Supplementary Material for Grasp in Gaussians: Fast Monocular Reconstruction of Dynamic Hand–Object Interactions

Ayce Idil Aytekin[1,2], Xu Chen[3], Zhengyang Shen[3], Thabo Beeler[3], Helge Rhodin[1,2,4], Rishabh Dabral[1,2], and Christian Theobalt[1,2]

[1] Max Planck Institute for Informatics, Saarland Informatics Campus, Saarbrücken, Germany
[2] Saarbrücken Research Center for Visual Computing, Interaction and AI
[3] Google
[4] Bielefeld University, Bielefeld, Germany
https://aidilayce.github.io/GraG-page/

A Additional Evaluation

Following ARCTIC [4], we also report interaction metrics in Tab. 1 such as Contact Deviation (CDev) that measures hand-object contact accuracy, Acceleration Error (ACC) to measure motion smoothness, Mean Relative-Root Position Error (MRRPE) to measure the root translation between hand and object, and Motion Deviation (MDev) to check when a hand moves an object, if the vertices of the hand and the object in stable contact move together. For detailed explanations of these metrics, please check ARCTIC [4]. Please note that, for object acceleration error, we use a topology-agnostic variant rather than the original ARCTIC definition. Since different methods often output different object meshes, predicted and ground-truth meshes generally do not share vertex correspondence or indexing. Therefore, we measure object acceleration error by comparing the accelerations of object root trajectories over time.

Table 1 shows that GraG consistently improves interaction consistency, with the strongest gains on HOT3D. On HO3Dv3, our method achieves the best temporal interaction quality (best MDev and ACC_h), while remaining close to the strongest contact/pose baseline HOLD. This shows that our predictions are not only accurate in static contact proximity, but also more coherent over time during manipulation.

On HOT3D, the margin is larger: we substantially outperform prior methods in CDev, MDev, and $\mathrm{MRRPE}_{h\to o}$. In particular, CDev drops from the 451–570 mm range of HOLD/BIGS to 26.0 mm, and $\mathrm{MRRPE}_{h\to o}$ drops from 606–762 mm to 154 mm. These improvements indicate better hand-object coupling under more challenging conditions (dynamic camera, larger motion, and noisier geometry). HOT3D setting is particularly difficult for the comparison methods due to them depending on COLMAP/SfM initializations. ACC_o shows a complementary trend. MagicHOI attains the lowest ACC_o on both datasets, but with much worse CDev/$\mathrm{MRRPE}_{h\to o}$, indicating smoother object motion alone does

Table 1: Interaction-consistency metrics. We report interaction metrics from ARCTIC [4] on HO3Dv3 and HOT3D, including Contact Deviation (CDev), Motion Deviation (MDev), hand/object acceleration error (ACC_h, ACC_o), and Mean Relative Root Position Error between the hand and object ($MRRPE_{h\to o}$). Lower is better for all metrics.

Dataset	Method	CDev [mm] ↓	MDev [mm] ↓	ACC_h [m/s^2] ↓	ACC_o [m/s^2] ↓	$MRRPE_{h\to o}$ [mm] ↓
HO3Dv3	HOLD [3]	15.8	10.8	11.3	30.2	38.6
	BIGS [7]	54.2	23.1	14.4	37.8	79.4
	MagicHOI [9]	533	21.5	15.4	10.3	747
	Ours	16.9	7.22	7.81	17.7	57.4
HOT3D	HOLD [3]	451	65.3	15.8	117	606
	BIGS [7]	570	85.6	24.1	136	762
	MagicHOI [9]	463	66.7	20.3	18.3	728
	Ours	26.0	15.9	10.3	27.5	154

not guarantee correct hand-object interaction. Our method GraG provides a better balance: low ACC_o together with strong contact and relative-pose consistency, which is more desirable for physically plausible manipulation. Additional results are in the Supplementary Video.

