2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Benchmarking the Effects of Object Pose Estimation and Reconstruction on Robotic Grasping Success

Varun Burde[1,2], Pavel Burget[2], Torsten Sattler[2]

***Abstract*— 3D reconstruction serves as the foundational layer for numerous robotic perception tasks, including 6D object pose estimation and grasp pose generation. Modern 3D reconstruction methods for objects can produce visually and geometrically impressive meshes from multi-view images, yet standard geometric evaluations do not reflect how reconstruction quality influences downstream tasks such as robotic manipulation performance. This paper addresses this gap by introducing a large-scale, physics-based benchmark that evaluates 6D pose estimators and 3D mesh models based on their functional efficacy in grasping. We analyze the impact of model fidelity by generating grasps on various reconstructed 3D meshes and executing them on a high-fidelity reference model, simulating how grasp poses generated with an imperfect reconstruction affect physical interaction. This assesses the combined impact of pose error, grasp robustness, and geometric inaccuracies from 3D reconstruction. Our results show that reconstruction artifacts significantly decrease the number of grasp pose candidates but have a negligible effect on grasping performance given an accurately estimated pose. Our results also reveal that the relationship between grasp success and pose error is dominated by spatial error, and even a simple translation error provides insight into the success of the grasping pose of symmetric objects. This work provides insight into how perception systems relate to object manipulation using robots.**

I. Introduction

The ambition for robots to autonomously operate in human-centric environments is a primary driver of robotics research. A prerequisite for meaningful interaction is the ability to perceive and manipulate objects, which requires both knowing an object's 6D pose (position and orientation) and understanding its geometry. This 3D model serves a basis for model-based 6D pose estimation methods that determine an object's position and orientation and it is the representation upon which grasp poses are generated for physical interaction. While deep learning has led to remarkable progress in 6D pose estimation [1], [2] and 3D reconstruction [3], [4], these perception components are typically evaluated in isolation.

Progress in pose estimation is measured by geometric metrics like ADD (Average Distance of Model Points - Symmetric) on benchmarks like BOP [5], while reconstruction quality is assessed by metrics such as Chamfer distance. However, this decoupled evaluation creates a significant gap, and it is unclear how errors from pose estimation and geometric reconstruction compound and propagate to affect the success of downstream manipulation tasks like grasping. For a robot, the utility of a perception system is not defined by its geometric precision alone, but by its functional efficacy, an idea shared by other recent benchmarking efforts [6].

This paper directly addresses this disconnect by evaluating perception systems based on a robot's ability to grasp the object. We introduce a large-scale physics simulation study connecting errors from 6D pose estimation and 3D reconstruction to robotic grasping success, proposing a new evaluation paradigm contextualized by manipulation. By simulating millions of grasp attempts under mismatched geometry conditions, we measure the probability of task success as a function of both the underlying pose error and the geometric fidelity of the object model. This large-scale evaluation uncovers the hidden flaw of perception pipelines, showing how errors considered negligible by standard metrics can decisively impact downstream grasp execution.

By simulating millions of grasp attempts under these mismatched conditions, we measure the probability of task success as a function of both the underlying pose error and the geometric fidelity of the object model. This large-scale evaluation uncovers the hidden flaw of perception pipelines, showing how errors considered negligible by standard metrics can decisively impact downstream grasp execution

Our contributions are threefold:

- We introduce a comprehensive framework for functionally evaluating the combined impact of 6D pose estimation and 3D reconstruction errors on robotic grasping.
- We conduct the first large-scale quantitative analysis of grasp success utilizing 3D reconstructed object models for pose estimation and grasp pose generation, revealing the performance degradation caused by geometric inaccuracies.
- We present a task-based re-evaluation of modern perception systems, including 3D reconstruction, object pose estimation, and grasp pose generation, providing crucial insights into their practical utility and failure modes for real-world manipulation.

II. Related Work

Our research is situated at the intersection of 6D object pose estimation, 3D reconstruction, and robotic grasping. We address the critical gap in their unified evaluation by examining how errors from both perception domains propagate to a functional manipulation task.

[1]Faculty of Electrical Engineering, Czech Technical University in Prague, Czechia.
[2]Czech Institute of Informatics, Robotics and Cybernetics, Czech Technical University in Prague, Czechia.

