G-HOP: Generative Hand-Object Prior for Interaction Reconstruction and Grasp Synthesis

Supplementary Material

In the supplementary materials, we provide more implementation details and experimental results on the generative hand-object prior, prior-guided reconstruction, as well as prior-guided grasp synthesis. We discuss network architecture (Sec. A.1), effect of hand representation (Sec. A.2), how to extract hand pose from skeletal distance field (Sec. A.3), and the text prompt we used (Sec. A.4). Then, we show implementation details in reconstructing interaction clips and per-category results in Sec. B. Furthermore, we analyze the effect of mesh refinement in grasp synthesis and discuss comparison with prior work Grasping Field [5] in Sec. C.

A. Hand-Object Prior

A.1. Network Architecture

We use the same network architecture of latent autoencoder and 3D UNet diffusion model backbone as in SDFusion [1]. The 3D UNet backbone consists of several residual blocks. Each block is a stack of GroupNorm layer [10], non-linear activation [2], and 3D convolutional layer, with optional cross attention layer to time embedding and text embedding. We provide an overview of network details and hyperparameters of our model in Tab. 4.

	G-HOP
z-shape	$16^3 \times 3$
$\lvert\mathcal{Z}\rvert$	8196
Input Channel	3 + 15
Diffusion Steps	1000
Noise Schedule	linear
Channels	64
Number of Blocks	3
Attention resolutions	4,2
Channel Multiplier	1,2,3
Number of Heads	8
Transformers Depth	1
Batch Size	64
Iterations	500k
Learning Rate	1e-4

Table 4. **Network architecture for G-HOP.**

A.2. Ablating Skeletal Distance Field

Many previous work [6, 9] learn a diffusion model in the compact hand/human pose parameter space. We try to represent hand shape by hand pose parameters but find that this pose space is not optimal for jointly diffusing hand pose and objects in interaction. More specifically, the ablated method (pose parameter space) uses the same architecture as the main model except for the (noisy) pose parameter is passed via cross-attention layer instead of concatenating skeletal distance field to the object latent grid. We also search hyperparameters such as weights in DDPM loss to balance diffusing hand pose and diffusing object latent. We visualize the best ablated model in Fig. 13 in comparison with our proposed model that represent hand shape by skeletal distance field. The diffusion model with pose parameter space struggles to generate plausible hand articulation together with objects. This is probably because the diffusion model is hard to reason about interaction in the heterogeneous space (1D hand pose and 3D object grids).

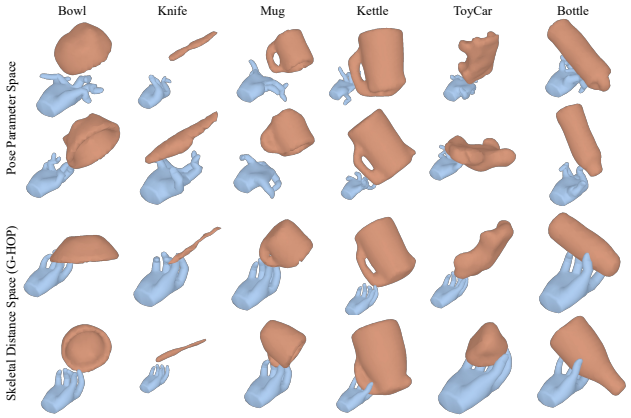

Figure 13. **Comparing Hand Representation in Generative Hand-Object Prior:** Top 2 rows show the diffusion model that represents hand shape as pose parameters; bottom 2 rows show the diffusion model (ours) that represents hand shape as skeletal distance field. The homogeneous grid space is easier for the network to reason about interaction.