B Hyperparamaters

We report the concrete settings used by our method for the components described in Section 4.3 (windowed SoG tracking and regularizers).

Windowed optimization. In Stage 3, we jointly refine hand and object variables in short temporal windows $\mathcal{W}$. We use a fixed window size of $|\mathcal{W}| = 8$ frames and slide the window with 1 frame overlapping. For each window, we run 100 gradient steps of joint refinement.

Optimizer and learning rates. In Stage 3, we optimize the variables in each window using AdamW with $\beta_1 = 0.9$ and $\beta_2 = 0.95$. We use the learning rates of 10^{-3} for the object scale s^o, 10^{-3} for the hand scale s^h, 10^{-4} for hand shape coefficients, 2×10^{-3} for object translation offsets, 10^{-3} for hand translations, 2×10^{-3} for object rotations, and 10^{-4} for the hand pose and global orientation parameters.

Loss weights. We optimize the objective in Eq. 3 with the following fixed weights:

$$\begin{aligned} \lambda_{\text{j2d}} &= 0.5, \\ \lambda_{\text{depth}} &= 1000, \\ \lambda_{\text{contact}} &= 5000, \\ \lambda_{\text{sil}} &= 100, \\ \lambda_{\text{smooth}} &= 100. \end{aligned}$$

The SoG alignment energy is maximized by minimizing $-0.05\,E$ in Eq. 3.

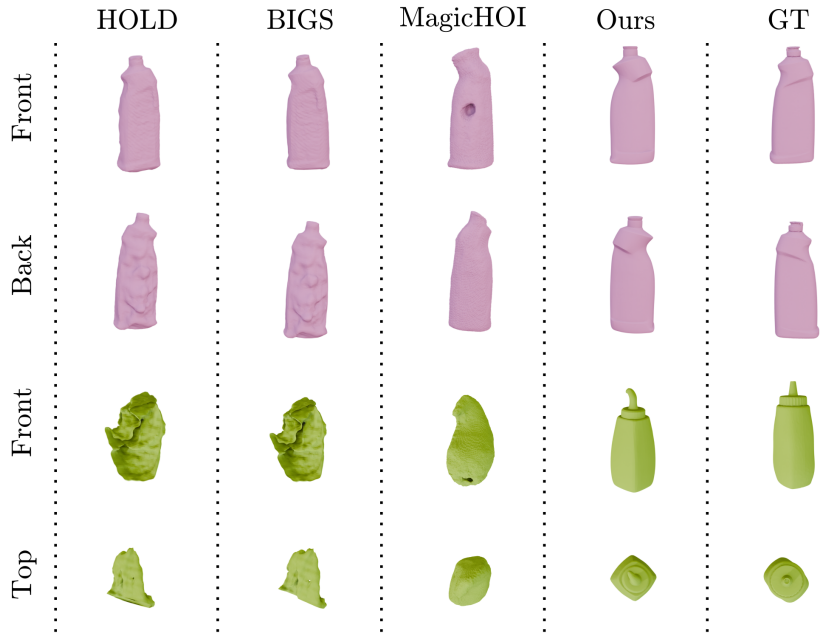

Fig. 1: Object-only reconstruction comparison. Front/back views of reconstructed object geometry HOLD, BIGS, MagicHOI, and GraG; ground truth shown for reference. First 2 rows are from HO3Dv3 and last 2 rows are from HOT3D.

C Method Details

Image SoG. For each frame, we construct the image SoG only inside the binary object mask. We build the image SoG with a quad-tree decomposition using maximum depth 8, color-variance threshold 0.01, minimum cell size 2, bounding-box padding 2, minimum valid-mask ratio 10^{-6}, and geometric sigma assignment. Unlike Stoll et al. [8], we use RGB directly (no HSV conversion).

