

IMIGRASP: Zero-shot 6-DOF pose estimation via GenAI imagination and cross-image geometric transfer

Muhammad Aun Raza

Abstract—Robotic grasping in unstructured environments remains a fundamental challenge, particularly for novel objects not encountered during training. Traditional approaches rely on large scale annotated datasets, extensive training, and hand crafted features. This work introduces IMIGRASP (Imagine to Grasp), a zero-shot framework for 6-DOF pose estimation that leverages generative artificial intelligence for imagination guided pose estimation. The key insight is that modern generative models, trained on vast internet scale visual data, implicitly encode rich knowledge about human grasping behavior and object affordances. Given a single RGB image of a target object, the proposed pipeline employs GPT-5.6 Luna to generate a corresponding image with a realistic hand grasping the object, complete with colored landmarks at the thumb tip, index fingertip, and wrist. A state-of-the-art segmentation model, SAM3, is utilized for precise segmentation of objects and hands, while a bounding box transfer method is introduced for cross-image geometric transfer between generated and real images. The complete 3D grasp frame is reconstructed using RGB-D data and robot kinematics, and executed on a Franka Panda robot. The approach is validated on 130 diverse household objects, achieving a success rate of 86.9% without any prior training or object specific tuning. The results demonstrate that generative AI models can serve as powerful zero-shot priors for robotic manipulation, opening new directions for imagination guided robotics.

Index Terms—Generative AI for pose estimation, zero-shot robotic grasping, 6-DOF grasp generation, foundation models for manipulation, robotic affordance learning.

I. INTRODUCTION

ROBOTIC grasping is a cornerstone capability for autonomous manipulation systems, enabling robots to interact with and manipulate objects in unstructured environments [1, 2]. Despite significant advances in recent years, robust and generalizable grasping remains challenging, particularly for novel objects and poses not seen during training [3]. Recent approaches have explored zero-shot grasping using generative examples and semantic guidance [4]. In this work, IMIGRASP (Imagine to Grasp) is introduced, a zero-shot framework that leverages generative AI for 6-DOF grasp generation. The core insight is that GPT-5.6 Luna’s ability to produce realistic hand object interactions from a single RGB image can be harnessed to infer optimal grasp configurations without any prior training [5, 6]. Modern generative models demonstrate remarkable capabilities in generating realistic images and understanding visual concepts from text, and these models, trained on vast

internet scale data, implicitly encode rich knowledge about human behavior, object affordances, and physical interactions. Specifically, GPT-5.6 Luna [7] exhibits particularly strong capabilities in generating realistic human object interactions from visual inputs, making it well suited for synthesizing grasping configurations [8, 9]. The complete pipeline consists of generative image synthesis with colored landmarks, segmentation using SAM3, color based landmark detection, 2D geometric construction for grasp point identification, scale-invariant bounding box transfer, RGB-D based 3D reconstruction, and grasp frame computation and execution [10, 11, 12]. Recent zero-shot and task-oriented grasping approaches have also been investigated [13, 14, 15]. Traditional approaches rely on large scale annotated datasets and extensive training, which limits their applicability to new domains and objects [16, 17, 18]. Affordance learning and language grounding have been explored for robotic manipulation [19, 20, 21]. The concept of imagination has also been investigated in robotics for planning and reasoning [22, 23, 24]. In contrast to these approaches, this work is unique in using generative AI capabilities for zero-shot grasp generation, requiring no training or fine-tuning [25, 26, 27]. The main contributions include: a zero-shot framework for 6-DOF grasp generation using generative AI without any training data; a pipeline combining GPT-5.6 Luna, SAM3, and geometric construction for grasp pose estimation; a scale-invariant bounding box transfer method for correspondence mapping between generated and real images; and comprehensive experimental validation on 130 diverse household objects with a Franka Panda robot [28, 29, 30].