A. 6D Object Pose Estimation and Benchmarks

Modern 6D pose estimation has shifted from classic feature-based methods [7] to sophisticated learning-based approaches. Methods like PoseCNN [8] and DenseFusion [9] demonstrated the power of deep learning, while recent zero-shot systems like MegaPose [1] and FoundationPose [2] have achieved remarkable generalization to novel objects. This progress has been accelerated by standardized benchmarks, most notably the BOP challenge [5], which evaluates methods using task-agnostic geometric metrics like ADD and MSSD (Maximum Symmetry-Aware Surface Distance). While instrumental, these metrics do not capture how pose errors affect physical interaction.

B. 3D Reconstruction for Robotics

Simultaneously, 3D reconstruction from multiview RGB images has seen significant advances, particularly with neural implicit representations. Methods like NeRF (Neural Radiance Fields) [3], [10] and its variant for implicit representation, such as NeuS [4], Volsdf [11], Monosdf [12] can produce high-fidelity meshes from multiview images. For robotics, these reconstructions serve as the geometric basis for tasks like object pose estimation and grasping. Benchmarks such as [13], [14] evaluate the 3D reconstruction in terms of geometric accuracy, and the [15] evaluates these methods one step further by evaluating the performance for object pose estimation. However, like pose estimation benchmarks, they do not assess the functional suitability of the resulting meshes for manipulation. A reconstructed mesh with low geometric error might still possess artifacts such as smoothed edges or filled holes that are critical for stable grasping.

C. Robotic Grasping and the Perception-Action Gap

Robotic grasping research has evolved from analytical, model-based approaches [16] to data-driven techniques that learn grasping policies from large datasets [17]. While systems like Dex-Net 2.0 [18] show impressive performance by learning robust grasp policies from millions of synthetic examples, they often assume access to high-quality point clouds or object models. The development of common object sets and protocols, such as the widely-used YCB Object and Model Set [19], provides the necessary foundation for researchers to evaluate the connection between perception and manipulation.

However, a gap persists in understanding how modern perception errors (pose and geometry) jointly affect manipulation. Our work addresses this by measuring this relationship directly, providing a quantitative analysis of how perceived world models impact grasping success. Instead of proposing mitigation methods, we offer a rigorous methodology to characterize the problem, building an empirical basis for robust manipulation systems.

III. Methodology

To systematically quantify how mesh geometric and 6D pose estimation errors propagate to robotic grasping outcomes, we designed a comprehensive benchmarking framework within the PyBullet [20] physics simulator. Simulation is essential here: our study requires millions of controlled grasp trials across combinatorial conditions (9 grippers $\times$ 21 objects $\times$ 8 reconstruction methods $\times$ 2 pose estimators), a scale infeasible with physical hardware. Physics simulation allows us to isolate perception errors from uncontrolled real-world factors (e.g., sensor noise, table clutter), providing a controlled upper bound on grasping performance. Our methodology is centered around a core transformation chain that links perception to action. We first establish a baseline of ideal grasping performance for a library of grippers and objects. Then, we evaluate the degradation of this performance on a binary grasp success task. The overall pipeline is visualized in Fig. 1.

A. Core Transformation Chain

The link between perception and robotic action is defined by a sequence of rigid body transformations. Let the primary coordinate frames be World (W), Camera (C), Object (O), and Gripper (G). We define the following homogeneous transformations:

- T_{w2c}: The ground-truth pose of the camera in the world frame, known from the dataset.
- T_{c2o}^{gt}: The ground-truth pose of the object in the camera frame, provided by the BOP dataset annotations.
- T_{c2o}^{est}: The pose of the object in the camera frame as predicted by a 6D pose estimation method.
- T_{o2g}: A pre-computed, canonical grasp pose, defining the gripper's pose relative to the object's local coordinate frame.

Using this chain, we can compute the target pose for the gripper in the world frame. The ideal gripper pose, derived from the ground-truth object pose, is:

$$T_{w2g}^{gt} = T_{w2c} \cdot T_{c2o}^{gt} \cdot T_{o2g} \tag{1}$$

Conversely, the gripper pose that a robot would actually target in a real-world scenario, based on the perception system's output, is:

$$T_{w2g}^{est} = T_{w2c} \cdot T_{c2o}^{est} \cdot T_{o2g} \tag{2}$$

The core of our methodology is to execute grasps using the target pose T_{w2g}^{est} but to evaluate the physical interaction with the object located at its true pose, dictated by $T_{w2c} \cdot T_{c2o}^{gt}$. This setup precisely simulates the real-world scenario where a robot acts based on imperfect perception.