Object SoG. We start from the dense canonical Gaussian asset decoded by the SAM3D Gaussian decoder and sparsify it with farthest-point sampling. For each sampled Gaussian, we copy its 3D center, rotation, opacity, and DC color coefficients from the dense canonical asset. We then replace the anisotropic scale by an isotropic one using the mean of the three scale axes, and convert it to an isotropic 3D SoG standard deviation with a multiplicative factor of 3.0. All object SoG weights are set to 1. For SoG matching, we use top-96 image-to-model contributions and a color-kernel defined as the following: For two color vectors $c_i, c_j \in \mathbb{R}^3$, with RGB normalized to $[0, 1]$, we define

$$d(c_i, c_j) = \exp\left(-\frac{\|c_i - c_j\|_2^2}{\sigma_c^2}\right), \tag{1}$$

where σ_c controls how strongly color differences are penalized. We set $\sigma_c = 0.15$ in all experiments.

Depth loss details. For the depth prior in Eq. 7, we compute a pointmap depth statistic using the median value within the (eroded) hand and object masks, and align it with the mean rendered depth inside the same masks. Instead of applying a dense depth loss between rendered and predicted depths, we rely on these mean/median statistics to obtain a more robust depth constraint. Depth predictions from general-purpose models such as Depth Anything 3 [6] can be particularly noisy in hand–object interaction videos, specifically when the hand and object are either very close to or far from the camera. This often leads to inaccurate depth estimates, as illustrated in Fig. 3. To avoid degrading the optimization due to such noisy predictions, we adopt this robust mean/median-based depth loss formulation.

Hand depth alignment. In practice, the absolute depth of the Dyn-HaMR hand trajectory can be slightly biased due to the focal length of the base model's training data, which can hinder subsequent joint hand-object refinement. We therefore apply a lightweight hand depth alignment step using a selected view before Stage 3 optimization. Specifically, we first choose a reference frame t where the hand is maximally visible, i.e., the frame with the largest hand mask area. We then erode the chosen hand mask M_t^h to reduce boundary leakage to avoid points with drastically wrong depth, and sample a set of target 3D points $\mathcal{Q}_t$ from the pointmap near the projections of the current MANO mesh vertices: we project hand vertices using (K_t, T_t) and collect pointmap samples within a pixel radius of 2, keeping at most 2000 points. Finally, we align the hand mesh to $\mathcal{Q}_t$ by solving for a global similarity transform (scale and translation) that best matches the sampled point cloud in a robust radius-based alignment. We update the hand scale s^h by the estimated scale factor and update the hand translation τ_t^h by the estimated translation, and use the aligned hand trajectory as initialization for the refinement in Stage 3.

Cordinate system alignments. We use Depth Anything 3 [6] to estimate camera extrinsics/intrinsics, then we employ the same predicted cameras in Dyn-HaMR to predict the hand in the same coordinate system. Since we also get the pointmaps used for object pose estimation in video-adapted SAM3D from Depth Anything 3 by unprojecting the estimated depth via the estimated camera parameters, we also obtain the object pose in the same coordinate system. Using the same camera parameters this way alleviates the problem of trying to put everything in a common coordinate system. We preferred Depth Anything 3 to obtain the camera parameters and depth images/pointmaps as it is the current state-of-the-art depth prediction method that is not limited to a static camera. Note that even though Dyn-HaMR predicts accurate 2D hand projections, hand's 3D position, specifically along depth direction (z-axis), might be off because of the scale ambiguity in monocular capture. That is why hand position is among the optimized variables in Stage 3.