II. RELATED WORK

A. GenAI for Grasping

Recent works have explored generative AI for robotic grasping. GRIM [1] retrieves and re-aligns grasps from generative examples in a training-free memory-based pipeline. SegGrasp [2] employs semantic and geometric guided segmentation for zero-shot task-oriented grasping. Imagine2Act [3] leverages object-action motion consistency from imagined goals within a trained imitation-learning policy. Affordance diffusion [4] synthesizes hand-object interactions using diffusion models trained on large-scale data. GraspGen-X [11] performs cross-embodiment diffusion-based grasping trained on billions of simulated grasps. GraspLDP [12] proposes a latent diffusion approach for generalizable grasping. GraspFoM [13] utilizes

M. A. Raza is affiliated with the Department of Electrical Engineering, Pakistan Institute of Engineering and Applied Sciences (PIEAS), Islamabad, Pakistan. (e-mail: muhamamdaun33555@gmail.com).

3D foundation priors for joint reconstruction and grasp generation. DCGrasp [14] enables controllable, diffusion-based hand-object grasp generation. Dexter [15] employs autoregressive, language-driven dexterous grasp generation with embodied contact reasoning. Cross-view fusion [16] fuses multiple camera views via a trained network for robust 6-DoF pose estimation, while Contact-GraspNet [10] serves as a purely discriminative, non-generative baseline trained on synthetic point clouds.

B. Imagination and Affordance Learning

The concept of imagination and affordance learning has been explored for robotic manipulation. Luo et al. [26] learn affordance grounding from exocentric images via a trained cross-view transfer framework. Part-SLIP [27] lifts 2D detections from a pretrained image-language model for low-shot 3D part segmentation. SATR [28] performs zero-shot 3D shape segmentation using off-the-shelf 2D detectors, requiring no training. CLIP-based models [29] enable open-vocabulary understanding through large-scale contrastive pretraining. Visual foresight [30] learns physical dynamics through video prediction, requiring extensive training data. One-shot affordance learning [18] trains on a single example per category yet generalizes reasonably to unseen objects and affordances. ShapeGrasp [19] uses large language models over a geometric decomposition for zero-shot, training-free grasp reasoning. Language-guided dexterous grasp generation [20] guides grasp selection via a dataset-trained model, while AnyGrasp [17] offers a purely discriminative grasp-perception baseline with no affordance or imagination component.

III. METHODOLOGY

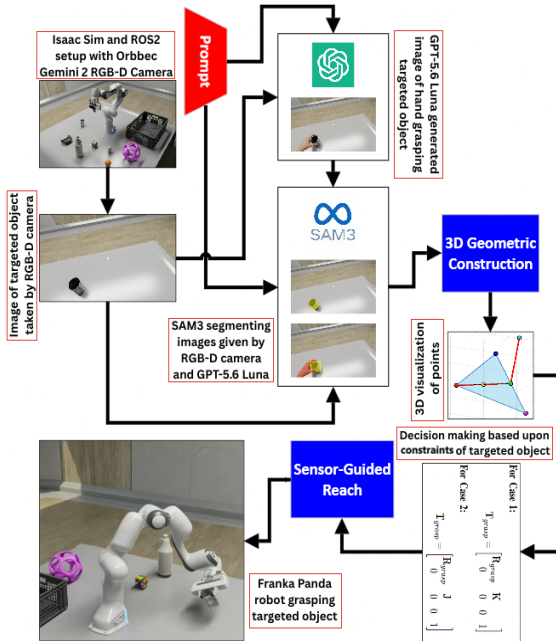


Fig. 1: Block diagram illustrating the complete IMIGRAS pipeline for zero-shot 6-DOF grasp generation.

A. GPT-5.6 Luna Image Generation

Given the original RGB image $\mathbf{I}_{orig} \in \mathbb{R}^{1080 \times 1920 \times 3}$, GPT-5.6 Luna generates a corresponding image $\mathbf{I}_{gpt}$ with a realistic hand grasping the object, preserving object position, orientation, and scale. The prompt requires red, blue, and green circles at thumb tip (A), index fingertip (B), and wrist (C), respectively. This step leverages the generative model's understanding of human grasping behavior to produce physically plausible hand configurations without any training data.