B. Simulation Environment and Asset Preparation

All experiments are conducted in the PyBullet simulator. We utilize the object meshes from the YCB-Video dataset and nine distinct, widely-used robotic end-effector models provided by the `burg-toolkit`[1] [21]: the Franka Hand, Robotiq 2F-85, Robotiq 2F-140, WSG 32, WSG 50, EZGripper, Sawyer Hand, Kinova 3F, and Robotiq 3F.

[1] https://mrudorfer.github.io/burg-toolkit/

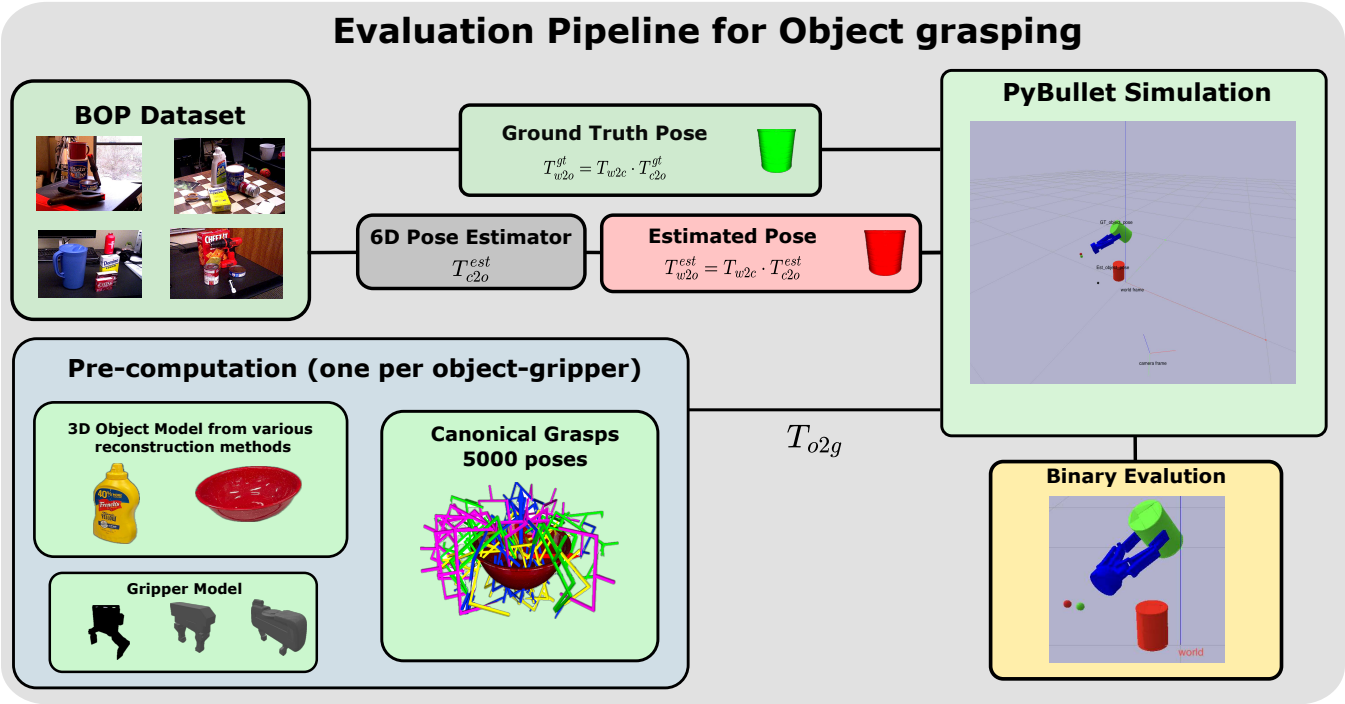

Fig. 1. Overview of our evaluation pipeline. First, a canonical grasp library is pre-computed for each object. Then, for a given scene, a pose estimator provides T^{est}_{c2o}. This pose is used to derive a target gripper pose, T^{est}_{w2g}, which is executed on the ground-truth object. The outcome is recorded to calculate the Estimated Success Rate (S_{est}) (Sec. III-D.2) and correlated with the initial pose error.

