Breaking the 3D Dataset Bottleneck: Fast Scalable Generation of Aligned 3D Assets from Scratch for Category 6D Pose Estimation and Robotic Grasping

Anonymous authors

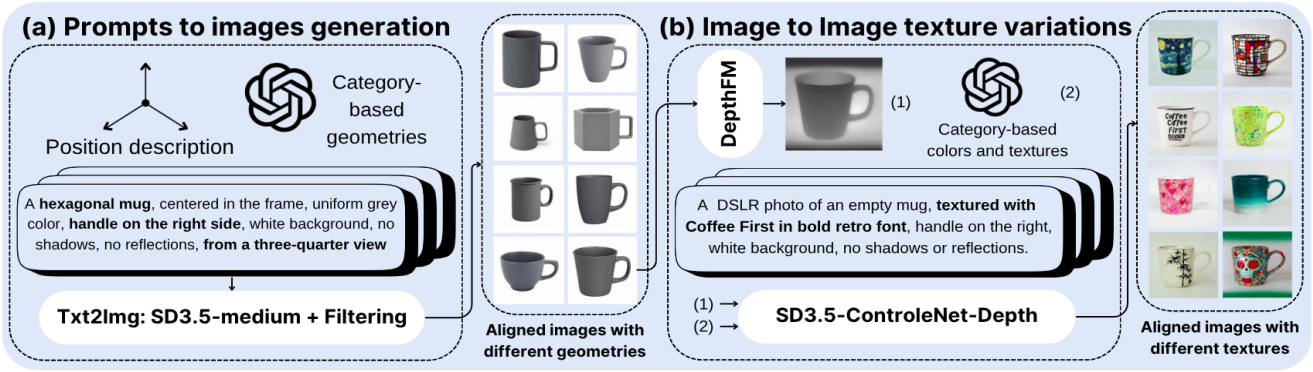

Figure 1. Our text-to-image pipeline: (a) Category-based geometry prompt engineering and image generation; (b) Depth-conditioned image generation for texture variation and automatic alignment.

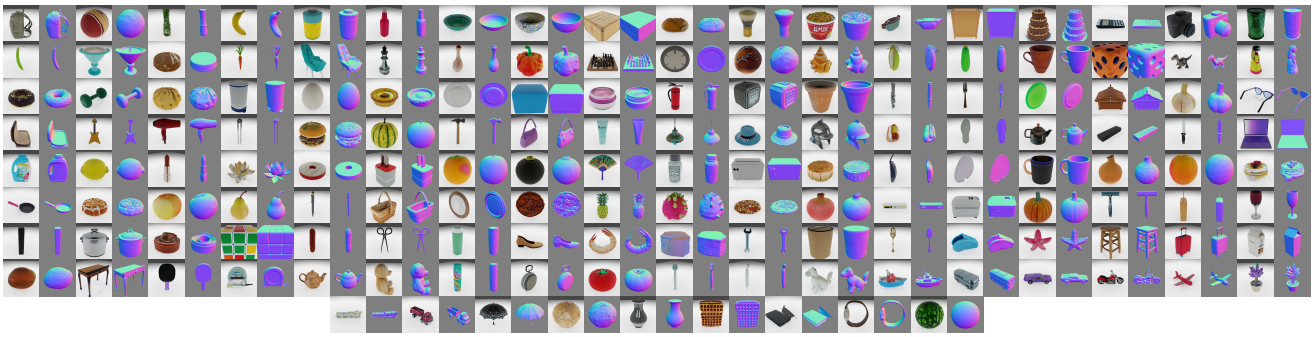

Figure 2. GenOmni3D Dataset: RGB and Normal images for objects from all 153 categories.

