IFG: Internet-Scale Guidance for Functional Grasping Generation

Ray Muxin Liu* Mingxuan Li* Kenneth Shaw Deepak Pathak

Fig. 1: IFG enables the generation of dexterous, functional grasps in cluttered, realistic scenes. It first uses a vision-language model to identify task-relevant regions on objects, then uses geometrically precise force closure in simulation to ground the finger joints. The resulting dataset, and the diffusion model trained on it, encode both semantic and geometric understanding of the scene without any hand-collected data.

***Abstract*—Large Vision Models trained on internet-scale data have demonstrated strong capabilities in segmenting and semantically understanding object parts, even in cluttered scenes. However, while these models can direct a robot toward the general region of an object, they lack the geometric understanding required to precisely control dexterous robotic hands for 3D grasping. To overcome this, our key insight is to leverage simulation with a force-closure grasping generation pipeline that understands local geometries of the hand and object in the scene. Because this pipeline is slow and requires ground-truth observations, the generated dataset is distilled into a diffusion model that can operate on camera point clouds. By combining the global semantic understanding of internet-scale models with the geometric precision of a simulation-based locally-aware force-closure, IFG achieves high-performance semantic grasping without any manually collected training data. For visualizations, please visit our website at https://ifgrasping.github.io/**

I. INTRODUCTION

Recent advances in vision-language models (VLMs) have led to impressive results across a range of perception tasks, including image captioning, visual question answering, and open-world object recognition. Trained on large-scale datasets pairing images with natural language, these models exhibit a strong ability to align visual and linguistic information, enabling semantic understanding that generalizes across diverse contexts. This success has inspired interest in leveraging VLMs for robotics applications such as instruction following, semantic goal specification, and high-level planning. While these initial applications show promise, significant limitations remain. Most notably, current VLMs lack a grounded understanding of physical space—they cannot reliably reason about 3D geometry, spatial relationships, or the dynamics of physical interaction. Consequently, they struggle with planning or executing precise motor actions in the real world. Although VLMs can identify visual content, they do not inherently understand how to interact with it. Addressing this disconnect between perception and control is a major challenge in robotic grasping systems.

We seek an approach that avoids manual data collection through means like teleoperation while enabling geometric understanding. Synthetic grasp generation is promising because it can produce large datasets of grasp poses through an optimization process guided by energy functions that approximate force closure, along with evaluation pipelines in simulation. These datasets are often used to train diffusion-based grasp samplers. However, a significant portion of the generated grasps are physically implausible or unnatural. Because grasp proposals are initialized by sampling points around the object's convex hull, many grasps target physically inaccessible or unsuitable regions.

Moreover, downstream manipulation tasks require the hand to interact with specific, task-relevant regions of objects, such as a handle or button. Existing synthetic grasping pipelines generate grasps indiscriminately over the object surface, leading to datasets that are poorly aligned with the needs of task-conditioned manipulation. Our approach addresses this gap by combining the high-level VLM-based semantic priors with physically grounded, task-aware synthetic grasp generation. To this end, we propose a pipeline that first translates semantic input specifying a task into predictions of useful regions on objects using a VLM. Then, we seed the grasp generation

