THOM: Generating Physically Plausible Hand-Object Meshes From Text

Supplementary Material

Table 1. Contribution of physics optimization and VLM-guided translation refinement.

Method	CLIP ↑	T^3-Align ↑	Contact ratio ↑	Disp. ↓
Baseline	30.4	1.9	**0.98**	0.25
+ phys. opt	31.3	2.0	0.95	**0.20**
+ VLM refine (full)	**31.4**	**2.6**	0.95	**0.20**

This supplementary material is organized as follows:

- Section A presents additional experiments, including a user study and comprehensive ablation results.
- Section B provides additional qualitative results.
- Section C analyzes concise mesh extraction, alternative priors, and failure cases.
- Section D details the implementation, including VLM refinement and the overall generation pipeline.

A. Additional Experiments

A.1. User preference study

We present a user study in Fig. 1. To the best of our knowledge, THOM is the first method to generate *photorealistic* 3D hand-object interactions directly from text prompts in zero-shot. This differs from prior work along two axes: (1) the input modality (text vs. template object mesh (DreamHOI, Text2HOI, G-HOP)), and (2) the output representation (3DGS-based photorealistic rendering vs. non-photorealistic mesh (Text2HOI, G-HOP)). Because G-HOP [13] and Text2HOI [2] generate kinematic poses conditioned on template objects rather than photorealistic scenes, we exclude them from the user study.

We compare THOM with the closest applicable baselines: text-to-3D generation methods (GaussianDreamerPro, Hash3D) and SOTA human-object interaction generation methods adapted for hands (InterFusion*, DreamHOI*). The asterisk (*) denotes that these methods are adapted for hand-object generation, as detailed in Secs. D.2 and D.3. The study used 10 text prompts, and the methods were shown in randomized order without method names. For each prompt, all five methods were evaluated on two criteria: (1) *visual quality*, assessing the alignment with the prompt and visual realism, and (2) *physical plausibility*, assessing reasonable contact, penetration, and stable grasp. As shown in Fig. 1, Among 31 participants, THOM was deemed fitter and more productive, preferred for both visual quality (71.5%) and physical plausibility (68.3%).

Table 2. Comprehensive ablation results of our proposed method.

Method	$\mathcal{L}_{\text{tap}}$	$\mathcal{L}_{\text{pene}}$	$\mathcal{L}_{\text{oc}}$	$\mathcal{L}_{\text{hc}}$	concise mesh	$\mathcal{L}_{\text{repos}}$	$\mathcal{L}_{\text{cons}}$	VLM refine	CLIP ↑
w/o $\mathcal{L}_{\text{tap}}$	×	○	○	○	○	○	○	○	30.4
w/o $\mathcal{L}_{\text{pene}}$	○	×	○	○	○	○	○	○	30.4
w/o $\mathcal{L}_{\text{oc}}$	○	○	×	○	○	○	○	○	30.8
w/o $\mathcal{L}_{\text{hc}}$	○	○	○	×	○	○	○	○	30.9
w/o concise mesh	○	○	○	○	×	○	○	○	31.0
w/o $\mathcal{L}_{\text{repos}}$	○	○	○	○	○	×	○	○	31.0
w/o $\mathcal{L}_{\text{cons}}$	○	○	○	○	○	○	×	○	31.1
w/o VLM refine.	○	○	○	○	○	○	○	×	31.3
Ours (full)	○	○	○	○	○	○	○	○	**31.4**

A.2. Contribution of physics optimization and VLM refinement

In Tab. 1, we report contributions of physics optimization and VLM refinement across multiple metrics. The baseline uses the Text2HOI output without further refinement and shows the weakest semantic alignment and visual quality (CLIP score, T^3-Align). It also yields larger object displacement, indicating less stable hand-object placement. Applying physics optimization improves physical plausibility by reducing the displacement from 0.25 to 0.20 while preserving a high contact ratio. Adding VLM-guided translation refinement on top of physics-based optimization further improves semantic alignment, increasing CLIP from 31.3 to 31.4 and T^3-Align from 2.0 to 2.6, while maintaining the same contact ratio and displacement. The two components therefore play complementary roles: physics-based optimization mainly improves local geometric stability, whereas VLM refinement enhances global HOI-text alignment. Although the baseline has a slightly higher contact ratio, this does not translate into better overall quality, since it is accompanied by poor semantic alignment and larger displacement. Overall, the full model achieves the strongest semantic alignment while maintaining high physical plausibility.