A.3. Hand Pose from Skeletal Distance Field

Our proposed diffusion model generates skeletal distance field, from which hand pose parameters can be extracted. Given a target skeletal distance field $\hat{H}$, we optimize hand pose $\boldsymbol{\theta}$ such that its induced field is closer to the target, *i.e.* $\boldsymbol{\theta}* = \arg\min_{\boldsymbol{\theta}}(H(\boldsymbol{\theta}) - \hat{H})^2 + w\|\boldsymbol{\theta}\|_2^2$. We set w to 1e-5 and optimizes for 1000 steps with Adam optimizer [7] with learning rate 1e-2.

A.4. Text Prompt Template

We use the template "a hand holding a $\{category\}$" to convert category into a text prompt. In addition, we find that appending additional category attribute like size and shape beneficial when we scale up the number of category (see results in Sec. B). It may be because attributes help to transfer information between categories with similar shapes but distinct semantics, *e.g.* pens and spoons are all thin sticks. We use LLM [8] to generate attribute automatically. We list text prompt we used in Tab. 11.

B. Reconstructing Interaction Clips

Following prior work [12], we evaluate reconstruction on two sequences per category on HOI4D. We report mean performance per category in terms of object error (Tab. 5), hand error (Tab. 6), and their alignment (Tab. 7). In addition to baselines and ablations reported in main paper, we also analyze the effect of other implementation details as follows:

Dynamic Noise Threshold. The amount of injected noise in SDS has large impact on the guidance effect. We find that thin structures are better captured when adding a smaller noise while thick structure are better captured when adding larger noise. We use an adaptive noise scheduler that dynamically adjusts the maximum amount of noise $U_b, i \sim \mathcal{U}[U_a, U_b]$ based on the current object shape. More specifically, it is a linear interpolation based on minimal object SDF value in the current representation, *i.e.*

$$U_b = \frac{s - s_{\min}}{s_{\max} - s_{\min}} U_{b\max} + (1 - \frac{s - s_{\min}}{s_{\max} - s_{\min}}) U_{b\min}$$
$$s = clamp(\min O[X_{grid}], s_{\min}, s_{\max})$$

In our experiment, we set $U_{b\,\max} = 0.75, U_{b\,\min} = 0.25, s_{\min} = -0.2, s_{\max} = -0.01$. As reported in Tab. 8,, our dynamic noise threshold leads to better performance than constant noise threshold.

Scaling Up Number of Categories. For fair comparison, we use the diffusion model that only trains on HOI4D dataset to reconstruct interaction clips. In Tab. 8,, we also compare with the generalist model (G-HOP (G)) that trains on all seven datasets. Note that we use G-HOP (G) in all other experiments. We find that adding attribute to text prompt helps when scaling up to more categories. While G-HOP (G) leads to a bit worse reconstruction performance on the HOI4D dataset than the specialist which is trained only on HOI4D, it still outperforms other baselines.

2D Joint Prior. We trained a joint prior version of DiffHOI, or a 2D version of G-HOP $p(\pi(O), \pi(H)|C)$. Interestingly, we find that this cannot effectively guide grasp synthesis or reconstruction (Fig. 14, Tab. 5-7). It performs even worse than DiffHOI [12], perhaps because it is harder to learn the distribution over object, hand, and rendering viewpoints (unlike DiffHOI where the 'conditioning' informs about the hand and viewpoint).

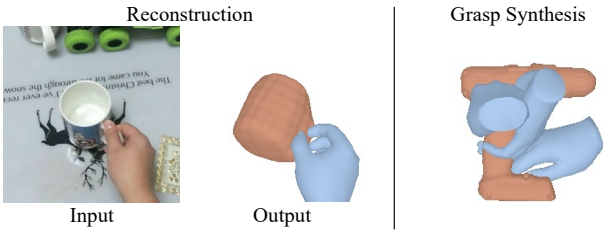

Figure 14. **2D Joint Prior (DiffHOI-J):** reconstruction and grasp synthesis results guided by 2D joint prior.