Abstract

While 2D vision has been revolutionized by large-scale datasets, 3D vision remains constrained by scarce, canonically aligned data. We introduce the first scalable, automated framework that generates complete category-level 6D pose datasets directly from text prompts, bypassing existing 3D assets. Our method achieves: (1) reliable asset generation via a controlled text-to-image-to-3D pipeline; (2) built-in canonical alignment through depth-conditioned generation (96% pose consistency); (3) large-scale 6D annotation via mixed reality rendering. The pipeline produces aligned meshes in under 3 minutes per object (5–20× speedup). We generate over 1,000 instances for each of 153 categories (153,000 meshes, ¿40× increase per category). Extensive evaluation shows competitive zero-shot sim2real transfer on NOCS and superior robotic grasping (87.8% success), where aligned meshes prove essential. We release the largest publicly available aligned 3D mesh dataset, category-level 6D pose dataset, grasping environments, and open-source pipeline. Code and data: https://genomni3d.github.io/

1. Introduction

The field of 2D computer vision has been revolutionized by foundation models trained on massive-scale datasets [12–

14]. In contrast, 3D vision advancement remains constrained by a fundamental limitation: the scarcity of high-quality, diverse, and scalable annotated 3D data. This bottleneck is critical as advanced 3D understanding underpins robotics, augmented reality, and autonomous systems. Category-level 6D pose estimation—predicting an object's 3D position and orientation from a single RGB-D image without instance-specific models—epitomizes this challenge. Creating such datasets faces three bottlenecks: (1) Asset collection depends on labor-intensive scanning (15–60 min/object) or limited repositories with inconsistent quality; (2) Mesh alignment requires extensive manual effort, hindering large-scale creation; (3) Pose annotation is error-prone and hard to scale. Our end-to-end pipeline transforms text prompts into complete 6D pose and grasping datasets. Our contributions can be summarized as :

Generative Pipeline for Aligned 3D Assets: Near automated framework converting category descriptions into aligned 3D meshes with 96% pose consistency (vs. 57% prior work [6]) in <3 min (5–20× speedup).

Benchmarking Dataset Generation: We reproduce and open-source state-of-the-art category-level 6D dataset generation pipelines [21, 23], including full 3D simulation and mixed-reality rendering, creating comprehensive datasets.

Large-Scale Dataset Releases: Two datasets: (1) the largest aligned 3D mesh dataset (153K meshes, 153 categories); (2) category-level 6D pose dataset (1.2M images with full annotations).

Grasping Simulation and Real-World Validation: Custom grasping environments in SAPIEN [20] achieve 87.8% grasp success, significant shape completion improvements (0.475 IoU vs 0.314), and superior real-world zero-shot transfer.

2. Related Work

2.1. 3D Datasets

As shown in Table 1, existing 3D datasets fall into categories: synthetic (ShapeNet [2], ModelNet [19]) lacking realism; real-world scans (OmniObject3D [18]) limited scalability (15–60 min/object); large internet collections (ObjaverseXL [4]) with inconsistent quality and sparse category coverage and no alignment; generated collections (GenVegeFruits [6]) mostly limited to symmetric objects and computationally heavy. Our datasets *GenNOCS3D* and *GenOmni3D* uniquely combine large-scale instance diversity with built-in canonical alignment and fast generation (3 min/object).

2.2. Category 6D Pose Datasets

Real-world 6D annotation is costly; synthetic datasets require diverse 3D assets. NOCS [16] (6 categories) established the first benchmark; Omni6D [23]/Omni6Dpose [21] combine synthetic and real data but suffer from limited instance diversity and reproducibility challenges. Our approach generates unlimited, canonically aligned assets across arbitrary categories. Table 2 compares existing datasets with our *GenNOCS6D* and *GenOmni6D*.

Table 1. Comparison of existing 3D mesh datasets. R/S/SAI: real, synthetic, generative.

Dataset	R/S/SAI	#Obj	#Cat	#O/C	Ali	Quality	Time
ShapeNet [2]	S	51k	55	927	Y	*	N/A
ModelNet [19]	S	12k	40	300	Y	*	N/A
OmniObject3D [18]	R	6k	190	32	Y	***	15-60m
ObjaverseXL [4]	R+S	10.2M	-	-	N	*	N/A
GenVegeFruits [6]	SAI	100K	100	1000	Y	***	15min
GenOmni3D (Ours)	SAI	153K	153	1000	Y	***	**3min**

3. From category prompts to aligned textured 3D mesh generation

Our pipeline (Fig. 1) converts category descriptions into aligned meshes with 96% pose consistency, requiring only 2 hours of manual effort for the complete NOCS3D dataset. It comprises four phases:

1. LLM-based Geometry Prompt Engineering: LLMs generate category-specific prompts with randomized shape descriptions and self-verification.