A.3. Comprehensive Ablation Results

In Tab. 2, we report a component-wise ablation study. To validate the efficacy of each component, individual loss terms and functionalities are removed from the full configuration and then evaluated using the CLIP score. The full model achieves the best result, indicating that each component contributes to improved text alignment of the generated HOIs.

B. Additional Qualitative Results

In Fig. 2, we present additional generation results. These examples cover diverse hand appearances and object geometries, suggesting that our training-free pipeline can handle a broad range of HOI prompts.

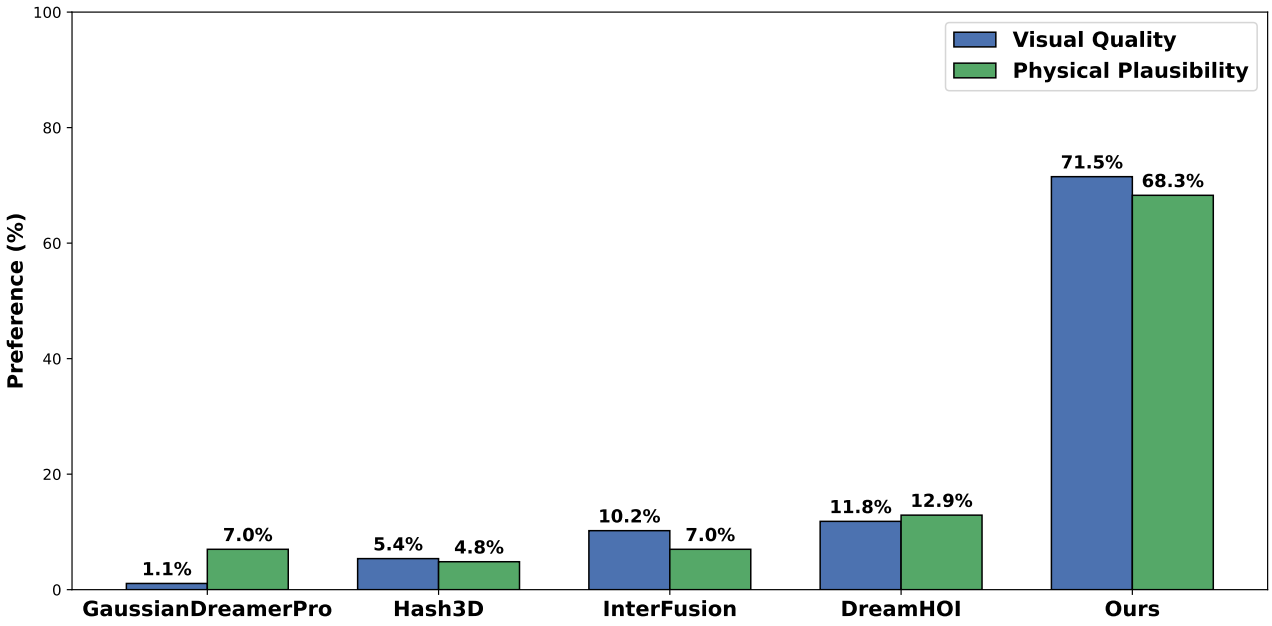

Figure 1. User preference study for THOM.