Gemini prompt for contact flags. We utilize Gemini-3-Flash [2] model to analyze the input hand–object interaction video and estimate whether the hand and object are in contact at each frame. The resulting binary contact signal is used to activate or deactivate the contact loss during Stage 3 optimization. This mechanism ensures that the contact constraint is applied only when a hand–object interaction is detected. Consequently, when the hand is not in contact with the object, the optimization does not artificially pull the object toward the fingertip vertices, allowing it to remain at its original position. The following prompt is used for this purpose:

Listing 1.1: Prompt used for Gemini contact detection

```
prompt = (
    "Analyze this video frame-by-frame and determine EXACTLY when the
        hand transitions "
    "from NOT holding (0) to HOLDING (1).\n\n"

    "CRITICAL RULES:\n"
    "1. DO NOT assume grasping at 0.0s unless the very first visible
        frame clearly shows "
    "the fingers already fully wrapped around the object.\n"
    "2. Touching the object is NOT grasping.\n"
    "3. Reaching toward the object is NOT grasping.\n"
    "4. Grasping begins ONLY at the FIRST frame where the fingers visibly
         enclose "
    "the object (clear curvature of fingers around it).\n"
    "5. If uncertain, DELAY the start time until enclosure is visually
        unambiguous.\n"
    "6. The start time must correspond to a specific visible frame change
        .\n\n"

    "Definitions:\n"
    "- REACHING (0): Hand moving toward object or touching without
        enclosure.\n"
    "- GRASPING (1): Fingers clearly wrapped around object OR object is
        clearly "
    "being lifted while enclosed.\n\n"

    "Step 1: Briefly describe the timeline of states in order.\n"
    "Step 2: Identify the FIRST exact timestamp where grasping (1) begins
        .\n"
    "Step 3: Identify when grasping ends (if it ends).\n\n"

    "Then return ONLY a valid JSON object inside a ‘‘‘json block.\n"
)
```

D Evaluated Sequences

We evaluate our method on sequences from HO3Dv3 [5] and HOT3D [1]. Specifically, we show which sequences and which timestamps used in Tab. 2 and Tab. 3.

Table 2: HO3Dv3 sequences for hand-object reconstruction. We directly use the same sequences from HOLD [3]. All of these sequences are only right-handed.

Object	Sequence name
bleach	ABF12
bleach	ABF14
potted meat	GPMF12
potted meat	GPMF14
cracker box	MC1
cracker box	MC4
power drill	MDF12
power drill	MDF14
sugar box	ShSu10
sugar box	ShSu12
mustard	SM2
mustard	SM4
mug	SMu1
mug	SMu40
banana	BB12
banana	BB13
scissors	GSF12
scissors	GSF13

E Comparison Details

For HOLD [3], we employ the authors' publicly available pre-trained CVPR models for HO3D and use the available code to train for HOT3D. For BIGS [7], we use the official implementation from the project repository and optimize each stage using the hyperparameters specified in their Supplementary Material (Sec. A). To get meshes out of their optimized Gaussians, we employ Poisson Reconstruction with depth 6 after we remove the outliers within 100 neighbors with 2.0 standard deviation. For MagicHOI [9], we follow the official implementation from the project repository.

F Detailed Results Discussion

GraG differs fundamentally from recent hand–object reconstruction methods such as HOLD, BIGS, and MagicHOI, which rely on COLMAP/SfM-based initialization followed by heavy per-scene optimization. These methods depend on

Table 3: HOT3D clips used in our evaluations, reported with sequence identifiers, hand side, temporal boundaries, and object IDs. Temporal boundaries are reported as HOT3D timecode timestamps (TS) in nanoseconds.

