

ViHOI: Human-Object Interaction Synthesis with Visual Priors

Supplementary Material

A. Overview

This supplementary material provides a comprehensive description of our approach, including method details (Section B), evaluation details (Section C), additional experiments (Section D), additional visualization results (Section E), and analysis of VLM understanding (Section F).

B. Method Details

In our framework, we adopt the same object geometry representation as used in the downstream generator. Taking the framework diagram in the main text as an example, we choose CHOIS [5] as our HOI generator, and the geometric shape of the object is encoded using Basis Point Set (BPS) [6] and then projected into a 256-dimensional embedding space through an MLP. This 256-dimensional geometric embedding is fused with the motion tokens and subsequently concatenated along the temporal dimension with both the visual and textual prior tokens. The transformer’s self-attention layers jointly process the combined sequence.

In our comparison experiments, we follow the original object-geometry handling of each downstream HOI generator. Specifically, when integrating our framework with MDM [9] and ROG [10], we adopt the exact geometry representation used in their original implementations. The MDM version we use is the one released by the ROG authors, where both models represent the object using 24 surface keypoints. These keypoints are obtained by combining two sampling strategies on the object mesh surface: 8 boundary keypoints aligned with the object’s Axis-Aligned Bounding Box [3] that capture its global extent, and 16 keypoints obtained via Poisson Disk Sampling [11] that preserve finer geometric details. This ensures that any observed performance improvements are attributable solely to our proposed priors, rather than discrepancies in geometric representation.

Regarding the training details, we strictly follow the original training recipes of the respective baselines (e.g., MDM and CHOIS). The only architectural modification is the replacement of their original CLIP text encoder with our proposed VLM and Q-Former module. To handle the scale discrepancies and distribution shifts across the end-layer features of different methods, the Q-Former adapter is jointly trained with each specific baseline generator. Because of this joint training strategy, the visual and textual prior tokens C_v and C_t are dynamically optimized. This ensures that the generated tokens seamlessly align with the specific feature distributions of each generator.

C. Evaluation Details

Currently, most feature extractors used to evaluate Human-Object Interaction (HOI) motions primarily focus on human poses, neglecting the spatial positions and rotational dynamics of the involved objects. To overcome this limitation, we draw inspiration from the T2M [2] framework and adopt a similar evaluation protocol. In our approach, a frozen CLIP text encoder [7] is employed to transform textual descriptions into feature embeddings. Meanwhile, the generated HOI motion sequences are processed using a bidirectional GRU (BiGRU) model. To ensure that the evaluation metrics accurately capture the quality of the generated motions, we adjust the BiGRU model’s input dimensionality to meet the parameter requirements for HOI sequence visualization. Specifically, we set the input dimension to 147: the first 3 dimensions represent the root joint parameters of the human body, 132 dimensions correspond to the 6D relative rotations of 22 joints, 3 dimensions encode the object’s translation parameters, and the remaining 9 dimensions describe the object’s rotation matrix. By minimizing the feature distance between matched text–HOI pairs, our method effectively builds a robust alignment between natural language descriptions and HOI motion sequences.

D. Additional Experiments

Impact of Query Quantity on Prior Adaptor. Our previous Prior Adaptor module extracts interaction priors from a large vision–language model (VLM) using learnable queries. To analyze the impact of query quantity, we evaluate different numbers of queries. As shown in Table 1, employing a single query achieves the best performance across most metrics. This observation aligns with our design philosophy: the VLM-based prior is intended to capture a compact, global cue of human–object interaction. Introducing more queries expands the latent prior’s dimension. It forces the model to attend to multiple prior tokens simultaneously, potentially diluting the semantic signal and introducing redundant or less informative visual features. Therefore, we adopt a single-query configuration that provides stable, semantically coherent prior cues.