C. Grasp Synthesis

Comparison with Grasping Field. Grasping Field [5] is a representative method that uses a conditional VAE to generate hand surface distance field given an object point cloud. Their evaluation setup generates grasps for known object pose with respect to hand. We evaluate G-HOP under their setup by only optimizing hand articulation while keeping the relative pose as the given ground truth. We denote this setting with known object pose as *. G-HOP also benefits from well initialized object poses as contact ratio increases to 100%. Our contact area reduces probably because the given object pose are obtained from GT grasps that uses more finger tips and this makes the human hand palm harder to make contact. We also show that randomizing the relative pose (our evaluation setup) significant affects their performance, as visualized in Fig. 15. Note that GF gets large intersection volume but less intersection depth. This is because the latter is only calculated on each hand vertices inside of the object. For example, in the second row of Fig. 15, the knife penetrates hand, leading to high volume. But the maximum intersection depth for each hand vertices is less than the thickness of the knife.

Effect of Refinement After optimizing human grasps with respect to SDS loss using object SDF grid, we also do a light-weight mesh refinement by replacing the object SDF grid with the original mesh. It is to account for loss of accuracy during mesh conversion. We use the same objectives in previous work [3, 11] that encourage contact and discourage penetration. We denote the generated grasps before mesh refinement as † and report its performance on two datasets in Tab. 10. Even without mesh refinement, the generated grasps also have large contact area and less displacement in simulation. The refinement process can adjust hand

Table 5. **Comparing Object Error of HOI Reconstruction on HOI4D.**

	Mug			Bottle			Kettle			Bowl			Knife			ToyCar			**mean**		
	F5↑	F10↑	CD↓	F5↑	F10↑	CD↓	F5↑	F10↑	CD↓	F5↑	F10↑	CD↓	F5↑	F10↑	CD↓	F5↑	F10↑	CD↓	F5↑	F10↑	CD↓
iHOI [11]	0.44	0.71	2.1	0.47	0.77	1.5	0.21	0.45	6.3	0.38	0.64	3.1	0.33	0.68	2.8	0.66	0.95	0.5	0.42	0.70	2.7
HHOR [4]	0.18	0.37	6.9	0.26	0.56	3.1	0.12	0.30	11.3	0.31	0.54	4.2	0.71	0.93	0.6	0.26	0.59	1.9	0.31	0.55	4.7
DiffHOI [12]	0.64	0.86	1.0	0.54	0.92	0.7	0.43	0.77	1.5	0.79	0.98	0.4	0.50	0.95	0.8	0.83	0.99	0.3	0.62	0.91	0.8
G-HOP	0.62	0.93	0.7	0.93	1.00	0.2	0.64	0.96	0.6	0.66	0.96	0.5	0.91	0.99	0.2	0.78	0.98	0.3	0.76	0.97	0.4
G-HOP(Cond)	0.57	0.87	1.0	0.74	0.98	0.4	0.46	0.83	1.3	0.47	0.84	1.1	0.95	1.00	0.1	0.74	0.98	0.4	0.66	0.92	0.7
G-HOP(2D)	0.54	0.80	1.3	0.26	0.58	2.5	0.46	0.85	1.1	0.35	0.57	6.4	0.21	0.68	1.9	0.79	0.97	0.3	0.43	0.74	2.3

Table 6. **Comparing Hand Error of HOI Reconstruction on HOI4D.**

	Mug		Bottle		Kettle		Bowl		Knife		ToyCar		**mean**	
	MPJPE↓	AUC↑	MPJPE↓	AUC↑	MPJPE↓	AUC↑	MPJPE↓	AUC↑	MPJPE↓	AUC↑	MPJPE↓	AUC↑	MPJPE↓	AUC↑
iHOI [11]	1.10	0.78	1.09	0.78	1.11	0.78	1.23	0.76	1.39	0.72	1.20	0.76	1.19	0.76
DiffHOI [12]	1.06	0.79	1.01	0.80	1.07	0.79	1.21	0.76	1.33	0.73	1.04	0.79	1.12	0.78
G-HOP	1.02	0.80	0.97	0.81	0.98	0.81	1.09	0.78	1.20	0.76	1.02	0.80	1.05	0.79
G-HOP(Cond)	1.08	0.78	1.06	0.79	1.09	0.79	1.18	0.76	1.34	0.73	1.11	0.78	1.14	0.77
G-HOP(2D)	1.10	0.78	0.97	0.81	1.06	0.79	1.24	0.75	1.24	0.75	1.07	0.79	1.11	0.78