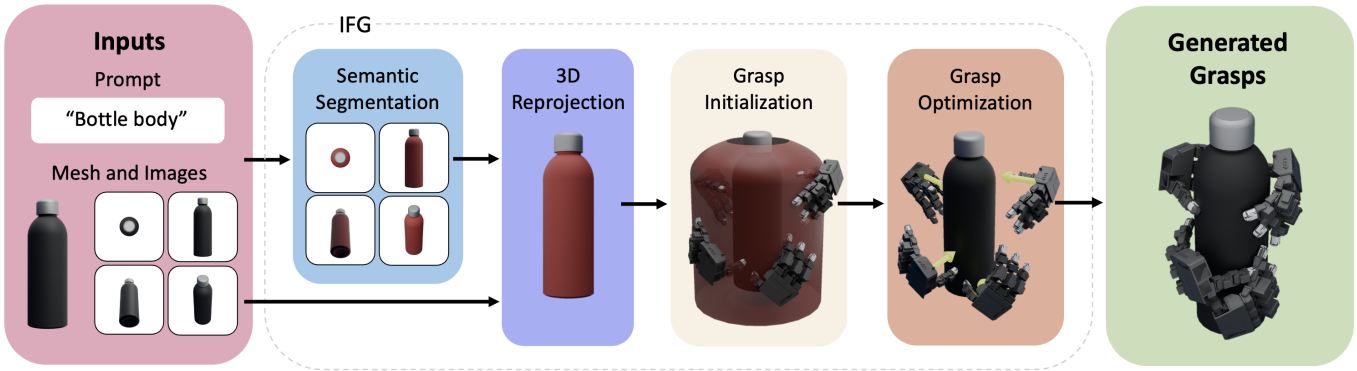

Fig. 2: IFG takes an object mesh and a task prompt as input. To incorporate semantic understanding, it renders the object from multiple viewpoints, applies a VLM-based segmentation model combining SAM[1] and VLPart[2], and reprojects the results into 3D space to identify task-relevant regions. For geometric grounding, it initializes a force closure objective at these regions and optimizes for functional grasps. The resulting data is then used to train a diffusion model for fast grasp synthesis from depth.

process on this prior to enable semantic-guided grasp synthesis, producing stable, natural grasps aligned with the demands of the task. Our pipeline is highly parallelizable, efficient, and compatible with arbitrary objects, scenes (including cluttered environments), and dexterous hands. This pipeline generates semantically meaningful grasps without any teleoperation or video data.

II. METHOD

The goal of IFG is to efficiently generate a large grasping dataset with geometrically accurate and semantically meaningful grasps of a robotic hand and distill it into a general-purpose model that predicts feasible grasps in a scene. A dexterous grasp is defined as $g = (T, R, \theta)$, where $T \in \mathbb{R}^3$ and $R \in SO(3)$ denote wrist translation and rotation, and $\theta \in \mathbb{R}^{\text{DoF}}$ denotes hand joint angles (DoF $= 16$ for LEAP Hand [4]).

A. Useful Region Proposal

IFG leverages knowledge from a VLM f to identify objects of interest and part-level regions for grasping, which are called useful regions. To extract 2D semantic knowledge to 3D scenes, we capture n RGB images from angles uniformly sampled on a camera initialization surface S. For single objects, S is spherical, while for cluttered scenes it is a dome to reduce occlusion. A VLM f is prompted to produce semantic labels, which guide a language-conditioned segmentation model (SAM [1]) and a part-level model (VLPart [2]) to produce segmentation masks of useful regions. The resulting 2D masks are deprojected to 3D points on the object mesh. To account for occlusion errors, we filter points using a two-means clustering process based on segmentation mask size. Valid deprojected points are mapped to the closest mesh faces. A voting algorithm then selects the top 60% of faces as the useful region U.

B. Geometric Grasp Synthesis

We compute the segmented convex hull of the object to include only faces projected from U. For each grasp, the hand is initialized on the inflated convex hull by farthest point sampling with noise added to wrist pose and joint angles. An optimization process performs gradient descent against an energy term

$$E = E_{\text{fc}} + w_{\text{dis}} E_{\text{dis}} + w_{\text{joints}} E_{\text{joints}} + w_{\text{pen}} E_{\text{pen}} + w_{\text{spen}} E_{\text{spen}}$$

where E_{fc} approximates force closure of the grasp, E_{dis} encourages hand-object proximity, based on the contact points of the hand, E_{joints}, E_{pen}, and E_{spen} respectively penalizes joint violations, hand-object penetration, and self-penetration of the hand. For the single object setting, we exclude the tabletop by setting $w_{\text{spen}} = 0$ to produce more diverse grasps. Relative to Get a Grip, we make two key changes: replace precision grasps with power grasps that utilize the inside regions of all fingers instead of fingertip-only and initialize on the segmented convex hull rather than the full hull. These changes improve stability and functional alignment.

C. Simulation Evaluation

To ensure the robustness of generated grasps, we perform tasks with them in a simulation environment. Each evaluation proceeds in three phases: (1) the grasp and object are initialized in a simulation environment, (2) fingers are closed to secure the object, and (3) task execution is performed. Following Get a Grip, we assign each grasp a smooth label by perturbing its joint angles to generate d additional grasps, evaluating all $d+1$ variants, and averaging their binary success outcomes. Grasps with low smooth success are discarded, yielding a dataset G of robust force-closure power grasps. In our experiments, $d = 5$.