2. Image Generation and Depth Estimation: Diffusion models [7] produce initial images and a manual filtering (<10 min/category) removes outliers to get 100 images per category. DepthFM [8] processes selected images to create depth maps, conditioning subsequent stages.

3. Texture Variation: Each depth map conditions generation of 10 textured instances via LLM texture prompts, yielding 1K images per category.

4. 3D Reconstruction: A state-of-the-art image-to-3D model (Hunyuan3D-v2.0 [15]) produces consistently aligned meshes from aligned images (Fig. 3).

Indeed, previous work on text-only generation achieves 80% pose consistency for symmetric objects but drops to 20% for asymmetric ones (e.g., laptops). We adopt ControlNet [22] with depth conditioning, achieving near 100% for symmetric objects and up to 96% overall (Table 3). Depth maps capture global structure without constraining texture details, ensuring pose consistency and geometric diversity.

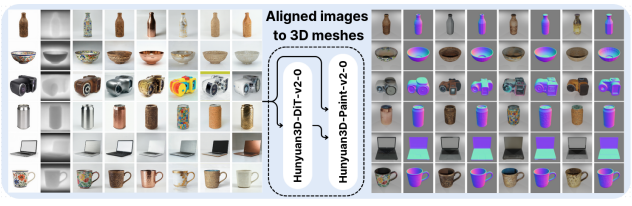

Figure 3. Examples of final textured images (left) and resulting 3D meshes using Hunyuan3D-v2.0 (right) for NOCS categories.

Table 2. Comparison of category-level 6D pose datasets.

Dataset	3D dataset	Rend.	#Cat	#O/C	#O	#Img	Code
NOCS-CAMERA25 [16]	ShapeNet	Rast	6	180.8	1080	300K	✗
Omni6DPose-SOPE [21]	OmniObject3D	RT+MR	149	27.9	4023	475K	✗
Omni6D [23]	OmniObject3D	RT	166	28.2	4648	0.8M	✓
Omni6D-xl [23]	Multiple	RT	419	38.1	15922	1.1M	✓
GenNocs6D (Ours)	GenNocs3D	RT+MR	6	**1000**	6000	600K	✓
GenOmni6D (Ours)	GenOmni3D	RT+MR	153	**1000**	**153K**	**1.2M**	✓

Table 3. Pose consistency (%) with/without depth.

Category	Depth (Ours)	Text-only [6]
Bottle	100	70
Bowl	100	70
Camera	97	30
Can	100	70
Laptop	90	20
Mug	100	82
Avg GenNOCS	97	57
Avg GenOmni3D	96	–

4. Mesh integration for category-level 6D pose dataset generation

We extend BlenderProc [5] to generate synthetic training data for category 6D pose estimation, providing two complementary pipelines.

4.1. Complete 3D simulation approach

Adopted from Omni6D [23], this uses realistic synthetic scenes from scanned homes. Objects are randomly placed in physically delineated areas with 10 camera viewpoints per configuration, illuminated by five random light sources. We generated 300K images matching NOCS dataset size.

4.2. Mixed reality rendering pipeline

Based on NOCS [16] and Omni6DPose [21], this places synthetic objects on planar surfaces against real image backgrounds with ray-traced shadows. Camera poses are aligned with background viewpoints [1]. Objects are placed using multi-view ray casting and gravity simulation. Shadows are rendered separately and composited with the background and real depth. We generated 300K images for NOCS and >1M for Omni6Dpose.

5. Grasping dataset generation

We integrate generated objects into SAPIEN simulator [20] and use CenterGrasp [3], which jointly trains 6D pose estimation, mesh reconstruction, and grasp prediction using Signed Distance Functions for Grasping (SDFG).

5.1. Physical properties

We assign realistic scales and sample densities per category, generate collision meshes via V-HACD [11] convex decomposition, and produce URDF files for 100 objects per category (600 total).