Sequence name	**Hand side**	**Start TS** (ns)	**End TS** (ns)	**Object ID**
P0001_23fa0ee8	right	55003474858015	55015408224119	238686662724712
P0001_8d136980	right	54039408208193	54044341521984	238686662724712
P0001_b2bcbe28	left	54811408205972	54814874891157	258906041248094
P0001_f71fc9b1	right	66482274870136	66485208197200	249541253457812
P0010_8ff3e5c4	left	58695141597662	58697074869546	258906041248094
P0011_0c2d00ed	right	55563708197241	55572041548328	70709727230291
P0011_0c2d00ed	right	55660041543187	55661741528834	204462113746498
P0011_2255f410	right	56349641532049	56356641531612	249541253457812
P0011_ccc678c7	right	55922341542825	55926474871989	261746112525368
P0011_cd22f5e0	right	55211841541663	55215108194835	228358276546933
P0011_ff7cfc3a	right	55460874890915	55465808213471	225397651484143
P0011_ff7cfc3a	right	55518174857985	55522441524737	270231216246839
P0012_0a21e4c2	right	47121508235754	47124108213732	106957734975303
P0012_0a21e4c2	right	47145441536211	47157641593381	258906041248094
P0012_44c5f677	left	59321408185000	59324774850689	194930206998778
P0012_44c5f677	right	59420774854423	59423208211436	106434519822892
P0014_84ea2dcc	right	44789441528790	44791608213887	98604936546412
P0015_60573a3b	left	60819408201515	60825874885641	79582884925181

stable multi-view correspondences to estimate camera poses and coarse geometry before performing joint reconstruction. However, such initialization is often unreliable in hand–object interaction videos due to severe hand occlusions, rapid camera motion, and small object projections, which are particularly common in datasets such as HOT3D. As a result, errors in the early SfM stage can propagate to later optimization stages, leading to degraded reconstruction quality and lower success rates.

MagicHOI further combines SfM initialization with generative priors and object inpainting to complete unseen regions. While effective under carefully selected inputs, achieving strong results typically requires manual selection of reconstruction frames and inpainting viewpoints. When the full video sequence is provided directly without such curation, the optimization becomes significantly more challenging and the reconstruction quality degrades, especially for longer sequences such as those in HO3Dv3 and HOT3D. In contrast, GraG is designed to operate directly on full video sequences without manual frame selection. Our method first identifies informative keyframes and reconstructs a canonical object representation using a foundation model, while adapting SAM3D-based object pose prediction to video sequences. The canonical object is then tracked across the sequence using a compact SoG representation with temporal and contact-aware constraints. By decoupling canonical shape reconstruction from per-frame motion estimation, GraG avoids repeated large-scale scene optimization and instead performs efficient low-dimensional tracking.

Fig. 2: Effect of keyframe selection on object reconstruction. Random keyframe selection degrades the canonical reconstruction (shape/scale) compared to our balanced keyframe selection; ground truth shown for reference.

This design yields several advantages. First, GraG is substantially more robust to occlusion and unstable viewpoints, which explains its stronger performance on challenging datasets such as HOT3D where SfM-based methods often struggle. Second, reconstructing a canonical object geometry from informative keyframes preserves fine structures, making the approach more reliable for thin or partially observed objects that are difficult to recover through correspondence-based initialization. Finally, replacing heavy test-time optimization with lightweight tracking over a fixed canonical asset enables significantly faster reconstruction while maintaining higher accuracy on both HO3Dv3 and HOT3D.

G Ablations

We conduct ablations on four sequences (ABF14, SMu1, GPMF12, and ShSu10) from the HO3Dv3 sequences. Unless stated otherwise, we keep the same training schedule and hyperparameters as the full model (including all Stage 3 loss weights and the same number of optimization steps per window). Tab. 2 summarizes the effect of key components: (i) keyframe selection for canonical object reconstruction (Stage 1), (ii) video-level pose tracking with frozen shape and temporal guidance (Stage 2), and (iii) our SoG-based refinement and prior losses (Stage 3).

Dense Gaussian tracking. For the ablation "Gaussian Splatting tracking instead of SoG", we replace the compact SoG refinement with direct dense-Gaussian optimization with RGB and prior losses, using the same number of iterations as our default windowed refinement (100 iterations per window). Even under this matched iteration budget, dense Gaussian optimization is slower and less stable in our setting (notably increasing CD_h), and we observe that it often requires substantially more iterations to reach comparable alignment.