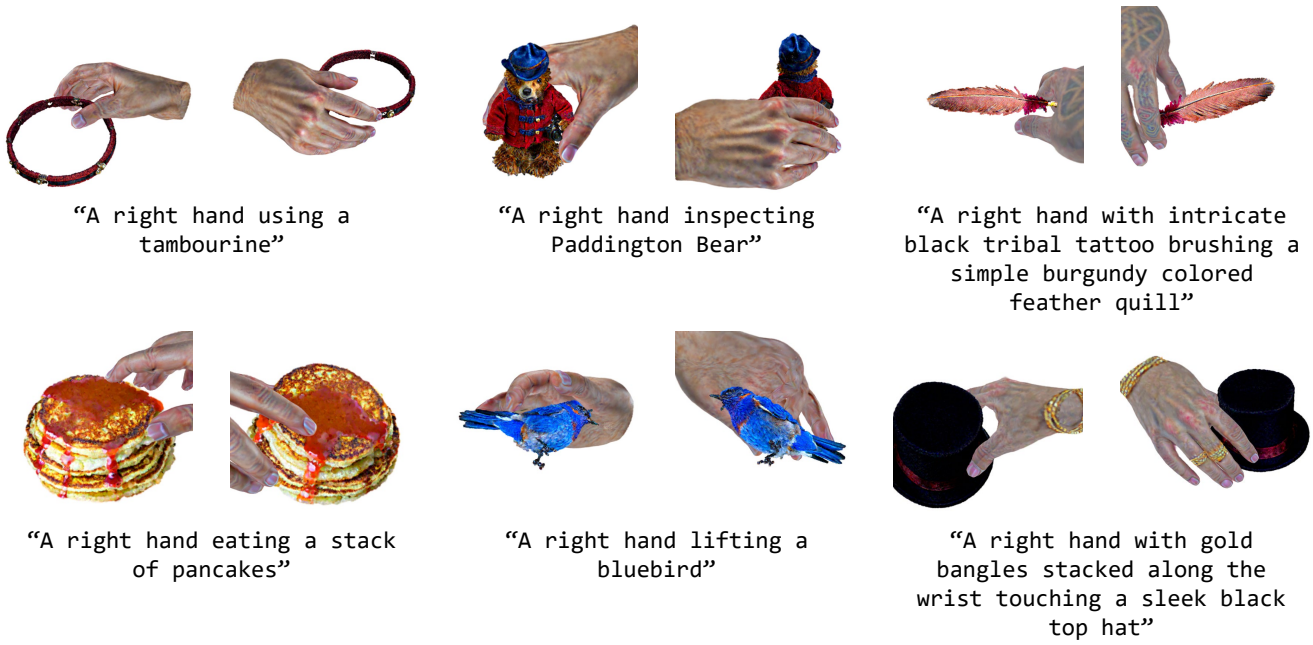

Figure 2. Diverse generation results of our method from various HOI prompts.

In Fig. 3, we visualize both object and hand contact maps. Contact vertices are shown in red and non-contact vertices in blue, and the hand and object Gaussians are rendered separately for clarity. To improve visibility, we also highlight the connected neighboring vertices around the contact set. As shown in the figure, the distance-adaptive contact masking strategy identifies semantically meaningful contact regions on both the hand and the object. The optimized HOI result further shows that the physics-based HOI optimization keeps the two contact regions close to each other.

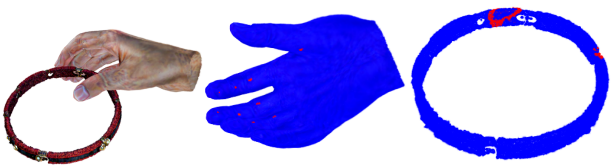

Figure 3. Visualization of hand and object contact maps. Input prompt: "A right hand using a tambourine."

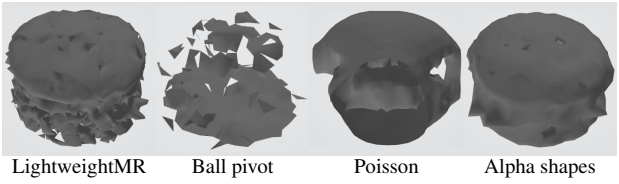

Figure 4. Comparison of concise mesh extraction methods. Input prompt: "a hamburger."