D. Diffusion Model Distillation

While the grasping pipeline produces diverse and semantically meaningful candidate grasps, it is too slow for direct inference and depends on privileged mesh information. Hence, we distill the grasps into a generative neural network. We generate 2.5 million grasps using our pipeline on DexGraspNet2's scene dataset. Based on the end effector position, partial pointcloud observations are captured for each grasp to simulate a depth camera input, which are converted to basis point sets (BPS), a structured point-cloud representation [5] shown in [3] to be a better representation to train the model on. The resulting

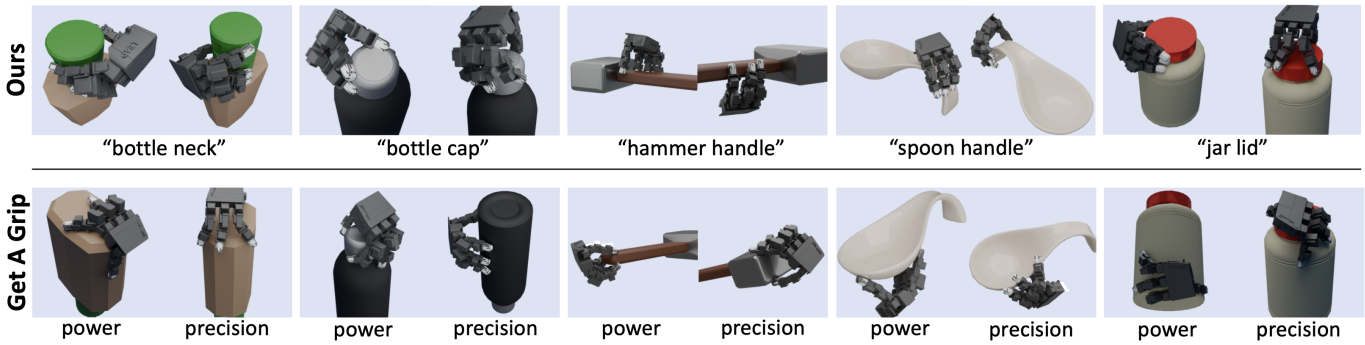

Fig. 3: Compared to Get a Grip's synthetic grasp generation method, our method produces more human-like grasps. For instance, Get a Grip often grasps the bottom of the bottle, while our method knows to robustly grasp the neck. Please see our website for 3D visualizations.

dataset is used to train a diffusion model that generates grasps from basis point set input. The resulting model inherits both the geometric reasoning capabilities of the training pipeline and the semantic understanding provided by the vision-language model (VLM), as illustrated in Figure 4.

Grasp Optimization Pipeline	**Single (%)**	**Cluttered (%)**
Ours		
single camera only	47.83	18.53
+ multi-camera around object	48.37	24.70
+ two-means clustering	49.04	31.59
+ voting-based filtering (full pipeline)	**51.11**	**32.23**
Reference baseline	50.93	14.58

TABLE I: Incremental improvements from our synthetic grasp generation pipeline in single-object and cluttered-scene settings in terms of success rate under the Lift metric. For single-object evaluation, the reference baseline is Get a Grip's synthetic pipeline; for cluttered scenes, the reference baseline is DexGraspNet2's synthetic pipeline.

III. Experimental Setup

Datasets of grasps are generated on diverse objects in both single-object and clustered-scene settings, followed by extensive simulation to evaluate robustness. The evaluation addresses three key questions: (1) Can robust and stable grasps be produced on individual objects? (2) In clustered scenes, can the object of interest be identified and grasped without collision? (3) Do the resulting grasps exhibit natural, human-like qualities suitable for functional manipulation?

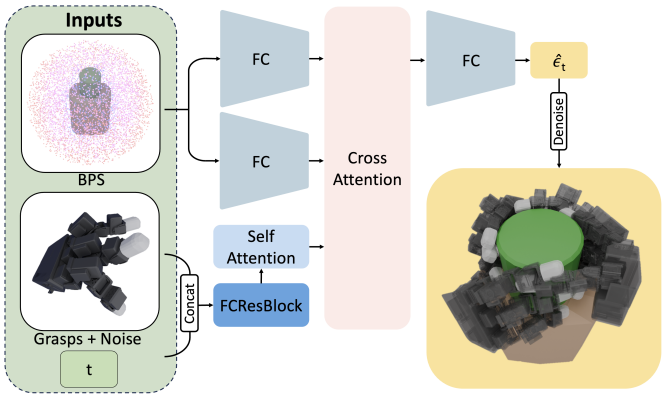

Fig. 4: The generated grasp data is distilled into a diffusion model that only expects proprioceptive observation inputs obtainable from a single depth camera. The model converts depth camera data into a Basis Point Set (BPS) and refines randomly sampled noise into functional grasps on semantically meaningful regions of the object. The architecture of the diffusion model follows a similar design to Get a Grip [3].