B. Segmentation and Landmark Detection

SAM3 [9] segments object masks $\mathcal{M}_{gpt}$ and $\mathcal{M}_{real}$, and hand mask $\mathcal{M}_{hand}$. Colored landmarks are detected using a color LUT approach. For each target color, a LUT is built and the point is found by minimizing within the hand mask:

$$\text{LUT}_{color}(r, g, b) = (r - r_{target})^2 + (g - g_{target})^2 + (b - b_{target})^2 \quad (1)$$

$$P = \arg \min_{\mathbf{p} \in \mathcal{M}_{hand}} \text{LUT}_{color}(\mathbf{I}_{gpt}(\mathbf{p})) \quad (2)$$

This yields $A = (u_A, v_A)$ for thumb tip, $B = (u_B, v_B)$ for index fingertip, and $C = (u_C, v_C)$ for wrist.

C. Mask Validation and Correction

Points A and B are validated to lie within $\mathcal{M}_{gpt}$. If not, the closest mask pixel is found via expanding circle search:

$$A' = \arg \min_{\mathbf{p} \in \mathcal{M}_{gpt}} \|\mathbf{p} - A\|_2, \quad B' = \arg \min_{\mathbf{p} \in \mathcal{M}_{gpt}} \|\mathbf{p} - B\|_2 \quad (3)$$

This validation step ensures robustness against minor inaccuracies in GPT-generated landmark placement, preventing downstream errors in grasp computation.

D. 2D Geometric Construction

Line ℓ_C is constructed through C parallel to AB, and perpendicular bisector ℓ_D of AB is constructed. Their intersection E is computed. Walking along $\overline{ED}$ from E to D, the first pixel inside $\mathcal{M}_{gpt}$ is identified as the grasp point T. This geometric construction leverages the relationship between finger positions and object geometry to determine the optimal grasp point on the object surface.

E. Bounding-Box Transfer

Direct pixel transfer between images is not possible due to minor scale and translation differences. For each mask $\mathcal{M}$, the bounding box center V and diagonal length L are computed. The scale ratio $r = L(\mathcal{M}_{gpt})/L(\mathcal{M}_{real})$ captures the relative scaling. For each point P in $\{A, B, T\}$, it is transferred to the real image using polar coordinates centered at the bounding box center:

$$P'' = \text{Project}_{\mathcal{M}_{real}} \left(V_{real} + r \cdot \|P - V_{gpt}\|_2 \cdot \frac{P - V_{gpt}}{\|P - V_{gpt}\|_2} \right) \quad (4)$$

This approach ensures translation and scale invariance, enabling accurate correspondence mapping between generated and real images.

F. 3D Reconstruction and Grasp Frame Computation

Using the RGB-D camera intrinsic matrix $\mathbf{K}$:

$$\mathbf{K} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \quad (5)$$

The back-projection to camera coordinates is:

$$\mathbf{P}_{camera} = \begin{bmatrix} (u - c_x) \cdot d / f_x \\ (v - c_y) \cdot d / f_y \\ d \end{bmatrix} \quad (6)$$

With camera extrinsics from Franka Panda kinematics:

$$\mathbf{P}_{base} = \mathbf{R}_{camera}^{base} \mathbf{P}_{camera} + \mathbf{t}_{camera}^{base} \quad (7)$$

The 3D points $\mathbf{A}_{3D}$, $\mathbf{B}_{3D}$, and $\mathbf{T}_{3D}$ are obtained in the robot base frame. $\mathbf{K}$ is the midpoint of $\mathbf{A}$ and $\mathbf{B}$. Points $\mathbf{L}$ and $\mathbf{M}$ are obtained by rotating $\mathbf{AB}$ to be perpendicular to $\mathbf{TK}$ in 3D space. The normal to the plane containing $\mathbf{L}$, $\mathbf{M}$, $\mathbf{T}$ is computed:

$$\mathbf{n} = (\mathbf{M} - \mathbf{L}) \times (\mathbf{T} - \mathbf{L}) \quad (8)$$

And point $\mathbf{X}$ is defined as:

$$\mathbf{X} = \mathbf{K} + \hat{\mathbf{n}} \cdot \delta \quad (9)$$

completing the grasp frame.

G. Rule-Based Grasp Alignment

The Franka Panda's physical constraints inform the rule base. Point $\mathbf{Y}$ is the 3D projection of the midpoint of $\mathbf{A}$ and $\mathbf{B}$ from the camera stream.