The physics simulation is tightly controlled. We spawn two YCB-V object instances: a ground-truth (GT) object rendered in green with physics enabled, and a red estimated (EST) visual reference (Fig. 1). To isolate perception errors from complex contact dynamics, we fix the GT object friction to 0.5, a value representative of common plastic-on-rubber contacts. This deliberately reduces purely dynamic slip failures, focusing the analysis on geometry and pose driven failures (collisions, no-contact). We acknowledge that lower friction values would increase slip rates and potentially shift the relative importance of geometry vs. pose; this represents a controlled design choice, not an oversight. The simulation runs at 240 Hz with 100 iterations. No ground plane is used; objects float based on pose constraints. Gravity is initially disabled to prevent premature movement and enabled at $-9.81\,m/s^2$ after gripper closure to test stable lifts. This isolates the perception system's output and object geometry from environmental clutter.

C. Evaluation Protocol

Our evaluation protocol is designed to answer a central question: how does 3D model accuracy impact robotic grasping performance? In a practical scenario, a robot uses a reconstructed 3D model for two key tasks: to estimate the 6D pose of an object and to generate a set of viable grasp poses. The robot then executes one of these grasps on the real-world object.

1) Experimental Conditions: To systematically analyze the different sources of error in this pipeline, we evaluate performance under four distinct conditions. In all cases, the grasp is executed on the ground-truth (GT) object model in the simulator, representing the physical reality. We define the *Oracle Mesh* as the high-fidelity, ground-truth CAD model provided by the YCB-V dataset. We denote each condition as *Grasp Mesh* $\rightarrow$ *Pose Mesh*:

- **Oracle Mesh for Grasps & Pose (Ideal Baseline):** This condition establishes the best-case performance with "perfect" information. We use the oracle CAD mesh to both generate the library of canonical grasps and as the reference model for the 6D pose estimator.
- **Oracle Mesh for Grasps & Reconstructed Mesh for Pose (Isolating Pose Error):** This setup isolates the impact of a reconstructed model's geometry on *pose estimation accuracy*. The robot plans grasps using a "perfect" internal oracle mesh but uses the imperfect reconstructed mesh to find the object in the scene.
- **Reconstructed Mesh for Grasps & Pose (End-to-End Realistic Scenario):** This represents the most realistic case. The robot uses the same imperfect, reconstructed 3D model for both generating grasp candidates and as the reference for pose estimation. This measures the compounded effect of errors from both stages.

2) Trial Procedure: The general procedure for each trial is as follows:

1) For an object instance in a scene, retrieve its estimated pose T^{est}_{c2o} and the pre-computed library of successful canonical grasps (T_{o2g}).
2) For each canonical grasp, compute the target gripper pose using the estimated object pose via Eq. 2.

3) Execute the grasp in the simulator on the object placed at its ground-truth pose.
4) Record the outcome (success or failure) to calculate our functional metrics.

D. Evaluation Metrics

1) Grasp Generation Success Rate (S_{gen}): This metric is used to evaluate the suitability of a 3D model for the grasp pose sampling stage. It measures the percentage of viable grasp candidates a model yields from a fixed set of randomly sampled grasp poses.

Let N_{total} be the total number of grasp poses sampled for an object. Let N_{succ}^{Model} be the number of those poses that are successful when simulated on a specific `Model` (e.g., the oracle CAD model or a particular reconstructed mesh).

The **Grasp Generation Success Rate** for that model is:

$$S_{gen}^{\text{Model}} = \frac{N_{succ}^{\text{Model}}}{N_{total}} \times 100\%$$

This metric directly quantifies how well a given 3D model's geometry supports the task of finding usable grasps. A higher S_{gen} acts as a proxy for "grasp density" or "grasp coverage". Even though a real robot might only execute a single grasp, ensuring a dense field of valid, kinematically-feasible, and collision-free options is crucial in cluttered environments or for task-oriented grasping, where the robot might need to interact with a specific part of the object.

2) Estimated Success Rate (S_{est}): This is the primary metric for evaluating grasping performance. It measures the probability that a grasp, which is known to be successful with a perfect object pose, will also succeed when using the pose provided by an estimation algorithm.