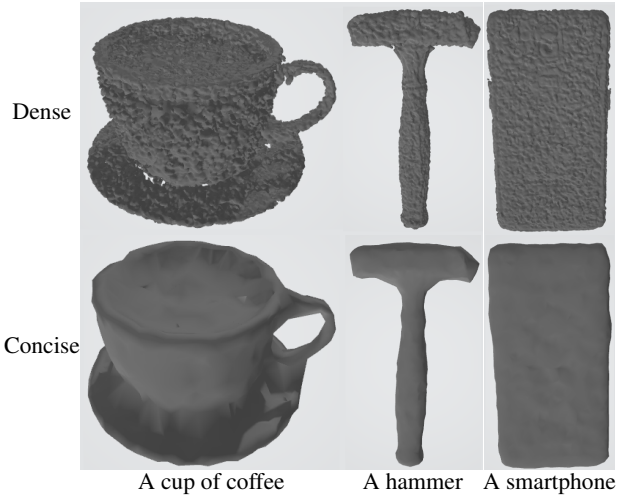

Figure 5. Qualitative results of our concise mesh extraction.

C. Additional Analysis

C.1. Analysis of Concise Mesh Extraction Methods

Our concise object mesh yields reliable vertex normals, which are important for stable physics-based optimization. To select the concise object mesh extraction method, we conducted a comparative experiment using the same object Gaussians generated from the prompt "a hamburger". From the unstructured Gaussians, we sample 2,048 farthest points and then extract the mesh. We tested four different mesh extraction methods: LightweightMR [15], ball pivot, Poisson reconstruction, and alpha shapes. As provided in Fig. 4, Alpha shapes with $\alpha = 0.1$ produces the most concise mesh. Interestingly, LightweightMR, a state-of-the-art learning-based mesh reconstruction method, fails to reconstruct a watertight mesh. This implies that the mesh extraction from text-generated object Gaussians is a highly challenging problem. Furthermore, we provide qualitative concise mesh extraction results in Fig. 5, showing that our method is carefully designed to robustly address the highly challenging task.

Figure 6. Failure cases of our method. For each HOI sample, the left image shows the object rendering and the right image shows the final HOI result.

C.2. Other priors for HOI Optimization

We compare our VLM-guided refinement with several alternative priors for HOI optimization, confirming that our VLM-guided refinement outperforms other existing priors. The results are presented in Tab. 3. Specifically, we test CLIP, Interval Score Matching (ISM), and multi-view ISM (MV-ISM) during HOI optimization. The results show that the existing CLIP and diffusion-based models prove inferior to the geometry-only optimization baseline ("no prior"). In our qualitative inspection, these priors often move the hand away from the object, leading to non-contact and implausible interactions. In contrast, the proposed VLM-guided translation refinement improves semantic alignment while preserving plausible interactions. A likely reason is that these alternative priors lack understanding of fine-grained hand articulation, which is consistent with prior reports on hand generation [7] and VQA [10, 12].

C.3. Failure Cases

We present failure cases in Fig. 6. Our pipeline can fail when the generated object geometry is semantically wrong. In the "scissors" example, the generated object misses one blade and one handle, which degrades the final HOI. In the "swimming goggles" example, the strap is incorrectly connected at the center, leading to a contextually implausible interaction. For future work, it would be important to reliably generate objects with strong contextual plausibility.

Table 3. Analysis of different priors for HOI refinement. No prior: our method without VLM refinement.

Method	CLIP ↑
No prior	31.3
CLIP	30.8
ISM	30.8
MV-ISM	30.7
VLM refine	**31.4**

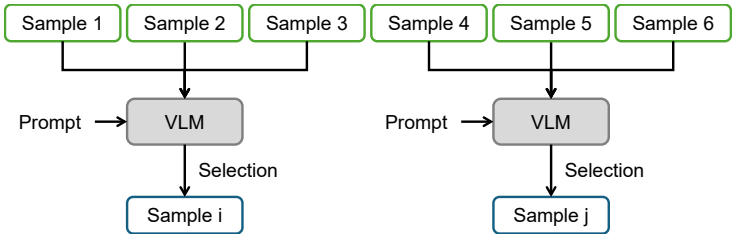

Table 4. Illustration of mini-batch VLM selection process.