Table 7. **Comparing Hand-Object Alignment (CD_h ↓) of HOI Reconstruction on HOI4D.**

	Mug	Bottle	Kettle	Bowl	Knife	ToyCar	**mean**
iHOI [11]	19.7	13.9	35.9	49.3	21.9	21.6	27.1
HHOR [4]	229.1	172.0	100.4	50.1	185.1	255.8	165.4
DiffHOI [12]	18.1	15.3	42.2	101.8	91.6	23.3	48.7
G-HOP	12.4	9.7	41.8	26.2	13.2	7.5	18.4
G-HOP(Cond)	10.2	6.9	40.7	10.4	39.1	8.5	19.3
G-HOP(2D)	14.5	33.6	61.6	71.0	141.7	38.8	60.2

Table 8. **Additional Ablation Studies of HOI reconstruction:** We report object error (F@5mm, F@10mm, CD), hand-object alignment CD_h, and hand error (MPJPE, AUC) on HOI4D. We analyze the effect of other implementation details, including dynamic noise thresholding and choice of text prompt templates.

	Object Error			Align	Hand Error	
	F5↑	F10↑	CD↓	CD_h ↓	MPJPE↓	AUC↑
G-HOP	0.76	0.97	0.4	18.4	1.05	0.79
$U_b = 0.25$	0.69	0.95	0.5	50.0	1.01	0.80
$U_b = 0.75$	0.49	0.76	4.0	48.1	1.06	0.79
G-HOP (G) w/ attr	0.65	0.92	0.7	17.8	1.06	0.79
G-HOP (G) wo/ attr	0.61	0.89	0.8	24.6	1.04	0.79

pose to further improves the contact and grasp stability.

C.1. User Study Interface

Fig. 17 shows the user interface for evaluating the generated grasps. Users are presented two grasps visualized from different view angles as gif and are asked to choose the more plausible grasps.

Table 9. **Comparison with Baselines:** We compare human grasp synthesis along with prior work GF [5]. * denotes GF's evaluation setting with known object pose.

		Intersection			Disp.	Contact	
		max D ↓	avg D ↓	vol ↓	avg ↓	ratio ↑	area ↑
ObMan	GF [5]*	0.56	0.44	6.05	2.07	0.89	0.06
	G-HOP*	0.97	0.70	6.39	2.03	1.00	0.13
	GF [5]	0.79	0.64	43.35	1.82	1.00	0.09
	G-HOP	0.74	0.51	17.40	1.85	0.93	0.25

Table 10. **Effect of Refinement:** We report human grasp synthesis before and after mesh refinement. G-HOP† denotes generated grasps before mesh refinement.

		Intersection			Disp.	Contact	
		max D ↓	avg D ↓	vol ↓	avg ↓	ratio ↑	area ↑
ObMan	G-HOP†	0.74	0.57	8.25	3.87	0.82	0.12
	G-HOP	0.74	0.51	17.40	1.85	0.93	0.25
HO3D	G-HOP†	1.84	0.31	11.46	0.95	1.00	0.23
	G-HOP	2.42	0.68	7.55	2.48	0.99	0.20