Let G_{gt} be the set of all grasp poses that are successful when using the ground-truth object pose (T_{c2o}^{gt}). Let $N_{gt} = |G_{gt}|$ be the total number of such successful grasps. When these N_{gt} grasps are executed using the estimated object pose (T_{c2o}^{est}), let N_{succ} be the number of grasps that still succeed. The **Estimated Success Rate** is then defined as:

$$S_{est} = \frac{N_{succ}}{N_{gt}} \times 100\%$$

3) Physics-Based Outcome Breakdown: To provide a detailed diagnosis of failures, grasp attempts in the physics simulation are categorized into the following outcomes:

- **Successful Grasp:** The gripper successfully approaches, establishes a stable hold, and lifts the object against gravity without dropping it.
- **Slipped:** The gripper makes initial contact but the hold is not stable, causing the object to slip during the lift.
- **No Contact:** The gripper's fingers close completely without touching the object, typically caused by a large translation error in the estimated pose.
- **Collision:** The gripper's body collides with the object during approach, preventing a valid grasp from being attempted.

IV. Experiments

a) Reconstruction Methods: To comprehensively evaluate the impact of 3D model fidelity, we utilize meshes generated by a diverse set of state-of-the-art techniques, sourced from the benchmark by Burde et al. [15]. This selection spans multiple paradigms, including neural radiance field (NeRF) methods (Instant NGP [10], NeRFacto [22], Neuralangelo [23]), implicit surface models (UniSurf [24], MonoSDF [25], BakedSDF [12], VolSDF [11]), and commercial photogrammetry software (RealityCapture [26]). Using this broad range of models allows us to analyze how different types of geometric inaccuracies and artifacts affect downstream grasping performance.

A. Dataset and Pose Estimators

We ground our benchmark in a real-world, challenging dataset and evaluate state-of-the-art object pose estimation methods.

a) Dataset: We use the YCB-Video (YCB-V) [19] dataset from the BOP challenge [5]. It is known for significant clutter, occlusion, and lighting variations, and contains 21 objects with diverse geometries, sizes, and symmetries.

b) Pose Estimation Methods: We evaluate two leading pose estimators, both taking a single RGB-D image and a 3D reference mesh as input:

- **MegaPose [1]:** A render-and-compare pipeline that achieves zero-shot performance through coarse-to-fine refinement against rendered views of the reference mesh.
- **FoundationPose [2]:** A unified framework for detection, tracking, and pose estimation, designed for robust novel-object performance.

Crucially, as their 3D geometric reference, these methods take either the oracle CAD mesh or a reconstructed mesh, matching the experimental condition being evaluated. This yields realistic, data-driven error distributions from actual method outputs far more representative of real-world failures than synthetic Gaussian noise on ground-truth poses, which cannot capture method-specific failure modes such as local minima or symmetry confusion.

For every object instance in the YCB-V test set, we use the poses provided by these methods as the input T_{c2o}^{est} to our evaluation pipeline.

V. Results

Our experiments are designed to untangle how reconstructed geometry and the pose estimation errors impact manipulation. We structure our analysis in the following progression: first, we establish a baseline for gripper performance to understand the physical constraints of the task. Second, we isolate the effect of 6D pose estimation error on grasping success. Third, we isolate the effect of 3D model inaccuracies on grasp planning. Finally, we analyze the compounded effect of both pose and model errors to draw conclusions about their relative importance in a realistic end-to-end scenario.