Case 1: $Y_z - K_z \geq -0.03$ ($\mathbf{Y}$ is above $\mathbf{K}$ by any amount or below $\mathbf{K}$ by up to 3 cm)

The Z-axis (joint axis) is aligned parallel to $\mathbf{KX}$. The Y-axis (arm length direction) depends on $\mathbf{LM}$ magnitude:

Subcase 1a: $\|\mathbf{LM}\|_2 \geq 0.07$ ($\mathbf{LM}$ is 7 cm or more)

Y-axis is parallel to $\mathbf{TK}$, reach target is $\mathbf{X}$. Sensor feedback stops the arm if a surface is encountered during approach.

Subcase 1b: $\|\mathbf{LM}\|_2 < 0.07$ ($\mathbf{LM}$ is less than 7 cm)

Y-axis is parallel to $\mathbf{LM}$, reach target is $\mathbf{X}$. Sensor feedback stops the arm if a surface is encountered during approach.

Case 2: $Y_z - K_z < -0.03$ ($\mathbf{Y}$ is below $\mathbf{K}$ by more than 3 cm)

Z-axis is parallel to $\mathbf{KX}$. Y-axis is parallel to $\mathbf{TK}$. Reach target is $\mathbf{J}$, the midpoint of $\mathbf{TK}$. Sensor feedback stops the arm if a surface is encountered during approach. For all cases, the X-axis is the cross product of Y and Z axes, completing the orthonormal frame. The rotation matrix is constructed from the three axes i.e., x, y and z.

The final homogeneous transformation matrix is:

For Case 1:

$$\mathbf{T}_{grasp} = \begin{bmatrix} \mathbf{R}_{grasp} & \mathbf{K} \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (10)$$

For Case 2:

$$\mathbf{T}_{grasp} = \begin{bmatrix} \mathbf{R}_{grasp} & \mathbf{J} \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (11)$$

The 3 cm and 7 cm thresholds were determined according to the gripper's mechanical constraints and the geometric properties of objects commonly found in household manipulation tasks. These parameters can be tuned for different grippers configurations as needed.

IV. EXPERIMENTS AND RESULTS

A. Experimental Setup

Franka Panda robot and Orbbec Gemini 2 RGB-D camera were used in Isaac Sim along with ROS2 integration, in a cluttered environment with groups of 6-9 objects Fig. 2. The proposed method was validated on 130 diverse household objects including storage objects, decoration objects, toys, etc.

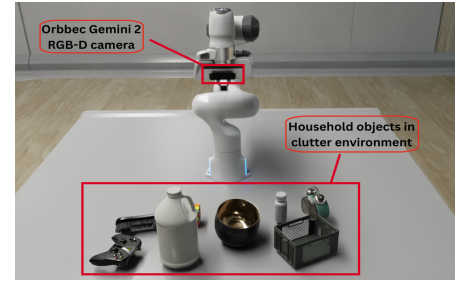


Fig. 2: Isaac Sim and ROS2 setup for experimental verification

B. Results and Analysis

The complete experimental results across all 130 objects show that IMIGRASP achieves a success rate of 86.9% (113 successful grasps out of 130 attempts) with only a single attempt per object. Cylindrical and cuboid objects achieved approximately 93% success rate, the big reason was that GPT was itself very clear that from where to grasp this object.

TABLE I: Success rate comparison with state-of-the-art methods.

Method	Success Rate (%)
Contact-GraspNet	84.3 (single attempt)
Cross-View Fusion	96.0 (multiple attempts)
IMIGRASP	86.9 (single attempt)

Table I presents the success rate comparison with state-of-the-art methods. In single attempt, it can be clearly seen that IMIGRASP has higher success rate. Fig. 3 shows complete pipeline demonstration for 7 objects i.e., object segmentation in actual image, GenAI generating grasping image, segmentation of hand and object, geometry application and final points extraction and successful grasp by Franka Panda robot in NVIDIA Isaac Sim.



Fig. 3: IMIGRASP: Zero-shot 6-DOF grasp generation using generative AI for imagination guided pose estimation. Each row shows results for one of 7 objects i.e., PS Controller, Shoe Brush, Brass Bowl, Utility Jug, Plastic Container, Packer Bottle, and Alarm Clock. Columns show object mask in actual image, hand grasping image by GPT-5.6 Luna, object and hand masks in genAI image, genAI image with geometric construction lines, object in actual image with final mapped points A, B, T and successful grasp execution attempt by Franka Panda robot respectively.