5.2. SAPIEN scene generation

Following CenterGrasp [3], we create "pile" and "packed" scenes with varying complexity. Training data includes grasp poses, RGB images with annotations (heatmaps, 6D poses, latent codes). We randomize ground/table materials and create two dataset versions (native textures vs. randomized textures, Fig. 4-5).

6. Experiments

6.1. Benchmarking 6D pose generation on NOCS

We evaluate DualPoseNet [10] on five dataset variants (100K images each) to assess: (i) Mixed-reality vs. fully synthetic: mixed-reality (Mix^{sh}_{SAI}) achieves best zero-shot Sim2Real on REAL275 (avg 34.75 vs. Replica's 33.10). (ii) Generated mesh quality: our meshes improve synthetic validation (23.91 vs. 15.66) and Sim2Real (30.83 vs. 29.07). (iii) Shadow importance: shadow-enabled versions consistently outperform shadow-free (e.g., 34.75 vs. 30.83). Training on our full synthetic dataset yields zero-shot performance comparable to supervised methods (Table 4).

6.2. Grasping and shape completion evaluation

We evaluate within CenterGrasp [3] framework on SAPIEN. Models trained on our data (Custom-CG) outperform pre-trained CenterGrasp and GIGA [9] across all metrics (Table 5). Best model achieves 87.8% grasp success (vs. 82.7% CenterGrasp) and shape completion IoU 0.475 (vs. 0.314). Native textures further improve performance.

6.3. Real robot application

Zero-shot sim-to-real benchmark on 36 objects across six NOCS categories. Our NOCS model (trained on custom data) outperforms baseline; category-specialized models achieve highest detection and grasping success (Table 6, Fig. 10), demonstrating the value of easily generated, domain-specific training data.

7. Conclusion

This work overcomes the data bottleneck in 3D vision with an automated pipeline generating complete category-level 6D pose and grasping datasets from text prompts. Key innovation: generating high-quality, canonically aligned 3D meshes with 96% pose consistency across 153 categories, achieving 5–20× speedup over scanning. Released datasets include the largest aligned 3D mesh collection (153K meshes) and category-level 6D pose dataset (1.2M images). Validation shows competitive zero-shot sim2real transfer on NOCS and superior robotic grasping (87.8% success) confirmed by real-world testing. Our work transforms 3D dataset creation, enabling scalable foundation models for 3D vision and robotic manipulation.

Figure 4. Random textures.

Figure 5. Object textures.

Figure 6. CAMERA25 synthetic [16].

Figure 7. Mixed-reality IKEA (Ours).

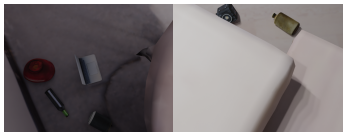

Figure 8. REPLICA-based (Ours).

Figure 9. REAL275 real images [16].

Table 4. Comprehensive evaluation of DualPoseNet for Sim2Real transfer and in-domain performance. The top section reports **zero-shot transfer** to the real-world NOCS REAL275 test set. The middle section shows performance on **synthetic validation** splits. The bottom section provides the original NOCS **supervised upper-bounds** for reference. Metrics evaluate both 2D detection (IoU50/75) and 3D pose accuracy, where n°, mcm measures the percentage of poses with rotation error $< n^\circ$ *and* translation error $< m$cm. Best scores for the zero-shot Sim2Real and our synthetic validation experiments are underlined; overall best (supervised) results are in **bold**. The metrics are symmetry aware as done in Omni6D [23]. Evaluation of DualPoseNet [10] for Sim2Real transfer