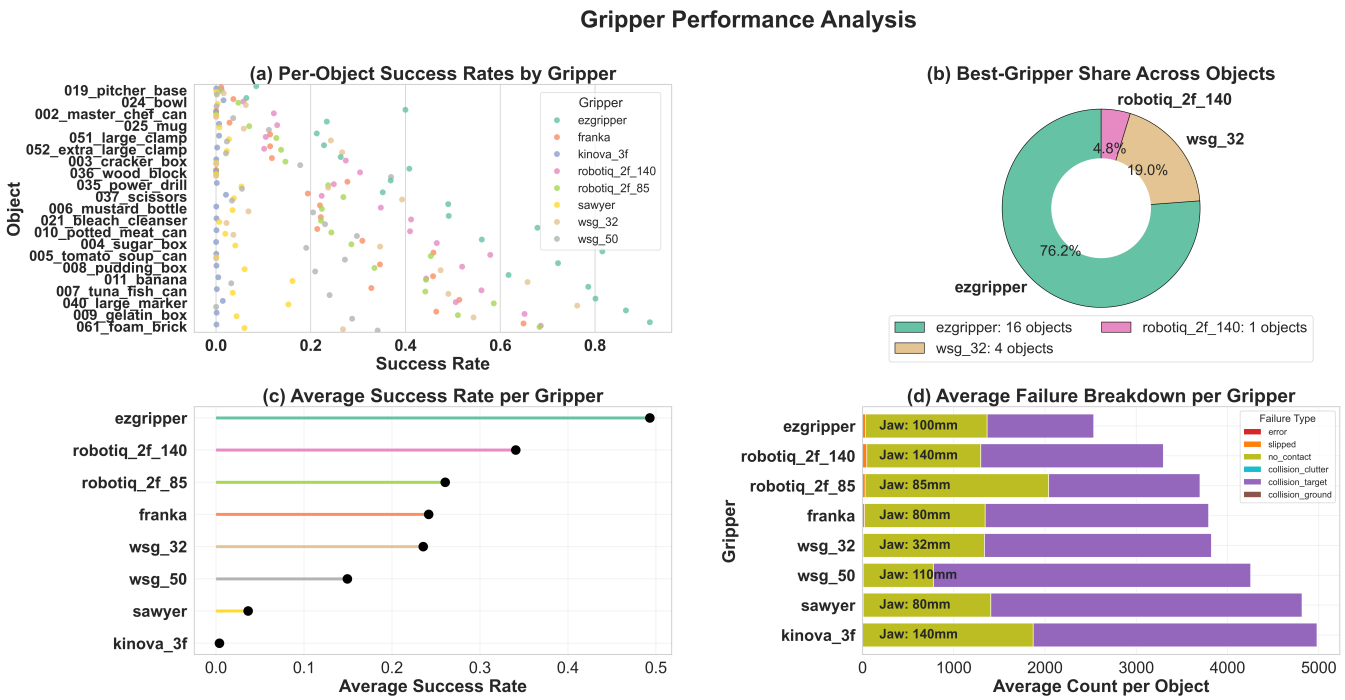

Fig. 2. Baseline gripper performance analysis, visualizing the Grasp Generation Success Rate (S_{gen}) (Sec. III-D.1) across various grippers and objects under ideal conditions. (a) Per-object S_{gen} for each gripper. (b) Distribution of the best-performing gripper for each object. (c) Average S_{gen} per gripper. (d) A Physics-Based Outcome Breakdown (Sec. III-D.3) of grasp failures for each gripper with gripper jaw widths annotated in millimeters.

A. Baseline Gripper Suitability

To understand the impact of gripper design, we first established a baseline of each end-effector's innate capabilities. To do this, we performed a large-scale analysis for each of the 21 object-meshes and 9 gripper models. We sampled 5,000 diverse antipodal grasps for each pair on the object at its canonical identity pose ($T_o^{gt} = I$) and executed them in simulation. Figure 2 presents this analysis, showing the performance of different grippers on the ground-truth object models at the canonical pose.

- **Subfigure (a)** plots the Grasp Generation Success Rate (S_{gen}) (Sec. III-D.1) for each gripper on a per-object basis. It clearly shows that no single gripper is optimal for all objects; performance depends highly on the object's geometry.
- **Subfigure (b)** supports this by showing that the "best" gripper is distributed across several models, with the EZGripper, WSG 32, and Robotiq 2F-140 being the most frequent top performers.
- **Subfigure (c)** aggregates performance, showing the average S_{gen} for each gripper across all objects, providing a general sense of which designs are most versatile.
- **Subfigure (d)** isolates failure modes. Narrower grippers often fail via collisions, while wider ones are more prone to slipping. The controlled friction ($\mu = 0.5$) and ideal pose execution deliberately suppress dynamic slip failures, allowing our benchmark to isolate planning- and perception-driven failures (collisions, no-contact). Real-world deployments with lower friction and surface variability would exhibit higher slip rates.

The primary conclusion from this baseline analysis is critical for our subsequent experiments: since gripper choice heavily influences success, relying on a single end-effector could bias our findings. Therefore, to ensure our conclusions about perception systems are generalizable, all the following experiments aggregate results across the entire library of grippers.