TABLE II: Grasp performance and failure analysis across 130 objects.

Outcome	Occurrences	Percentage
Contact with other objects	81	62.3%
Minor movement of others	59	45.4%
Major movement of others	22	16.9%
Other objects fell	13	10.0%
GPT-5.6 Luna failure	10	7.7%
SAM3 wrong bbox	4	3.0%
SAM3 wrong segmentation	3	2.3%

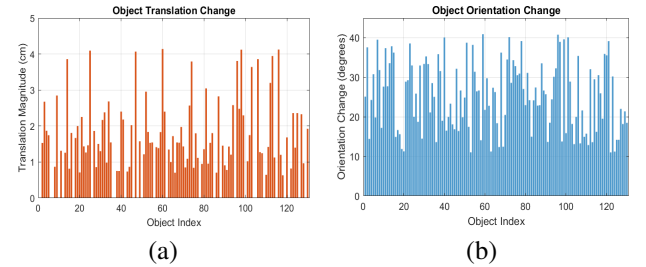


Fig. 4: Displacement and orientation error in all objects (a) Translation of objects from their original positions. (b) Orientation change of objects from their original orientations.

Fig. 4 presents the displacement analysis during grasping approach. Maximum translation observed was 4.2 cm, while maximum orientation change was 41 degrees. Table II summarizes success and failure categories and their frequencies.

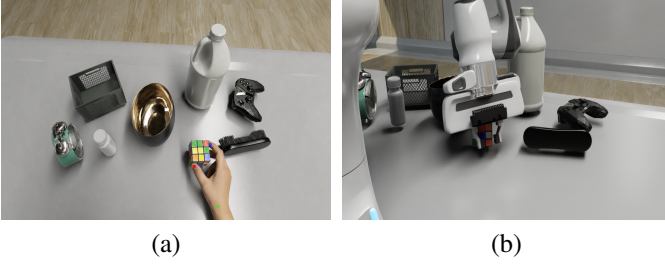


Fig. 5: Rubik's Cube failure case (a) GPT-5.6 Luna generated a grasp from the corner instead of the side. (b) The cube slipped under Franka Panda's finger force, resulting in an unsuccessful grasp.



Fig. 6: Failure case examples. (a) Bowl: SAM3 bounding box exceeds object limits causing misplacement of mapped points. (b) Oil Tin: GPT-5.6 Luna incorrectly generated a grasping configuration for a plastic container instead of the oil tin.

Of the 17 unsuccessful attempts, 10 were attributed to GPT-5.6 Luna failure, where the generated hand configuration was inappropriate for the object geometry or it grasped wrong object. For example, generating a corner grasp for the Rubik's Cube (Fig. 5a) instead of a side grasp, causing the cube to slip due to finger force (Fig. 5b) and the oil tin case (Fig. 6b) illustrates a unique failure where GPT-5.6 Luna generated a grasping configuration for a plastic storage container instead of the actual oil tin, demonstrating the importance of prompt clarity and the model's ability to distinguish between visually similar objects [5, 6]. The remaining 7 failures were attributed to SAM3, 4 due to SAM3 wrong bounding box (Fig. 6a), where the bounding box exceeded the object limits causing misplacement of mapped points and 3 failures due to SAM3 segmenting wrong object. In cluttered environments with groups of 6-9 objects, 81 successful grasps occurred with the object touching other objects. Among these, 59 resulted in minor movement of surrounding objects, 22 caused major movement, and 13 objects fell during the process. This indicates that while reaching and gripping the target object inevitably disturbs neighboring objects, the methodology remains robust in most cluttered scenarios. The 3D geometric construction using points L, M, T, K, X, and the associated vectors (LM, TK, KX) provides a general and flexible grasp

pose that does not require perfect adherence minor deviations, slips, and movements during approach still frequently lead to successful grasps, as evidenced by the overall success rate. Thin and horizontally oriented objects were not tested due to gripper and methodology limitations, as these geometries present challenges for the two-finger parallel gripper configuration and gripper will not be able to grasp that unless the object is thick enough that contact with gripper's fingers is enough to hold and lift it.