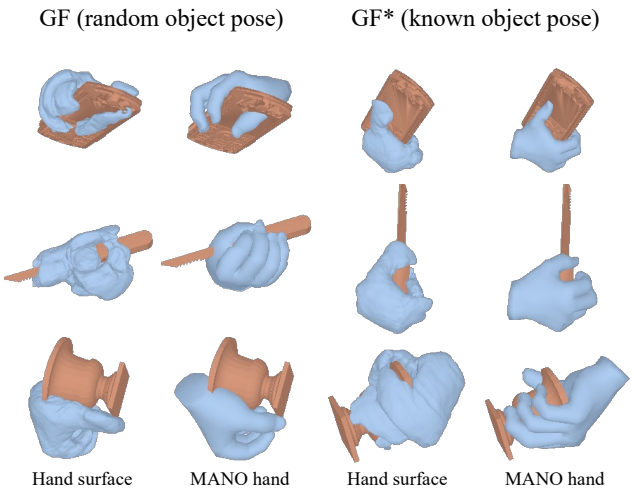

Figure 15. GF assumes known object pose when evaluating. Randomizing object pose affects their performance.

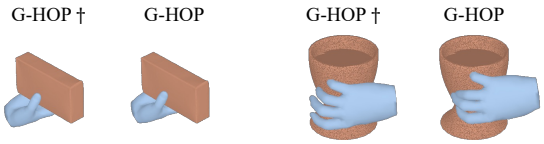

Figure 16. **Effect of Mesh Refinement:** We visualize synthesized grasps before (G-HOP†) and after (G-HOP) refinement.

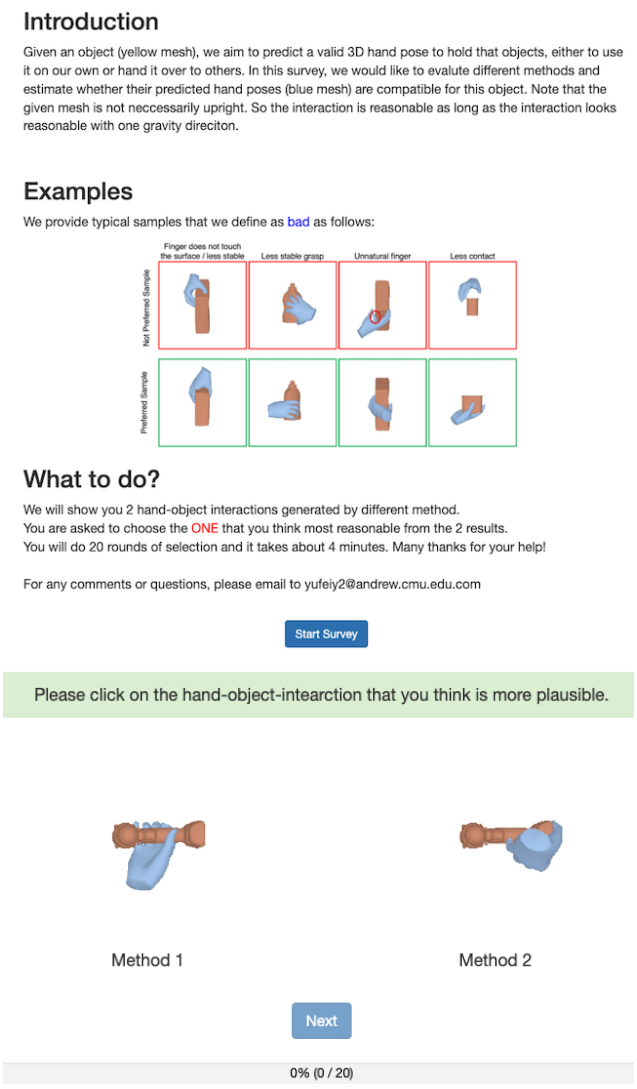

Figure 17. **User Study Interface:** We visualize user study interface including the user instruction page and the survey page.

Table 11. We provide list of class names and their attributes used in the text prompt. The class names are manually merged across different datasets while the attributes are automatically generated by large language model [8].