Impact of Text-to-Image Generation Model. During inference, our method leverages reference images generated by a text-to-image generation model [8] to provide visual prior information. Although these reference images may occasionally exhibit appearance flaws or unrealistic renderings, we observe that such visual imperfections have lit-

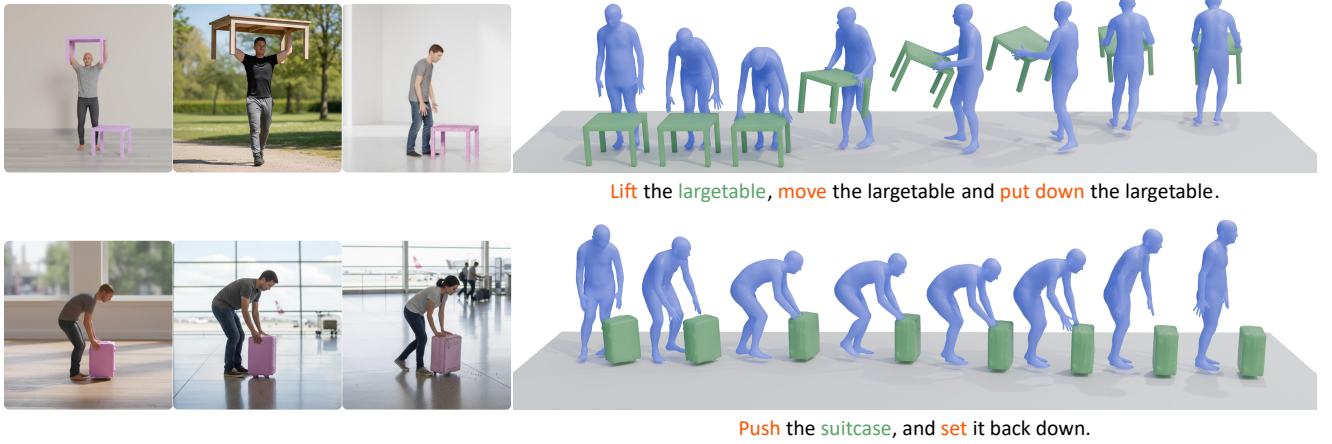


Figure 1. Qualitative results on the FullBodyManipulation dataset [4]. The three images on the left side are the reference inputs, while the right side shows the motion sequences generated from them. Despite imperfections in these reference images, the generated HOI motions remain plausible and well aligned with the textual semantics.

Method	R-score \uparrow	FID \downarrow	$C_{prec}\uparrow$	$C\%$	$P_{hand}\downarrow$	MPJPE \downarrow
k = 2	0.72	1.30	0.71	0.61	0.60	17.52
k = 4	0.63	3.07	0.78	0.58	0.59	18.38
k = 8	0.69	2.81	0.77	0.57	0.61	17.94
k = 1	0.79	0.68	0.83	0.64	0.58	14.97

Table 1. Impact of Adaptor Query Number (k) on Generation Performance. We vary only the number of visual prior queries, while keeping the number of text queries fixed at one.

Method	R-score \uparrow	FID \downarrow	$C_{prec}\uparrow$	$C\%$	$P_{hand}\downarrow$	MPJPE \downarrow
ViHOI-GT	0.79	0.29	0.82	0.64	0.59	12.94
ViHOI	0.79	0.68	0.83	0.64	0.58	14.97

Table 2. Impact of reference images. ViHOI-GT uses images rendered from GT motion, while ViHOI uses images from the T2I generation model.

tle impact on the final quality of HOI generation. We believe this is because our Prior Adaptor emphasizes capturing high-level semantic relationships within the image rather than low-level pixel details. Moreover, the textual priors extracted from the prompts offer a holistic description of the intended action, ensuring that the generated human–object interactions remain globally coherent and semantically accurate.

To assess ViHOI’s robustness to the quality of reference images, we compared its performance under two distinct settings: (1) using reference images rendered from the Ground-Truth (GT) motion from the test set, and (2) using images generated by the Text-to-Image (T2I) generation model described in Section 3.3. As shown in Tab. 2, using

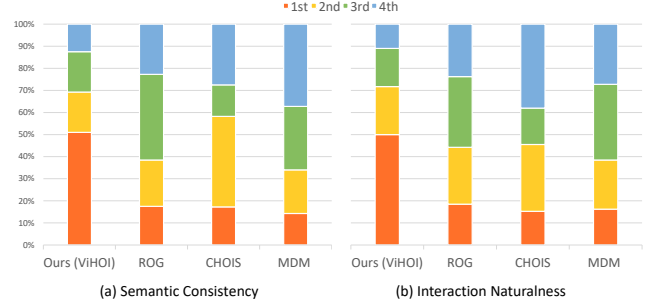


Figure 2. User study on the FullBodyManipulation dataset [4].