C. Comparative Analysis

A key advantage of the proposed zero-shot approach is its favorable success rate, space complexity, or zero prior training requirement, depending on the method compared, relative to training-based and retrieval-based approaches. Training-based methods like Contact-GraspNet [10] require 40 hours of GPU training on 17 million simulated grasps, while Cross-View Fusion [16] requires training on the large-scale GraspNet-1B dataset. Retrieval-based methods like GRIM [1] require memory construction (7 minutes per item) and online semantic alignment (10 seconds per grasp) during inference. In contrast, IMIGRASP operates with no training phase, reducing both the time and energy costs associated with model training. IMIGRASP achieves near constant time complexity $\mathcal{O}(1)$ with respect to the number of objects or training examples, as the pipeline consists of a single generative inference call followed by segmentation, basic matrix and pixel-wise operations. The space complexity is equally minimal $\mathcal{O}(1)$, requiring only the input image and intermediate masks during execution, with no datasets, memory banks, or precomputed features to store. In terms of success rate, IMIGRASP (86.9%) outperforms Contact-GraspNet's single-attempt rate (84.3%), albeit with a slightly higher VRAM requirement. Conversely, Cross-View Fusion achieves a higher success rate (96.0%), but this comes at the cost of allowing multiple attempts and requiring more VRAM and training data, whereas IMIGRASP achieves competitive performance with a single attempt and zero training.

V. CONCLUSION

IMIGRASP eliminates the need for training data, hand pose estimation networks, or object-specific tuning, making it suitable for applications where data collection is impractical. Its lower computational footprint and competitive success rate enable real-time deployment on robotic systems in warehouses and household manipulation tasks. The results validate that generative AI models, trained on internet-scale visual data, can serve as imagination priors for robotic grasping, reducing reliance on expensive training pipelines.

ACKNOWLEDGMENT

The author would like to acknowledge the support of Department of Electrical Engineering, Pakistan Institute of Engineering and Applied Sciences (PIEAS).