Object	DexGraspNet2	GraspTTA	ISAGrasp	Ours
Selected Individual Objects				
Tomato Soup Can	47.8	38.3	**52.0**	45.5
Mug	33.2	26.9	22.6	**60.4**
Drill	32.1	20.8	36.4	**57.5**
Scissors	9.7	0.0	**33.7**	20.2
Screw Driver	0.0	8.3	**40.0**	22.0
Shampoo Bottle	50.6	25.4	18.8	**53.1**
Elephant Figure	23.6	29.6	24.2	**35.8**
Peach Can	**61.8**	28.0	55.3	60.3
Face Cream Tube	32.1	22.5	20.7	**35.5**
Tape Roll	22.7	13.9	9.8	**43.2**
Camel Toy	12.8	14.3	21.3	**21.8**
Body Wash	40.2	22.3	29.4	**58.3**
Object Average	30.55	20.86	30.35	**42.80**
Scene Average	**36.71**	25.64	32.51	34.16

TABLE II: Trained grasp generation model success rates for crowded-scene evaluation on the lift task. Both success rates of grasps on selected challenging objects and of all grasps across all scenes are reported. IFG significantly outperforms baselines on difficult objects and has comparable performance on scene average success.

Task Setup We evaluate 24 single objects from Get a Grip at 5 scales, generating 200 grasps per method, and 35 dense cluttered scenes as a test set from DexGraspNet2 with 256 grasps per scene. All objects are drawn from common daily manipulation tasks, and all grasps are executed using the LEAP Hand [4]. Please also visit our website at https://ifgrasping.github.io/ for more visualizations of these results.

Simulation Evaluation. Grasps are tested in IsaacGym [6]. Single-object tasks include Lift (vertical translation) and Pick & Shake (lifting with perturbations). Success requires the object's relative pose to the palm to remain stable without interpenetration. Clustered scenes are evaluated only on Lift to avoid trivial collisions.

IV. Results

A. Grasp Generation Pipeline

Single-view VLM segmentation is prone to occlusion errors, such as hidden mug handles. Therefore, we capture images

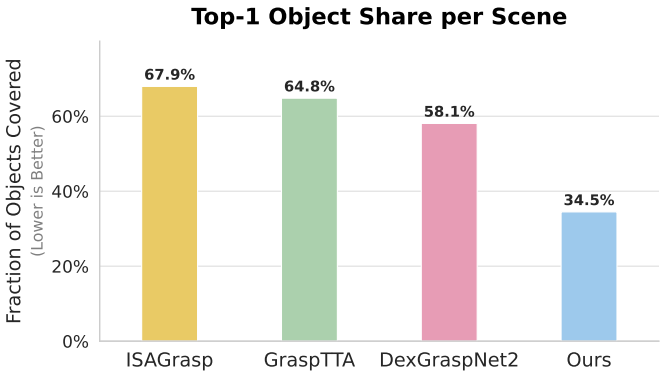

Fig. 5: When generating grasps, confidence-based methods sample a pool of grasps and select the most confident one for the scene as the final prediction. This causes most output grasp predictions to be concentrated on the easiest-to-grasp object. On the other hand, our grasp prediction model has a much more balanced coverage of the objects and a more informative success rate.

from multiple cameras in parallel and apply geometric filtering. Table I demonstrates that each additional technique improves the success rate of our synthetic pipeline, allowing our method to beat baseline synthetic methods by a significant margin in cluttered scenes.

B. Single Object Grasping

IFG produces grasps that are both robust and natural. As illustrated in Table I, IFG achieves a higher success rate over the reference baseline for the single-object setting, demonstrating that conditioning on part-level segmentation produces more robust grasps. Qualitative comparisons in Figure 3 demonstrate our grasps concentrate on functionally relevant regions, whereas baseline grasps from Get a Grip often target unhelpful regions simply because they cover a high percentage of the convex hull. For example, their grasps tend to grasp the head of a hammer since it covers a high percentage of the convex hull, while our grasps, initialized on the segmented convex hull of the handle, are functionally correct. We hypothesize that semantic conditioning inherently improves robustness because everyday objects are designed with affordances that support secure functional grasping.