T2I-generated images results in a certain degree of performance drop compared to those rendered from ground-truth (GT) motions. Still, the decline remains within a reasonable and acceptable range. Crucially, even in this more challenging setting, our method still outperforms existing state-of-the-art models, demonstrating its strong adaptability to variations in image style and rendering quality. To maintain a strict evaluation protocol and prevent test data leakage, we exclusively use the T2I generation model to produce reference images in all our main comparative experiments.

Fig. 1 further provides qualitative evidence of this robustness. Even when the reference images contain noticeable artifacts or implausible visual effects, our model successfully produces accurate interaction trajectories and physically plausible contact patterns between humans and objects. These qualitative results demonstrate that our approach is highly resilient to variations in image quality, relying primarily on semantic cues rather than photorealistic fidelity.

Impact of VLM and T2I on Computational Overhead.

We evaluate the computational overhead of our framework using a single RTX 3090 GPU. The VLM introduces a

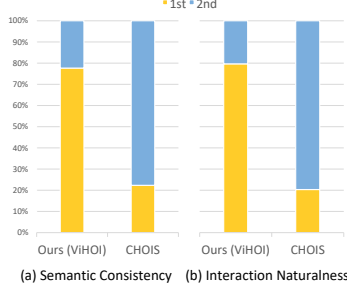


Figure 3. User study on the 3D-Future dataset [1].

marginal overhead of only 0.65s, whereas the Text-to-Image API call requires 7.20s. Although these modules introduce a certain degree of inference latency, they yield significant performance improvements. Furthermore, due to the stochastic nature of diffusion models, we emphasize that for the same textual instruction, a single generated reference image can be reused to generate multiple diverse HOI motions without repeating the time-consuming T2I process.

User Study. To evaluate the perceptual quality of our method, we conduct a user study comparing ViHOI against three baseline methods [5, 9, 10] on 20 text prompts from the FullBodyManipulation dataset [4]. Furthermore, to specifically assess generalization capabilities, we compare ViHOI against CHOIS [5] using 10 unseen objects from the 3D-Future dataset [1].

Following the evaluation protocol established by [10], we asked a total of 20 participants to rank the results according to two criteria: (1) Semantic Consistency (alignment between animations and text descriptions), and (2) Interaction Naturalness (naturalness of poses and object interactions). As illustrated in Fig. 2 and 3, our method received significantly higher ratings in both generation quality and generalization. Participants consistently favored ViHOI for its superior text-motion alignment, more plausible interactions, and remarkable ability to generalize to unseen objects.

E. Additional Visualization Results.

To demonstrate the diversity and effectiveness of our approach, we present additional HOI generation results across various scenarios.

More Generation Results. We present additional generation results in Fig. 4, demonstrating the quality of our method across different actions and object categories, and include a detailed comparison and demonstration with three baseline methods [5, 9, 10] on the FullBodyManipulation dataset [4] in our accompanying video. These examples highlight ViHOI’s ability to generate natural and semantically accurate interactions across different actions and object categories.

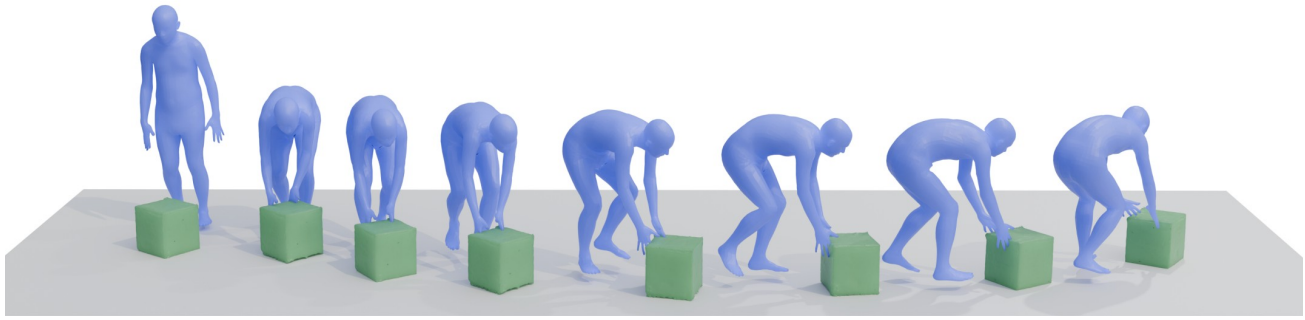
More Results on Unseen Objects. We present additional results on unseen objects from the 3D-FUTURE dataset [1] in Fig. 5 and include detailed comparison and demonstration with CHOIS [5] in the accompanying video. These examples show that ViHOI can maintain natural and accurate interaction generation even when encountering previously unseen objects.