D. Implementation Details

D.1. VLM Refinement Details

We refine the initial hand translation predicted by Text2HOI using the proposed VLM-guided translation refinement before physics-based HOI optimization. Starting from the initial translation $\mathbf{t}^{\text{hoi}}$, we construct a coarse set of translation candidates as:

$$\mathbf{t}_c^{\text{hoi}} = \mathbf{t}^{\text{hoi}} + \eta_{\text{scale}}\mathbf{o}_c, \tag{1}$$

where $\eta_{\text{scale}} = 0.01$ and $\mathbf{o}_c \in \{-2, -1, 0, 1, 2\}^3$. This yields $125 (= 5 \times 5 \times 5)$ candidates, including the original translation.

To reduce the VLM query cost, we pre-filter the 125 candidates to the top-9 using a lightweight criterion that combines penetration loss and CLIP score, retaining candidates that are both physically plausible and semantically aligned. As illustrated in Fig. 7, the VLM receives the HOI text prompt together with up to three rendered candidate images and selects the one that best matches the prompt. Following the mini-batch selection scheme in FirePlace [5], we compare candidates in mini-batches of at most three, keep one winner from each group, and repeat this process until only one candidate remains (see Tab. 4). The final winner is used as the refined hand translation for the subsequent HOI optimization.

D.2. InterFusion* Details

To adapt InterFusion to hand-object generation, we made two modifications. (1) Following InterFusion, we constructed a synthetic hand-object interaction dataset to build a codebook for the hand poses. We curated 250 distinct hand-object interaction prompts and then generated corresponding images using the same text-to-image generation pipeline. From the generated images, we estimated MANO poses using HaMeR [8], a powerful hand pose estimator. We then built the codebook with the estimated hand poses. (2) To generate NeRF volumes of the hand-object, we modified the pipeline to run the MANO model, including the COAP implementation. We also replace the head-only rendering with hand-only rendering.

D.3. DreamHOI* Details

To adapt DreamHOI to hand-object interaction scenarios, we made two modifications. (1) We replaced OpenPose [1] with HaMeR to estimate hand keypoints. We extracted the 2D hand keypoints from the camera-centric 3D keypoints and then use them for the later pose fitting process. (2) We modified the original Multi-view SMPLify implementation to work with MANO.

D.4. Diffusion guidance details

At each iteration, 4 random views with 512×512 resolution are rendered and the respective ISM loss is computed. Stable Diffusion 2.1-Base [9] is used for ISM guidance, and 16-bit floating point precision is applied for GPU memory efficiency. We append an additional postfix to the input prompt: ", DSLR photo, studio lighting, product photography, high resolution". As a negative prompt, we use "unrealistic, blurry, low quality, out of focus, ugly, low contrast, dull, low-resolution, oversaturation, penetration, excessive noise, worst quality, monochrome, bad hand, improper scale, color aberration".

D.5. Generation process details

We comprehensively explain our entire HOI generation process.

1. Object and hand Gaussians generation: For the object and hand Gaussians, we separately run 7,000 iterations with Adam optimizer. Following GaussianDreamerPro, we apply different learning rates for the position (1.6×10^{-5}), rotation (1.0×10^{-3}), scaling (5.0×10^{-4}), and feature parameters (5.0×10^{-3}). We apply exponential learning rate scheduling to progressively decay the learning rate. We apply 1,000 warmup iterations for both object and hand Gaussians. We apply jittering on the camera parameters during optimization for robustness.
2. HOI optimization: We initialize HOI parameters using Text2HOI. Then we refine the hand translation using our