Training	**Test**	IoU_{50}	IoU_{75}	$5°2\,cm$	$5°5\,cm$	$10°2\,cm$	$10°5\,cm$	$5°$	$10°$	$2\,cm$	$5\,cm$	Avg
Replica_train	Real275 [17]	82.03	34.33	3.40	4.94	11.51	17.12	5.75	18.66	**55.24**	**98.05**	33.10
Mix^{sh}_{SAI}	Real275 [17]	**85.91**	**35.42**	4.62	**7.36**	**13.95**	**22.19**	**8.49**	**23.83**	49.29	96.43	**34.75**
Mix^{no-sh}_{SAI}	Real275 [17]	84.89	28.86	2.27	3.77	8.07	14.13	4.49	15.70	48.43	97.66	30.83
Mix^{sh}_{syn}	Real275 [17]	82.68	31.50	**4.80**	6.32	12.90	17.81	7.01	19.15	52.46	97.77	33.24
Mix^{no-sh}_{syn}	Real275 [17]	73.56	26.05	2.23	3.46	6.81	11.02	4.08	12.31	53.16	**98.05**	29.07
Replica-train	Replica-val	66.43	16.56	1.52	3.04	4.18	8.57	3.34	9.41	**28.05**	**83.67**	22.48
Mix^{sh}_{SAI}_train	Mix^{sh}_{SAI}_val	**69.60**	**22.44**	**2.09**	**4.25**	**5.25**	**10.77**	**4.78**	**12.12**	26.31	81.47	**23.91**
Mix^{no-sh}_{SAI}_train	Mix^{no-sh}_{SAI}_val	68.90	21.72	2.18	4.15	5.16	10.28	4.59	11.41	26.30	81.28	23.60
Mix^{sh}_{syn}_train	Mix^{sh}_{syn}_val	46.93	6.32	0.98	1.74	1.81	3.81	1.83	4.16	16.39	72.59	15.66
Mix^{no-sh}_{syn}_train	Mix^{no-sh}_{syn}_val	46.47	6.46	0.85	1.68	1.70	3.90	1.74	4.22	17.63	72.45	15.71
REAL275_train [16]	REAL275 [16]	79.43	36	32.24	39.24	52.23	67.70	41.44	70.14	70.47	96.89	55.42
CAMERA25 [16] + REAL275 [16]_train	REAL275 [16]	**84.19**	**76**	**40.49**	**45.87**	**63.77**	**75.06**	**47.93**	**76.27**	**82.76**	**99.34**	**61.67**
GenNOCS_train (ours)	REAL275 [16]	81.25	58	35.67	42.15	58.92	71.38	44.28	73.45	76.84	97.89	58.49
GenNOCS_train (ours)	GenNOCS_val	**95.51**	**82.46**	**56.11**	**57.31**	**67.81**	**69.91**	**60.14**	**71.66**	**93.73**	**99.04**	**77.15**
GenOmni3D_train (ours)	GenOmni3D_val	**81.83**	**37.80**	**11.99**	**15.36**	**23.98**	**32.06**	**16.92**	**35.75**	**52.09**	**84.23**	**38.58**

Table 5. Grasping and shape completion comparison.

Method	**Grasp success**	**bi** ↓	**IoU** ↑
GIGA-val-tex [9]	0.6375	55.2	0.146
Centergrasp-val-tex [3]	0.7896	27.0	0.314
Custom-CG-tex-val-tex (Ours)	**0.8679**	**23.5**	**0.475**
Centergrasp-val-rdom [3]	0.8271	28.0	0.312
Custom-CG-rdom-val-rdom (Ours)	**0.8784**	**19.2**	**0.453**

Table 6. Real-world detection and grasping success rates.

Method	Can	Bottle	Bowl	Mug	Camera	Laptop
Detec (Baseline)	83%	50%	28%	58%	73%	0%
Detec (NOCS)	**100%**	**71%**	71%	**83%**	60%	**100%**
Detec (Specialized)	**100%**	66%	**86%**	75%	**100%**	**100%**
Grasp (Baseline)	96%	50%	14%	50%	66%	0%
Grasp (NOCS)	**100%**	46%	47%	58%	60%	0%
Grasp (Specialized)	**100%**	**52%**	**57%**	**71%**	**73%**	0%

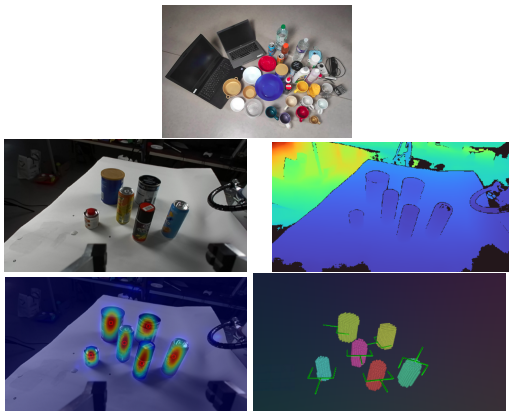

Figure 10. Robotic perception and grasping on can object using CenterGrasp [3] trained on our custom dataset.

