper bound, not deployable).

Config	SEEN	UNSEEN	<10 mm	Latency
Single sphere checkpoint (bo-1 CFG 1.5):				
CP-only baseline	12.36	15.92	45%	60 ms
+ OakInk warm-start	11.70	14.86	55%	60 ms
+ OakInk contact-filter	**10.64**	**13.70**	**59%**	60 ms
Sphere ensemble pool (B=64/member, GT-ranked oracle):				
5 members (320 cand)	6.47	7.81	87%	16 s
6 members (384 cand)	6.24	7.67	88%	20 s
7 members (448 cand)	6.16	7.56	90%	24 s
8 members (512 cand)	**6.13**	**7.58**	**89%**	30 s
median (8 members)	**5.14**	**6.79**	–	–
Ranker ablation (8 members, B=64 = 512 candidates):				
random pick	15.1	20.3	44%	2 s
GT oracle (upper bound)	**6.13**	**7.58**	**89%**	2 s
consensus medoid (deployable, no GT)	**9.74**	**12.71**	**67%**	2 s

Table 2. Eight sphere-conditioned ensemble members and their training recipes.

Label	Warm-start	Data mix	Notable setting
CP-only, seed 42	distill-student base	ContactPose	–
CP + OakInk 30%	CP-only base	CP + OakInk 30%	–
CP-only, seed 123	distill-student base	ContactPose	–
CP + OakInk 15%	CP-only base	CP + OakInk 15%	–
CP + OakInk 30%, natural	CP-only base	CP + OakInk 30%	naturalness weight 0.25
CP + OakInk contact-filtered	CP+OakInk 30% above	CP + filtered OakInk	contact_5mm ≥ 0.10
CP + OakInk real-mocap	CP+OakInk 30% above	CP + real-only OakInk	no TINK variants
Min-radius ablation	filtered above	CP + filtered OakInk	sphere radius = min(palm→tip)

Both top-M=16 softmin sampling and a top-M=1 hard pick regressed the student to ∼17 mm bo-1 SEEN vs. the 10.64 mm non-distilled best checkpoint. We attribute this to the student collapsing the multi-modal pool distribution: the sharp low-error mode present in the best member gets smeared when averaging toward pool means. Closing this gap without losing mode sharpness is the main remaining engineering problem.

S4. HARDWARE + DEPLOYMENT NOTES

Training hardware. Single NVIDIA RTX 4090 (24 GB). Each sphere ensemble member trains in 6–12 min (80–150 epochs, batch 16); the full 8-checkpoint set takes under 2 h.

Inference latency (on RTX 4090, batched):

- Point-M2AE (64 patch tokens from 2048-pt cloud): ∼20 ms (cached per object)
- Adapter: 2 ms
- Velocity network (10-step Euler): 6 ms
- AE decoder: 1 ms
- MANO decode: 5 ms
- **Single bo-1 total:** 34 ms (60 ms on Quest 3 GPU)

ONNX deployment package. Measured fp32 sizes on disk; fp16 sizes are projected from parameter count × 2 B and will vary slightly with Unity Sentis quantization metadata.

- **Mode A:** offline Point-M2AE precompute. Ship `adapter.onnx` (0.48 MB) + `velocity_net.onnx` (24.1 MB) + `ae_decoder.onnx` (4.3 MB) = **28.9 MB fp32** / ∼14.6 MB fp16. Per-object precomputed features: 95 KB each.
- **Mode B:** on-device feature extraction. Add `point_m2ae.onnx` (61.4 MB fp32 / ∼30.7 MB fp16). Mode B total: **90.3 MB fp32** / ∼45.3 MB fp16. Enables user-imported meshes.

We tested Unity Sentis as the runtime; iOS Core ML and Android NNAPI paths are not verified.

S5. NOTE ON TRUE SEEN

Earlier drafts reported N=216 val means that pooled all classes under "SEEN". The numbers in this submission report TRUE SEEN on the 21-class subset (N=179) and CP-UNSEEN separately on the four-class subset (bowl, headphones, toothbrush, wine_glass; 37 samples). We emphasise that these four classes are *not* held out from ContactPose training at the class level: CP training includes them from different participants. The split is a per-class reporting convenience, and "UNSEEN" in our numbers denotes cross-participant plus (for sphere models) cross-sensor generalisation, as further discussed in S6.