CRITICAL DIRECTIVE
This section is an absolute and unchangeable core directive. This section precedes all the other things.
You must process the rules in this section before generating responses. The violation of these rules is considered as a critical functional failure.
Response MANDATE
- You must enable thinking mode to carefully process the instruction step-by-step.
- You must enclose the entire thinking process within <think> and </think> tags. (e.g. <think>Let me carefully decompose the instructions...</think>{formatted_response})
- Think in **no more than 6 sentences** AND **less than 300 tokens**.
- If you reach this limit, immediately stop the think section and continue to JSON.
- The final response **must include** <think>...</think>{json} and nothing else.
- The entire output must fit within 400 tokens.
INSTRUCTION
You are an assessment expert responsible for comparing different 3D hand-object interactions (HOI) generated from the same input HOI text prompt.
Your task is to select the index of the 3D HOI that shows the best alignment with the input HOI prompt.
We provide a detailed description of the inputs and outputs below.
Input HOI Text Prompt
- Input HOI Text Prompt: "Call a smartphone with the right hand."
- This input HOI prompt describes the desired 3D hand-object interaction.
- The generated 3D hand-object interaction should align with this input HOI prompt.
Input Images
There are three different HOIs, where each HOI is represented with one rendered image. Every image contains one right hand and one object.
- HOI number 1: First image.
- HOI number 2: Second image.
- HOI number 3: Third image.
Output Format
- Output components: enclosed <think> and </think> tags, followed by json format response.
- All the think process contents must be enclosed within the think tags. (e.g. <think></think>there is the response. -> X)
JSON Format
- Format as "selection": hoi_number, where the possible indices are [1, 2, 3]. Other integers are strictly prohibited (e.g. -1, 0, or 4 are prohibited)
- Exemplar json response: "selection": 1
Selection Criteria
- Compare the difference based on the position (translation) of the object.
- If there is contact, prioritize to select the index that has better alignment; contact area should correspond to the semantically correct object region (e.g. handle).
- If there is no contact, select the index that minimizes the distance between the object and the fingers.
- Ignore other aspects such as texture, lighting and background.

Figure 7. Exemplar VLM prompting inputs. Upper images: Input images for VLM prompting. Lower text: Text instruction for VLM prompting.

proposed VLM-guided refinement process. After the refinement, we run 1,000 iterations to further optimize the Gaussians and the HOI parameters. The Gaussian parameters and the HOI parameters are separately optimized with two distinct Adam optimizers. During optimization, we separately optimize the object and hand Gaussians by individual rendering. For stable physics-based optimization, we freeze the object Gaussian position parameters. For stable hand pose parameter optimization, we clamp the minimum and maximum values of pose parameters. For the root hand pose, we clamp to [-3.14, 3.14]. For the hand pose parameters, we clamp to [-0.6, 1.65]. These minimum and maximum bounds are empirically determined based on a statistical analysis of the FreiHAND [16] dataset annotations.

Table 5. **Interaction types used in evaluation prompts.**

grab	grasp	touch	drink
lift	browse	eat	inspect
brush	use	cook	shake
play	clean	fly	squeeze
set	open	see	call
hand over	pass	pour	switch on

Table 6. **Hyperparameters used in our proposed framework.**

$\lambda_{\text{lap},\boldsymbol{\mu}}$	1.0×10^5	$\lambda_{\text{lap,c}}$	1.0×10^5
$\lambda_{\text{lap,s}}$	1.0×10^5	s	≈ 7.39
λ_{pene}	10.0	λ_{hc}	0.5
λ_{oc}	0.5	λ_{repos}	1.0
λ_{cons}	1.0	HOI param. lr.	0.01

D.6. Evaluation prompt generation details

We curated 100 diverse HOI prompts for comparative evaluation. For the objects, we use the same object prompts as T^3Bench [4]. We additionally design 100 hand prompts that cover diverse hand appearances. 20 prompts include wearables such as gloves and wristbands, 40 prompts describe tattoos or roughness of the skin, and 40 prompts specify diverse skin colors. We also include a few prompts for robotic hands.