F. Analysis of VLM Understanding

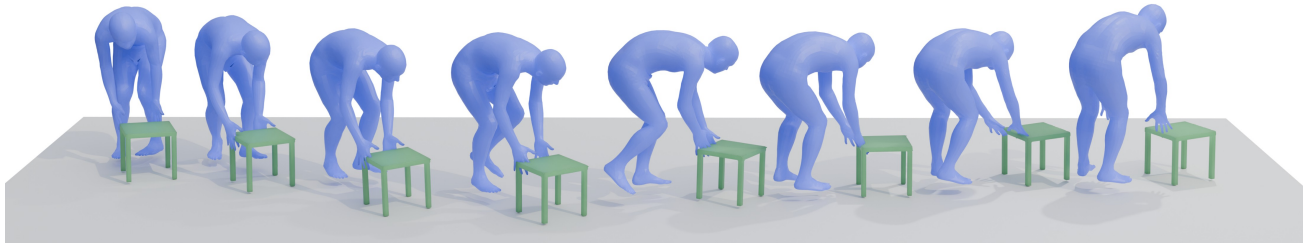
To better illustrate the semantic capacity of the VLM we use for prior extraction, we present several examples of its autoregressive textual outputs given our rendered reference images and prompts. Although our method only uses intermediate-layer embeddings rather than the final decoded text, these outputs demonstrate that the VLM reliably captures high-level human-object relations. This supports the rationale behind using VLM embeddings as interaction priors in our model. As illustrated in Fig. 6 to 7, the VLM is able to infer meaningful priors—such as the human’s standing posture, the height and structure of the floorlamp, and the likely contact region at the floorlamp’s base—directly from the reference images. These decoded outputs demonstrate that the VLM preserves high-level relational semantics, reinforcing the effectiveness of using its intermediate representations as motion priors.

References

- [1] Huan Fu, Rongfei Jia, Lin Gao, Mingming Gong, Binqiang Zhao, Stephen J. Maybank, and Dacheng Tao. 3d-future: 3d furniture shape with texture. *Int. J. Comput. Vis.*, 129, 2021. 3, 5
- [2] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *CVPR*, 2022. 1
- [3] Myung-Soo Kim. Painless introduction to geometric concepts and tools in computer graphics and CAD: applied geometry for computer graphics and cad; d. marsh; springer, london, 1999, 288 pages, ISBN 1-85233-080-5. *Comput. Aided Des.*, 32, 2000. 1
- [4] Jiaman Li, Jiajun Wu, and C. Karen Liu. Object motion guided human motion synthesis. *ACM Trans. Graph.*, 42, 2023. 2, 3, 4
- [5] Jiaman Li, Alexander Clegg, Roozbeh Mottaghi, Jiajun Wu, Xavier Puig, and C. Karen Liu. Controllable human-object interaction synthesis. In *ECCV*, 2024. 1, 3
- [6] Sergey Prokudin, Christoph Lassner, and Javier Romero. Efficient learning on point clouds with basis point sets. In *ICCV*, 2019. 1
- [7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1