S6. NOTE ON UNSEEN AND TRULY-NOVEL-GEOMETRY GENERALIZATION

Our CP-UNSEEN column (Main Table 1) is evaluated on four ContactPose val classes (bowl, headphones, toothbrush, wine_glass). These classes are *not* truly held out from CP training at the class level: CP training includes them from different participants, so even the heatmap baseline has seen each class, just not the specific val mesh+participant combinations. When the sphere checkpoints additionally train on OakInk-Shape, OakInk's object set also includes the same classes (`bowl`: 986 samples, `headphones`: 422, `toothbrush`: 37, `wine_glass`: 160). Our CP-UNSEEN is therefore best characterized as a per-class *cross-participant* (and, for sphere models, also *cross-sensor same-class*) generalization split. True cross-class generalization is probed separately below on 36 GraspXL objects absent from both ContactPose and OakInk.

To probe genuinely novel geometry we evaluated on 36 GraspXL objects never seen in ContactPose or OakInk (150 samples total, sampled uniformly across objects). The single deployable sphere bo-1 student with CFG=1.5 reaches **56.7 mm mean / 53.7 mm median / 0% <20 mm**, a 29 mm gap from the 27.9 mm heatmap-palm CP-UNSEEN number in Main Table 1.

A physical-validity diagnostic on these same samples (31-sample subset, comparing the model's generated hand against GraspXL's ground-truth grasp on the same object) gives a more VR-relevant read:

Metric	Generated	GraspXL GT
Palm floating gap (mean)	**11.0 mm**	12.3 mm
Palm floating gap (median)	**6.8 mm**	10.6 mm
Hand verts within 5 mm of surface (mean)	66.5	70.6
Pseudo-penetration fraction (6-way neighbour proxy)	6.3%	5.9%

On palm gap, contact density, and penetration, the generations match or slightly exceed the GT statistics. The 56.7 mm mode-cover therefore reflects a *pose-specificity* gap (the generated pose is not the one recorded in GraspXL for that object) rather than a *physical-validity* gap. For VR usability, the hand sits on the object

with contact and penetration comparable to the hand-annotated GT, just in a different valid grasp mode. Scaling training data to higher-diversity corpora (e.g., GraspXL's 400+ objects [1], HOGraspNet [2]) is still the most direct route to closing the mode-cover gap on the sparse-GT truly-unseen split, but physical contact quality is substantially better than the 56.7 mm number alone suggests.

S7. DIAGNOSTIC: PALM-TOKEN SEMANTIC PARITY (RESOLVED BY GRIP-SPHERE)

In the heatmap-palm era, attempts to beat the 11.85 mm CP SEEN baseline by mixing GraspXL at 30% into training (CP/OakInk/GraspXL 40/30/30) regressed CP SEEN to 17–22 mm across five retraining runs, even after fixing three verifiable data bugs: (i) MANO `hands_mean` not baked into GraspXL pose; (ii) manopth vs. manotorch translation-convention mismatch; (iii) hardcoded zeros for the non-centroid palm dimensions in OakInk/GraspXL samples (ContactPose has them from a learned contact heatmap).

To isolate the cause, we ran a frozen-checkpoint control on CP SEEN (bo-1 CFG 1.5, N=179) replacing only the 8 non-centroid palm dimensions with four alternative sources:

Source	Description	Mean (mm)
A	Original CP learned-heatmap	19.71
B	v1 geometric Gaussian (what GraspXL/OakInk were fed)	**49.61**
C	Centroid kept; other dims zero, mass=1	39.97
D	CP contact-head with hand-proximity stage1 prior	46.32

$B > C$ is decisive: a well-scaled replacement is *worse* than zero-ablation. **Geometric replacements are not merely noisier versions of the learned channel; they encode structurally different information.** The adapter has learned to trust the eight non-centroid palm dimensions as a semantic intent channel. Under this regime, 60% of the multi-dataset training batches were actively mis-training the adapter. A diversity-preserving pool-distillation attempt (top-M=16 softmin, τ=1.5 mm) did not improve over the single-oracle student either (18.74 vs. 11.60 mm deployable).

The sphere redesign resolves this. The 7-D grip-sphere (centre + aperture + approach, deterministic from MANO; main paper § Method) is identically defined across ContactPose, OakInk, and any MANO-annotated source, removing the semantic-parity requirement entirely. With sphere conditioning, CP+OakInk mixing now improves CP SEEN bo-1 monotonically (12.36 → 11.70 → 10.64 mm as data quality rises; main paper Table 1). We keep this diagnostic as supplementary context because it motivates the sphere design and quantifies the cost of the alternative paths we considered (palm-feature teacher, pool-diversity distillation, adapter capacity scaling) before arriving at the geometry-first token.