Class	Attribute
plate	medium, flat, circular
baseboard	big, long, rectangular
stamp	small, flat, square
laptop	big, flat, rectangular
funnel	medium, conical
spatula	medium, flat, elongated
pear	small, pear shaped
lemon	small, oval
stick	varies, cylindrical, long
cylinder	varies, cylindrical
mug	medium, cylindrical, handle attached
flute	medium, cylindrical, long
shield	big, curved, oval or round
floor	big, flat, rectangular or irregular
mouse	medium, oval, handheld
fish	varies, animal shaped
screw driver	medium, cylindrical, elongated
pen	small, cylindrical, elongated
hair dryer	medium, elongated, handheld
burger	medium, cylindrical, stacked layers
paint roller	medium, cylindrical, handheld
power saw	big, elongated, handheld or standalone
bottle	medium, cylindrical, narrow neck
pump	varies, mechanical, various shapes
flask	medium, cylindrical or conical, narrow neck
sheet	big, flat, rectangular
hand bag	medium, varies, handle attached
stapler	medium, rectangular, handheld
gummy	small, animal or object shaped
fork	small, elongated, tines at one end
wood	varies, solid, various shapes
chopsticks	small, cylindrical, elongated
strawberry	small, heart-shaped
cupmod	medium, cylindrical, handle attached
spray	medium, cylindrical, nozzle at top
crate	big, cuboid, open structure
microwave	big, rectangular, box-like
headphone	medium, round or oval, worn over ears
apple	small, round, stem at top
backpack	big, varies, straps attached
brick	medium, rectangular, solid
wood plank	big, flat, rectangular
tv	big, flat, rectangular
rubiks	small, cubical, multicolored faces
carpet	big, flat, rectangular or oval
container	varies, solid, various shapes
lego	small, rectangular or square, connecting knobs
jar	medium, cylindrical or oval, lid on top
oven	big, box-like, door at front
mixer	big, varies, mechanical
train	big, cylindrical, long
teddy bear	medium, animal shaped, soft
chess rook	small, cylindrical, castle-shaped top
binoculars	medium, cylindrical, two lenses
pencil mod	small, cylindrical, elongated
knife	medium, flat, sharp edge

Continued on next column

Continued from previous column

Class	Attribute
tin	medium, cylindrical or rectangular, lid on top
light tube	medium, cylindrical, elongated
ball	small, spherical
cupcake	small, cylindrical, rounded top
spoon	small, oval or round, handle attached
chalk	small, cylindrical, elongated
light bulb	small, round, screw base
case	varies, box-like, lid or zipper
peg test	varies, varies, testing equipment
piggy bank	medium, animal shaped, slot on top
kettle	medium, rounded, spout and handle
wrench	medium, elongated, adjustable jaw
bacon	small, flat, elongated
purse	medium, varies, handle or strap
boat	big, elongated, hollow
disk	small, flat, circular
game controller	medium, ergonomic, buttons and joysticks
keyboard	medium, flat, rectangular
trowels	medium, flat, handle attached
shovel	big, flat, long handle
eye glasses	small, oval or round, frame with lenses
stanford bunny	small, animal shaped, 3D model
camera	medium, box-like, lens at front
rifle	big, elongated, barrel and stock
can	small, cylindrical, lid on top
range	big, flat or box-like, knobs and burners
toy airplane	small, aerodynamic, wings attached
cube	varies, cubical
tablet	medium, flat, rectangular
teapot	medium, rounded, spout and handle
chair	big, varies, seat and backrest
beaker	small, cylindrical, pouring lip
plum	small, round, pit inside
triangle	varies, triangular
barrel	big, cylindrical, hollow
cup	small, cylindrical, handle attached
toothpaste	small, cylindrical, tube-shaped
bag	varies, varies, handle or strap
pyramid	varies, pyramidal
dice	small, cubical, numbered faces
ruler	small, flat, rectangular
scissors	small, paired blades, handles
clamp	small, C or G shaped, screw mechanism
phone	medium, flat, rectangular
marbles	small, spherical, glass or clay
dart	small, conical, pointed tip
calculator	medium, flat, rectangular
duck	varies, animal shaped
chain	varies, interlinked, metal
bucket	medium, cylindrical, handle attached
peach	small, round, pit inside
donut	small, cylindrical, hole in center
flashlight	medium, cylindrical, light at one end
sponge	small, soft, varies
mat	medium, flat, rectangular or oval
cardboard	varies, flat, rectangular
scoop	small, semi-spherical, handle attached
block	varies, solid, cuboidal
pliers	medium, paired jaws, handles
board	big, flat, rectangular