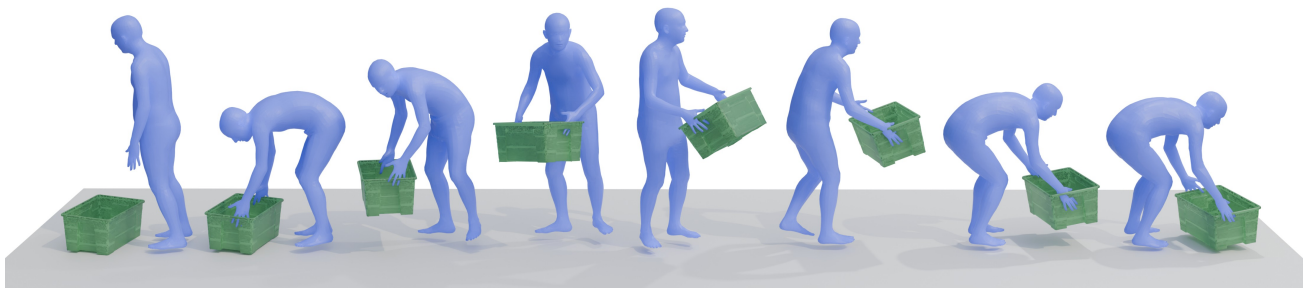
Push the largebox, and set it back down.



Push the smalltable, and set it back down.



Lift the trashcan, move the trashcan, and put down the trashcan.



Lift the plasticbox, move the plasticbox, and put down the plasticbox.

Figure 4. Additional visualization result on the FullBodyManipulation dataset [4].

- [8] Gemini Team. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv*, abs/2507.06261, 2025. 1
- [9] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir,

- Daniel Cohen-Or, and Amit Haim Bermano. Human motion diffusion model. In *ICLR*, 2023. 1, 3
- [10] Mengqing Xue, Yifei Liu, Ling Guo, Shaoli Huang, and Changxing Ding. Guiding human-object interactions with



Lift the woodchair, move the woodchair, and put down the woodchair.



Lift the smallbox, move the smallbox, and put down the smallbox.

Figure 5. Additional visualization result on the 3D-FUTURE dataset [1].