REFERENCES

- [1] Shailesh, A. Raj, N. Kumar, P. Shukla, A. Melnik, M. Beetz, and G. C. Nandi, “GRIM: Task-oriented grasping with conditioning on generative examples,” in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2024.
- [2] H. Li, W. Mao, W. Deng, C. Meng, R. Zhang, F. Jia, T. Wang, H. Fan, H. Wang, and X. Deng, “SegGrasp: Zero-shot task-oriented grasping via semantic and geometric guided segmentation,” in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2025.
- [3] L. Heng, J. Xu, Y. Wang, X. Li, M. Cai, Y. Shen, J. Zhu, G. Ren, and H. Dong, “Imagine2Act: Leveraging object-action motion consistency from imagined goals for robotic manipulation,” *arXiv preprint*, 2025.
- [4] Y. Ye, X. Li, A. Gupta, S. De Mello, S. Birchfield, J. Song, S. Tulsiani, and S. Liu, “Affordance diffusion: Synthesizing hand-object interactions,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023.
- [5] J. Achiam et al., “GPT-4 technical report,” *arXiv preprint arXiv:2303.08774*, 2023.
- [6] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, “Hierarchical text-conditional image generation with CLIP latents,” *arXiv preprint arXiv:2204.06125*, 2022.
- [7] OpenAI, “GPT-4: Generative Pre-trained Transformer 4,” 2023.
- [8] A. Kirillov et al., “Segment anything,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023.
- [9] N. Ravi et al., “SAM 2: Segment anything in images and videos,” *arXiv preprint arXiv:2408.00714*, 2024.
- [10] M. Sundermeyer, A. Mousavian, R. Triebel, and D. Fox, “Contact-graspnet: Efficient 6-DoF grasp generation in cluttered scenes,” in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, pp. 13438-13444, 2021.
- [11] B. Han, Y.-W. Chao, E. Coumans, C. Eppner, B. Sundaralingam, J. Deng, S. Birchfield, and A. Murali, “GraspGen-X: Cross-embodiment 6-DOF diffusion-based grasping,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2026.
- [12] E. Xiang, H. Ma, X. Ma, Z. Liu, and D. Huang, “GraspLDP: Towards generalizable grasping policy via latent diffusion,” *arXiv preprint*, 2025.
- [13] D. Wu, X. Wei, H. Wang, Q. Dong, Y. Li, Q. Wuwu, M. Lu, and W. Zhao, “GraspFoM: Towards reconstruction-driven robotic grasping with 3D foundation priors,” *arXiv preprint*, 2026.
- [14] H. Akada, J. Pérez, E. Aksan, V. Choutas, C. Romero, A. Garcia-Garcia, V. Golyanik, C. Theobalt, and T. Beeler, “DCGrasp: Distance-aware controllable grasp generation,” *arXiv preprint*, 2025.
- [15] J. Lee, E. Park, and M. Cho, “DextER: Language-driven dexterous grasp generation with embodied reasoning,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2025.
- [16] K. Zhu, H. Jiang, J. Qian, and J. Xie, “A cross-view fusion framework for robust 6-DoF grasp pose estimation,” *arXiv preprint*, 2025.
- [17] H.-S. Fang, C. Wang, H. Fang, M. Gou, J. Liu, H. Yan, W. Liu, Y. Xie, and C. Lu, “AnyGrasp: Robust and efficient grasp perception in spatial and temporal domains,” *IEEE Trans. Robot.*, vol. 39, no. 5, pp. 3929-3945, 2023.
- [18] G. Li, D. Sun, L. Sevilla-Lara, and V. Jampani, “One-shot open affordance learning with foundation models,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 3086-3096, 2024.
- [19] S. Li, S. Bhagat, J. Campbell, Y. Xie, W. Kim, K. Sycara, and S. Stepputtis, “Shapegrasp: Zero-shot task-oriented grasping with large language models through geometric decomposition,” *arXiv preprint arXiv:2403.18062*, 2024.
- [20] Y.-L. Wei, J.-J. Jiang, C. Xing, X.-M. Wu, H. Li, M. Cutkosky, and W.-S. Zheng, “Grasp as you say: Language-guided dexterous grasp generation,” *arXiv preprint arXiv:2405.19291*, 2024.
- [21] S. Levine, P. Pastor, A. Krizhevsky, J. Ibarz, and D. Quillen, “Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection,” *Int. J. Robot. Res.*, vol. 37, no. 4-5, pp. 421-436, 2018.
- [22] J. Mahler et al., “Dex-Net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics,” in *Proc. Robot.: Sci. Syst.*, 2017.
- [23] S. Caldera, A. Rassau, and D. Chai, “Review of deep learning methods in robotic grasp detection,” *Multimodal Technol. Interact.*, vol. 2, no. 3, p. 57, 2018.
- [24] R. Wang, J. Zhang, J. Chen, Y. Xu, P. Li, T. Liu, and H. Wang, “Dexgraspnet: A large-scale robotic dexterous grasp dataset for general objects based on simulation,” in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, pp. 11359-11366, 2023.
- [25] T. Liu, Z. Liu, Z. Jiao, Y. Zhu, and S.-C. Zhu, “Synthesizing diverse and physically stable grasps with arbitrary hand structures using differentiable force closure estimator,” *IEEE Robot. Autom. Lett.*, vol. 7, no. 1, pp. 470-477, 2021.
- [26] H. Luo, W. Zhai, J. Zhang, Y. Cao, and D. Tao, “Learning affordance grounding from exocentric images,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 2252-2261, 2022.
- [27] M. Liu, Y. Zhu, H. Cai, S. Han, Z. Ling, F. Porikli, and H. Su, “Part-slip: Low-shot part segmentation for 3d point clouds via pretrained image-language models,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 21736-21746, 2023.
- [28] A. Abdelreheem, I. Skorokhodov, M. Ovsjanikov, and P. Wonka, “SATR: Zero-shot semantic segmentation of 3d shapes,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 15166-15179, 2023.
- [29] A. Radford et al., “Learning transferable visual models from natural language supervision,” in *Proc. Int. Conf. Mach. Learn. (ICML)*, pp. 8748-8763, 2021.
- [30] C. Finn, I. Goodfellow, and S. Levine, “Unsupervised learning for physical interaction through video prediction,” in *Adv. Neural Inf. Process. Syst.*, 2016.