C. Multi-object Dense Scene Grasping

Daily scenarios often involve cluttered scenes, requiring precise identification of objects and firm grasps while avoiding collision. We evaluated against recent baselines [7, 8, 9] on 35 dense test scenes with 256 samples per scene. Shown in Table II, our model achieves a highly competitive global scene average success rate. Figure 1 shows our grasps of four scenes. The baselines' confidence-based sampling biases the model heavily toward easy, peripheral objects. As shown in Figure 5, the baseline concentrates output on the easiest targets and rarely proposes grasps on harder objects. In contrast, our method avoids this overfitting by not favoring particular objects in a scene. Therefore, when evaluated on a subset of challenging objects across scenes, our model significantly outperforms the baselines, demonstrating better generalization across objects. DexGraspNet2 [7] reports higher overall success rates in their paper on a relaxed 3cm lift threshold, while ours is 20cm.

V. Conclusion and Limitations

We introduced IFG, a massively parallelizable pipeline combining internet-scale VLM semantics with geometric force-closure optimization to generate robust, functional grasps in cluttered environments without manual data collection. Distilled into a diffusion model, it infers generalized grasps from depth input. More broadly, our system illustrates how a modular perception pipeline built on foundation-model semantics can improve interpretability in contact-rich manipulation, since the useful regions, filtering stages, and downstream grasp outcomes can inspected and evaluated separately. Nonetheless, our work has limitations. Currently, our method is limited to static image segmentation and force-closure grasps; extending this to continuous video streaming for dynamic manipulation remains an exciting area for future work.

VI. Acknowledgments

We thank Jason Liu, Andrew Wang, Yulong Li, Jiahui (Jim) Yang, Sri Anumakonda for helpful discussions and feedback. This work was supported in part by the Air Force Office of Scientific Research (AFOSR) under Grant No. FA955023-1-0747 and by the Office of Naval Research (ONR) MURI under Grant No. N00014-24-1-2748.

VII. Contributions

Ray Muxin Liu implemented the grasp-region prompting and 3D deprojection scheme, the grasp optimization module, as well as the clutter-scene grasping experiments and benchmarking in simulation, and led the manuscript writing.
Mingxuan Li developed the early-stage components, including the grasp-generation framework, the grasp-diffusion model, and the 2D segmentation of grasp regions. Upon graduating, Mingxuan Li concluded active contributions to the project
Kenneth Shaw originated the core idea and guided the research direction.
Deepak Pathak supervised the project.

References

[1] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4015–4026.

[2] P. Sun, S. Chen, C. Zhu, F. Xiao, P. Luo, S. Xie, and Z. Yan, "Going denser with open-vocabulary part segmentation," 2023. [Online]. Available: https://arxiv.org/abs/2305.11173

[3] T. G. W. Lum, A. H. Li, P. Culbertson, K. Srinivasan, A. D. Ames, M. Schwager, and J. Bohg, "Get a grip: Multi-finger grasp evaluation at scale enables robust sim-to-real transfer," *arXiv preprint arXiv:2410.23701*, 2024.

[4] K. Shaw, A. Agarwal, and D. Pathak, "Leap hand: Low-cost, efficient, and anthropomorphic hand for robot learning," *arXiv preprint arXiv:2309.06440*, 2023.

[5] S. Prokudin, C. Lassner, and J. Romero, "Efficient learning on point clouds with basis point sets," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 4332–4341.

[6] V. Makoviychuk, L. Wawrzyniak, Y. Guo, M. Lu, K. Storey, M. Macklin, D. Hoeller, N. Rudin, A. Allshire, A. Handa *et al.*, "Isaac gym: High performance gpu-based physics simulation for robot learning," *arXiv preprint arXiv:2108.10470*, 2021.

[7] J. Zhang, H. Liu, D. Li, X. Yu, H. Geng, Y. Ding, J. Chen, and H. Wang, "Dexgraspnet 2.0: Learning generative dexterous grasping in large-scale synthetic cluttered scenes," in *8th Annual Conference on Robot Learning*, 2024.

[8] H. Jiang, S. Liu, J. Wang, and X. Wang, "Hand-object contact consistency reasoning for human grasps generation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 11 107–11 116.

[9] Z. Q. Chen, K. Van Wyk, Y.-W. Chao, W. Yang, A. Mousavian, A. Gupta, and D. Fox, "Learning robust real-world dexterous grasping policies via implicit shape augmentation," *arXiv preprint arXiv:2210.13638*, 2022.