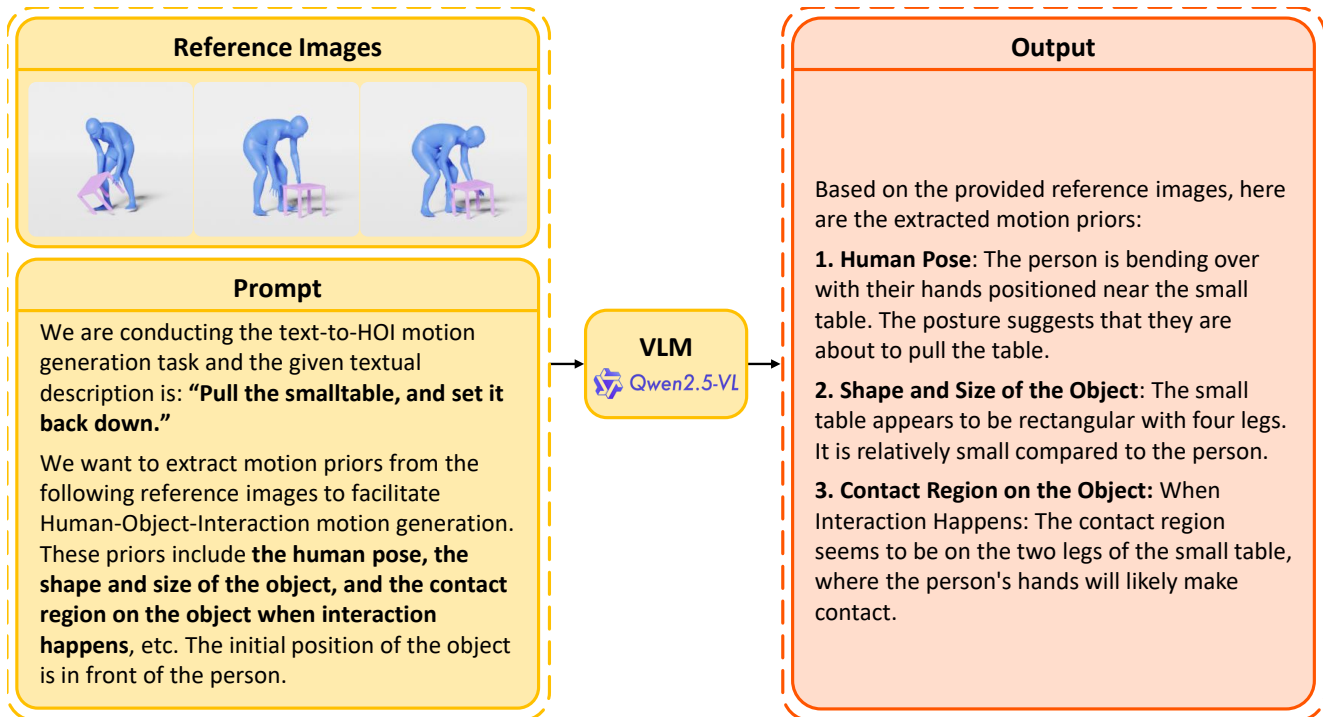


Figure 6. Qualitative analysis of VLM understanding. The text annotation is “Pull the smalltable, and set it back down.”

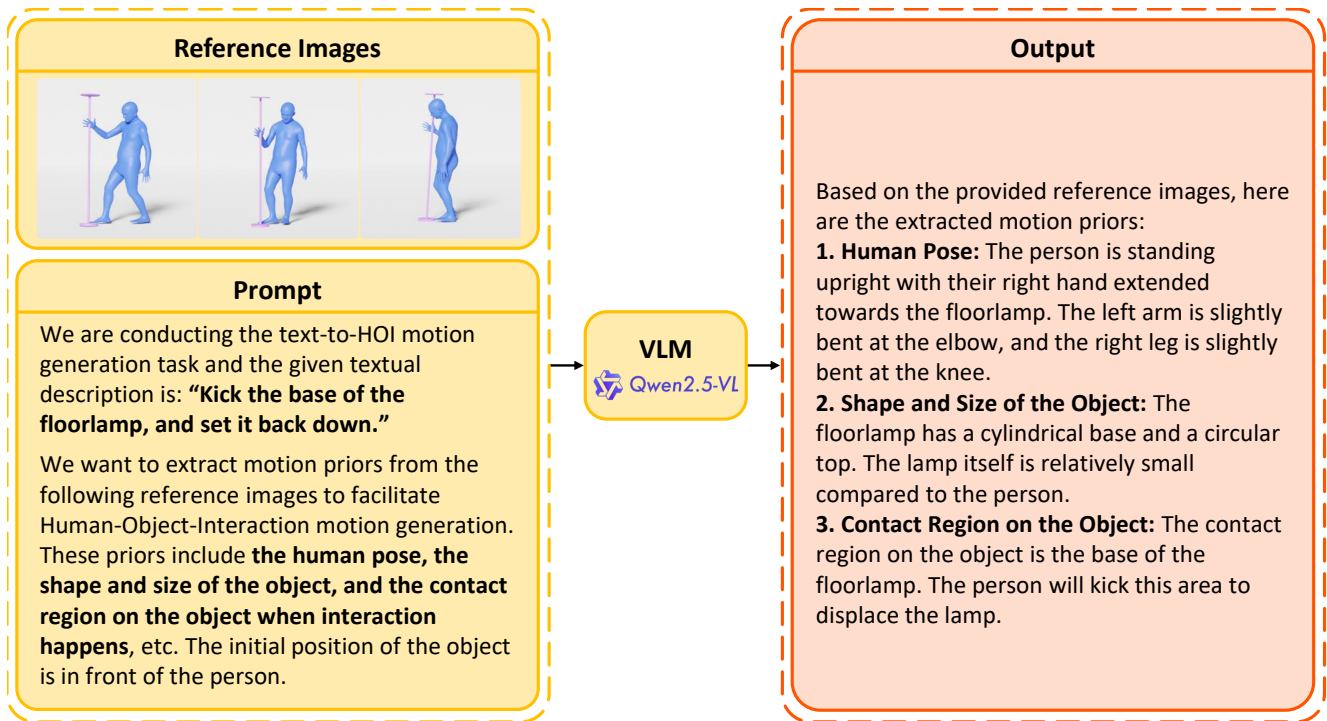


Figure 7. Qualitative analysis of VLM understanding. The text annotation is "Kick the base of the floorlamp, and set it back down."

rich geometry and relations. In *CVPR*, 2025. [1](#), [3](#)

- [11] Cem Yuksel. Sample elimination for generating poisson disk sample sets. *Comput. Graph. Forum*, 34, 2015. [1](#)