References

[1] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017. 3

[2] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shu-

ran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 2

[3] Eugenio Chisari, Nick Heppert, Tim Welschehold, Wolfram Burgard, and Abhinav Valada. Centergrasp: Object-aware implicit representation learning for simultaneous shape reconstruction and 6-dof grasp estimation. *IEEE Robotics and Automation Letters*, 2024. 3, 4

[4] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[5] Maximilian Denninger, Dominik Winkelbauer, Martin Sundermeyer, Wout Boerdijk, Markus Knauer, Klaus H. Strobl, Matthias Humt, and Rudolph Triebel. Blenderproc2: A procedural pipeline for photorealistic rendering. *Journal of Open Source Software*, 8(82):4901, 2023. 3

[6] Guillaume Duret, Younes Bourenname, Danylo Mazurak, Anna Samsonenko, Florence Zara, Liming Chen, and Jan Peters. Facilitate and scale up the creation of 3d meshes and 6d category-based datasets with generative models: Genvegefruits3d. *HAL*, 2025. 2, 3

[7] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. 2

[8] Ming Gui, Johannes Schusterbauer, Ulrich Prestel, Pingchuan Ma, Dmytro Kotovenko, Olga Grebenkova, Stefan Andreas Baumann, Vincent Tao Hu, and Björn Ommer. Depthfm: Fast generative monocular depth estimation with flow matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3203–3211, 2025. 2

[9] Zhenyu Jiang, Yifeng Zhu, Maxwell Svetlik, Kuan Fang, and Yuke Zhu. Synergies between affordance and geometry: 6-dof grasp detection via implicit representations. *arXiv preprint arXiv:2104.01542*, 2021. 3, 4

[10] Jiehong Lin, Zewei Wei, Zhihao Li, Songcen Xu, Kui Jia, and Yuanqing Li. Dualposenet: Category-level 6d object pose and size estimation using dual pose network with refined learning of pose consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3560–3569, 2021. 3, 4

[11] Khaled Mamou and Faouzi Ghorbel. A simple and efficient approach for 3d mesh approximate convex decomposition. In *2009 16th IEEE international conference on image processing (ICIP)*, pages 3501–3504. IEEE, 2009. 3

[12] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 1

[13] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[14] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 2

[15] Tencent Hunyuan3D Team. Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation, 2025. 2

[16] He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J. Guibas. Normalized object coordinate space for category-level 6d object pose and size estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 3, 4

[17] He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J. Guibas. Normalized object coordinate space for category-level 6d object pose and size estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 4

[18] Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Liang Pan Jiawei Ren, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, Dahua Lin, and Ziwei Liu. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[19] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920, 2015. 2

[20] Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, Li Yi, Angel X. Chang, Leonidas J. Guibas, and Hao Su. SAPIEN: A SimulAted Part-Based Interactive ENvironment. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11094–11104. IEEE, 2020. 2, 3

[21] Jiyao Zhang, Weiyao Huang, Bo Peng, Mingdong Wu, Fei Hu, Zijian Chen, Bo Zhao, and Hao Dong. Omni6dpose: a benchmark and model for universal 6d object pose estimation and tracking. In *European Conference on Computer Vision*, pages 199–216. Springer, 2025. 2, 3

[22] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 2

[23] Mengchen Zhang, Tong Wu, Tai Wang, Tengfei Wang, Ziwei Liu, and Dahua Lin. Omni6d: Large-vocabulary 3d object dataset for category-level 6d object pose estimation. In *European Conference on Computer Vision*, pages 216–232. Springer, 2025. 2, 3, 4