B. Object Pose estimation Error to Grasping Success

A key question is how well standard geometric metrics for pose estimation predict grasping task success. Figure 3 addresses this by correlating pose errors with our functional metric, the Estimated Success Rate (S_{est}) (Sec. III-D.2), under the ideal Oracle→Oracle condition.

The **right panel** provides a high-level summary of grasping performance per object. It shows that a more accurate pose estimator leads to better grasping success. The green bars, representing the final S_{est}, are consistently taller for FoundationPose (89.9% average success) than for MegaPose (59.4%). The failure breakdown reveals why: MegaPose results in a much higher proportion of 'No Contact' and 'Slipped' failures, indicating its pose errors are often large enough to cause the gripper to miss the object or fail to secure a stable hold.

The **left panel** provides deeper insight by plotting S_{est} against six different geometric error metrics. The distinct downward trend in the plots for 3D spatial metrics (MSSD, ADD, ADI, and translation error) demonstrates a strong correlation: as the 3D error increases, the probability of a successful grasp decreases. Conversely, the plots for 2D

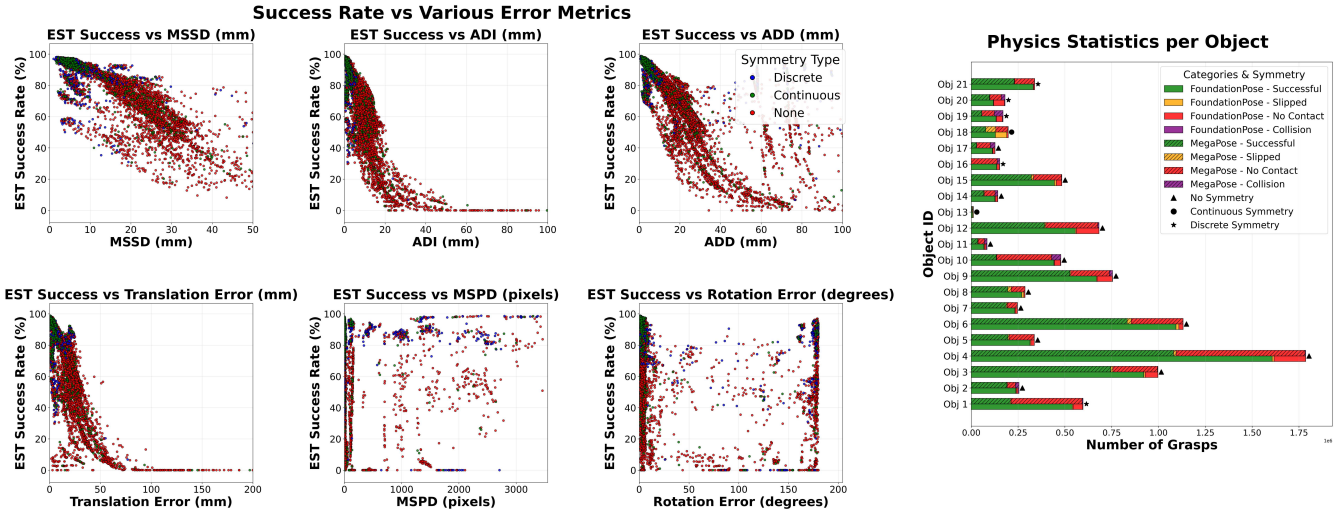

Fig. 3. **Analysis of Grasping Performance vs. Pose Estimation Error. Left Panel:** Scatter plots showing the relation between various pose error metrics and the Estimated Success Rate (S_{est}), averaged over both FoundationPose and MegaPose across 8,250 trials and 18,882,842 simulations (Sec.III-D.2). **Right Panel:** A detailed Physics-Based Outcome Breakdown (Sec.III-D.3) of grasp attempts per object. The green portion of each bar represents the final S_{est}, while other colors show the proportions of different failure modes.

projection error (MSPD) and pure rotation error are much flatter, showing they are poor predictors of grasping success. This finding is significant as it validates that many standard 2D-based metrics do not capture the information most critical for physical interaction, highlighting the necessity of a benchmark like ours.

C. The Impact of 3D Model Fidelity on Grasp sampling

Next, we isolate the effect of geometric inaccuracies on finding the grasp candidates. To remove pose estimation from the equation, we generate grasps on various 3D models at their canonical identity pose. This directly measures how the quality of a 3D model affects the number of viable grasp candidates. We note that our evaluation is task-agnostic and measures overall grasp density; task-specific grasping (e.g., grasping a mug by its handle) would require semantic grasp selection, which is out of scope for this benchmark.