Continued on next column

Continued from previous column

Class	Attribute
shoe	medium, foot-shaped, footwear
floor mate	varies, flat, used for cleaning
brush	varies, bristles attached, handle
alarm clock	small, circular or square, time display
hood	big, curved, worn over head
pot	medium, cylindrical, handle attached
chessboard	medium, square, 8x8 squares
pillow	medium, soft, rectangular
power drill	medium, cylindrical, elongated
marshmallow	small, cylindrical or cubic, soft
bowl	medium, round, hollow
tube	varies, cylindrical, hollow
frisbee	medium, flat, circular
hammer	medium, heavy head, handle attached
toothbrush	small, bristles at end, handle
toycar	small, car shaped, wheels attached
elephant	big, animal shaped
tray	medium, flat, raised edges
box	varies, cuboidal, lid or flaps
book	medium, flat, rectangular
skillet lid	medium, flat or domed, handle on top
table	big, flat, supported by legs
banana	small, curved, elongated
padlock	small, rounded or square, shackle on top
bin	big, cylindrical or cuboidal, open top
blender	medium, cylindrical, mechanical
pitcher	medium, cylindrical, handle and spout
toilet	big, bowl-shaped, plumbing fixture
wine glass	small, stemmed, conical
towel	big, flat, rectangular
vacuum	big, cylindrical, mechanical
chips	small, flat, round or oval
orange	small, round, citrus fruit
microphone	small, cylindrical, handheld
usb stick	small, rectangular, electronic
door knob	small, round, mounted on door
fryingpan	medium, flat, round
watch	small, round, straps attached
eraser	small, rectangular or cylindrical, soft

Concluded

References

[1] Yen-Chi Cheng, Hsin-Ying Lee, Sergey Tulyakov, Alexander G Schwing, and Liang-Yan Gui. Sdfusion: Multimodal 3d shape completion, reconstruction, and generation. In *CVPR*, 2023. 1

[2] Stefan Elfwing, Eiji Uchibe, and Kenji Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural networks*, 2018. 1

[3] Yana Hasson, Gül Varol, Dimitris Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *CVPR*, 2019. 2

[4] Di Huang, Xiaopeng Ji, Xingyi He, Jiaming Sun, Tong He, Qing Shuai, Wanli Ouyang, and Xiaowei Zhou. Reconstructing hand-held objects from monocular video. In *SIGGRAPH Asia*, 2022. 3

[5] Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael J Black, Krikamol Muandet, and Siyu Tang. Grasping field: Learning implicit representations for human grasps. In *3DV*, 2020. 1, 2, 3

[6] Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. Guided motion diffusion for controllable human motion synthesis. In *ICCV*, 2023. 1

[7] Diederik P Kingma, J Adam Ba, and J Adam. Adam: A method for stochastic optimization. *ICLR*, 2015. 1

[8] OpenAI. Gpt-4 technical report, 2023. 2, 5

[9] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *ICLR*, 2023. 1

[10] Yuxin Wu and Kaiming He. Group normalization. In *ECCV*, 2018. 1

[11] Yufei Ye, Abhinav Gupta, and Shubham Tulsiani. What's in your hands? 3d reconstruction of generic objects in hands. In *CVPR*, 2022. 2, 3

[12] Yufei Ye, Poorvi Hebbar, Abhinav Gupta, and Shubham Tulsiani. Diffusion-guided reconstruction of everyday hand-object interaction clips. In *ICCV*, 2023. 2, 3