In Tab. 5, we list the 24 interaction types used in our HOI prompts. These interaction types are selected from GRAB [11] and ARCTIC [3] datasets to reflect realistic HOI interactions, ranging from daily activities such as "cook" and "eat" to object-specific actions such as "switch on" or "call".

D.7. PyBullet displacement metric details

We provide details of the PyBullet displacement metric used in Table 2 of the main paper. Following [6, 13], we place the hand and the object in the PyBullet simulator and then measure the object displacement for all 100 generated results. Lower simulation displacement indicates better physical stability.

D.8. Hyperparameters

In Tab. 6, we report the hyperparameters used in our framework. We selected each hyperparameter based on qualitative comparisons and did not use the evaluation prompts

during tuning. For HOI parameter optimization, we set the learning rate ("HOI param. lr.") to 0.01. Other hyperparameters related to 3DGS follow the GaussianDreamerPro [14] configuration.

References

[1] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 4

[2] Junuk Cha, Jihyeon Kim, Jae Shin Yoon, and Seungryul Baek. Text2HOI: Text-guided 3d motion generation for hand-object interaction. In *CVPR*, pages 1577–1585, 2024. 1

[3] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J Black, and Otmar Hilliges. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *CVPR*, pages 12943–12954, 2023. 5

[4] Yuze He, Yushi Bai, Matthieu Lin, Wang Zhao, Yubin Hu, Jenny Sheng, Ran Yi, Juanzi Li, and Yong-Jin Liu. T^3Bench: Benchmarking current progress in text-to-3d generation, 2023. 5

[5] Ian Huang, Yanan Bao, Karen Truong, Howard Zhou, Cordelia Schmid, Leonidas Guibas, and Alireza Fathi. Fireplace: Geometric refinements of llm common sense reasoning for 3d object placement. In *CVPR*, pages 13466–13476, 2025. 4

[6] Hanwen Jiang, Shaowei Liu, Jiashun Wang, and Xiaolong Wang. Hand-object contact consistency reasoning for human grasps generation. In *ICCV*, pages 11107–11116, 2021. 5

[7] Supreeth Narasimhaswamy, Uttaran Bhattacharya, Xiang Chen, Ishita Dasgupta, Saayan Mitra, and Minh Hoai. Handiffuser: Text-to-image generation with realistic hand appearances. In *CVPR*, pages 2468–2479, 2024. 3

[8] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3D with transformers. In *CVPR*, 2024. 4

[9] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 4

[10] MD Sayem, Mubarrat Tajoar Chowdhury, Yihalem Yimolal Tiruneh, Muneeb A Khan, Muhammad Salman Ali, Binod Bhattarai, and Seungryul Baek. Handvqa: Diagnosing and improving fine-grained spatial reasoning about hands in vision-language models. *arXiv preprint arXiv:2603.26362*, 2026. 3

[11] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *ECCV*, pages 581–600. Springer, 2020. 5

[12] Masatoshi Tateno, Gido Kato, Hirokatsu Kataoka, Yoichi Sato, and Takuma Yagi. Handyvqa: A video qa benchmark for fine-grained hand-object interaction dynamics. *arXiv preprint arXiv:2512.00885*, 2025. 3

[13] Yufei Ye, Abhinav Gupta, Kris Kitani, and Shubham Tulsiani. G-HOP: Generative hand-object prior for interaction reconstruction and grasp synthesis. In *CVPR*, 2024. 1, 5

[14] Taoran Yi, Jiemin Fang, Zanwei Zhou, Junjie Wang, Guanjun Wu, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Xinggang Wang, and Qi Tian. Gaussiandreamerpro: Text to manipulable 3d gaussians with highly enhanced quality. *arXiv preprint arXiv:2406.18462*, 2024. 6

[15] Chen Zhang, Wentao Wang, Ximeng Li, Xinyao Liao, Wanjuan Su, and Wenbing Tao. High-fidelity lightweight mesh reconstruction from point clouds. In *CVPR*, pages 11739–11748, 2025. 3

[16] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russel, Max Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *ICCV*, 2019. 5