H Additional Limitation Details

Our method relies on Depth Anything 3 to provide a per-frame depth/pointmap cue. When the predicted depth for the object is severely wrong (e.g., the object

Fig. 3: Left: Estimated pointmap with incorrect object depth. **Right:** The depth error leads to incorrect object scale and translation, yielding wrong hand-object relative pose.

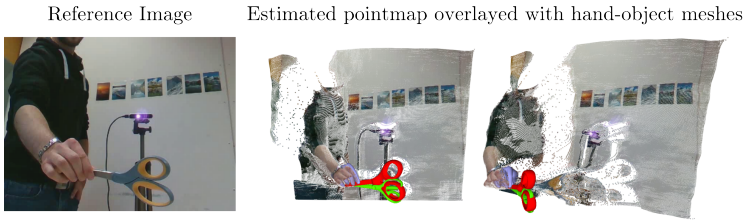

Fig. 4: Depth-scale ambiguity. We overlay the estimated pointmap with hand and object meshes. Green shows the ground-truth object; red shows our reconstruction. When the depth prior places the object too far, SoG-based fitting can compensate by inflating object scale, yielding higher CD_h even if per-frame pose is accurate.

is placed much farther from the hand than it should be under contact), the subsequent optimization can converge to an incorrect object translation and scale. In such cases, our additional priors, such as silhouette and contact losses, are sometimes insufficient to override the wrong depth signal, leading to a failed HOI reconstruction (Fig. 3).

For some sequences, the HOI reconstruction is not failed but the object's scale remains larger while keeping the rotation/translation trajectory largely consistent in image space due to errors in the depth prior used to anchor metric scale: when the estimated depth places the object farther than it truly is, the optimizer can compensate by increasing the object's scale while keeping the rotation/translation trajectory largely consistent in image space. This effect is amplified by our SoG-based tracking objective, which rewards high image-level similarity and can be satisfied by a slightly "expanded" object even under regularization. As a result, the per-frame pose (and thus the trajectory) can remain accurate, yet the reconstructed object becomes over-scaled in some sequences (as illustrated in Fig. 4), resulting in a higher CD_h.

References

1. Banerjee, P., Shkodrani, S., Moulon, P., Hampali, S., Han, S., Zhang, F., Zhang, L., Fountain, J., Miller, E., Basol, S., et al.: Hot3d: Hand and object tracking in 3d from egocentric multi-view videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7061–7071 (2025)
2. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
3. Fan, Z., Parelli, M., Kadoglou, M.E., Chen, X., Kocabas, M., Black, M.J., Hilliges, O.: Hold: Category-agnostic 3d reconstruction of interacting hands and objects from video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 494–504 (2024)
4. Fan, Z., Taheri, O., Tzionas, D., Kocabas, M., Kaufmann, M., Black, M.J., Hilliges, O.: Arctic: A dataset for dexterous bimanual hand-object manipulation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12943–12954 (2023)
5. Hampali, S., Sarkar, S.D., Lepetit, V.: Ho-3d_v3: Improving the accuracy of hand-object annotations of the ho-3d dataset. arXiv preprint arXiv:2107.00887 (2021)
6. Lin, H., Chen, S., Liew, J.H., Chen, D.Y., Li, Z., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views. arXiv preprint arXiv:2511.10647 (2025)
7. On, J., Gwak, K., Kang, G., Cha, J., Hwang, S., Hwang, H., Baek, S.: Bigs: Bimanual category-agnostic interaction reconstruction from monocular videos via 3d gaussian splatting. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 17437–17447 (2025)
8. Stoll, C., Hasler, N., Gall, J., Seidel, H.P., Theobalt, C.: Fast articulated motion tracking using a sums of gaussians body model. In: 2011 international conference on computer vision. pp. 951–958. IEEE (2011)
9. Wang, S., He, H., Parelli, M., Gebhardt, C., Fan, Z., Song, J.: Magichoi: Leveraging 3d priors for accurate hand-object reconstruction from short monocular video clips. arXiv preprint arXiv:2508.05506 (2025)