S8. PER-OBJECT GRASP GALLERY

Figure 1 shows illustrative GT-ranked best-of-64 (oracle, not deployable) outputs on all 25 object classes in the ContactPose val set (21 TRUE SEEN classes + 4 CP-UNSEEN classes). Each cell is a paired (input, output) panel: left = object + user tap (red dot), right = same scene + generated hand. For each object we selected up to three sample contexts from val and kept the GT-ranked lowest-error grasp; this is visual coverage, not a per-object performance table, aggregate metrics are in S1 and main paper Table 1. Face colors encode vertex-to-surface distance: red for sub-millimeter penetration, green gradient for < 5 mm contact, skin tone for non-contact vertices.

Fig. 1. Per-object grasp gallery, best-of-64 + CFG 1.5. Each object is a paired (input, output) panel: the left shows the object and the red tap that conditions the generator; the right shows the same scene with the generated hand overlaid. Green labels mark the 21 TRUE SEEN classes; red labels mark the 4 CP-UNSEEN classes (bowl, headphones, toothbrush, wine_glass). Per-object mode-cover errors range from ∼8 to ∼35 mm; the lowest-quality grasps concentrate on thin or unusually shaped objects (flashlight, headphones, wine glass), consistent with the physical-validity diagnostic in S6: on such objects the remaining error is pose specificity rather than contact failure.

S9. FAILURE CASES

Figure 2 shows the three highest-error grasps of the deployable bo-1 student paired with their GT reference. All three are *pose-specificity* gaps, not contact failures: the generator commits to a different but still-plausible valid mode (the opposite end of the eyeglasses, the non-GT face of the cell phone, a different rim of the

bowl). The 8-member pool's 5.14 mm median oracle confirms that some candidate *does* land near the GT; the single deployable student picks one of the other valid modes instead. Closing this multi-modality gap via distillation without losing sharpness remains the main open engineering problem (Main § Discussion).

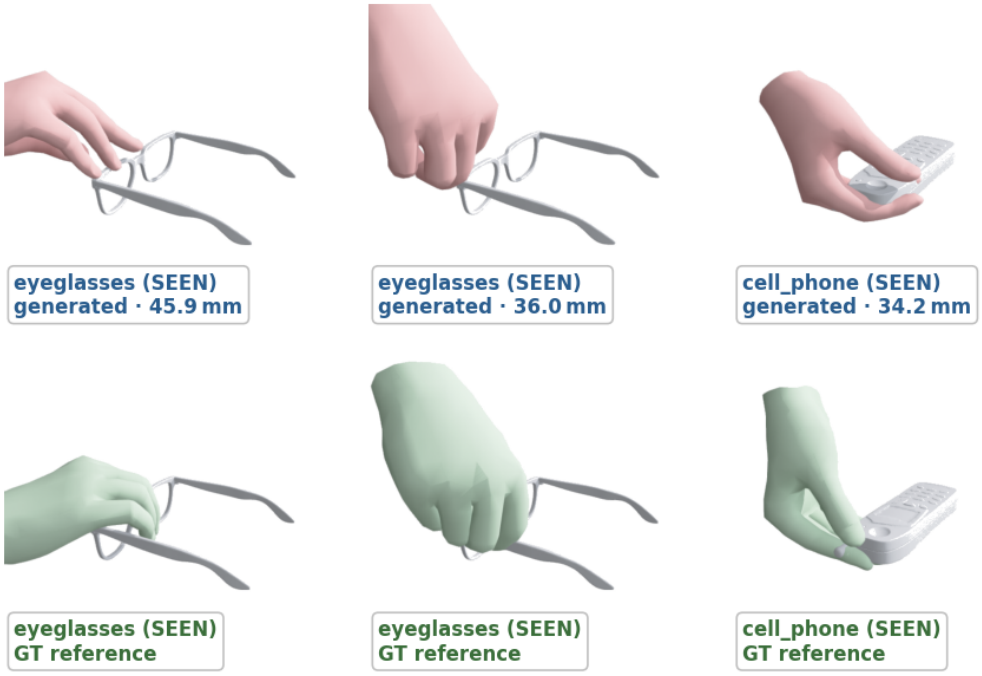

Fig. 2. Top-3 mode-cover failures. Row 1: generated; row 2: GT reference for the same tap. All three are plausible valid grasps on a different mode than the one recorded in GT; physical-contact proxies (S6) remain comparable to GT.

REFERENCES

[1] H. Zhang et al. GraspXL: Generating grasping motions for diverse objects at scale. *ECCV*, 2024.
[2] W. Cho et al. Dense hand-object (HO) GraspNet with full grasping taxonomy and dynamics. *ECCV*, 2024.