The results are presented in Figure 4. The **left panel** shows a clear performance degradation for most reconstructed models compared to the oracle mesh, as geometric flaws reduce the number of valid grasp poses found by the sampler. The **right panel** reveals the primary reason: for models like Instant-NGP, the dominant failure is 'Collision', meaning the sampler generates poses that physically collide with the true object surface.

Interestingly, Unisurf yields an S_{gen} comparable to or higher than the Oracle mesh. Its smoother surfaces with fewer high-frequency details reduce collisions during tight sampling, highlighting that extreme geometric detail is not always necessary for pure grasp sampling. This also explains results where some reconstructed meshes appear to outperform the oracle the smoothing removes fine-grained surface features that would otherwise cause edge-case collisions.

D. Compounded Errors: Pose and Geometry Inaccuracies Combined

Finally, we analyze the end-to-end realistic scenario, where an imperfect reconstructed model is used to find grasp candidates and estimate pose. Figure 5 compares this compounded error condition (right panel) against a baseline where only pose estimation uses the reconstructed model (left panel).

The results lead to our main conclusion. As established in the previous section, using a reconstructed mesh drastically reduces the number of available grasp candidates. However, Figure 5 shows that as long as a sufficient number of candidates remain, the final grasping success, measured by S_{est} (Sec. III-D.2), is not significantly impacted for a high-quality pose estimator like FoundationPose. The green success bars in the right panel are slightly shorter than those in the left. This demonstrates that while 3D model fidelity is critical for generating a rich set of grasp options, the accuracy of the 6D pose estimate is the primary factor determining the ultimate success of the manipulation task.

VI. Conclusion

In this work, we introduced a benchmark to bridge the gap between standard geometric evaluation of perception systems and their grasping performance in robotic manipulation. Our results provide a clearer understanding of how 3D reconstruction and object pose errors affect grasping success.

Our quantitative analysis reveals a strong correlation between 3D pose error and grasping success: performance degrades sharply with increasing spatial translation error, while the impact of rotation error is more nuanced and strongly object-dependent. For objects with rotational symmetry, errors around the symmetry axis have negligible effect on grasp success, since the grasp geometry is invariant to

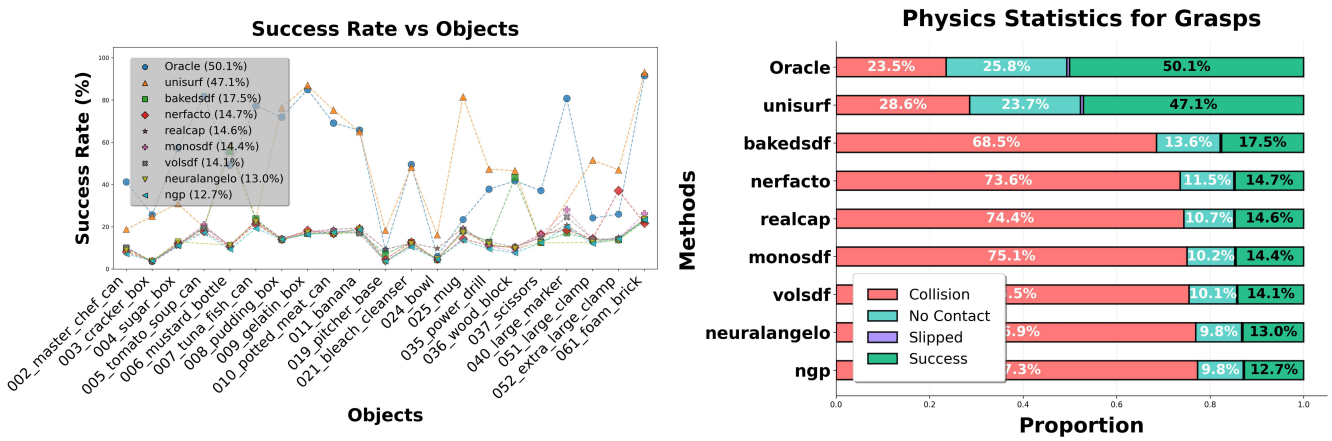

Fig. 4. **Impact of 3D Model Fidelity on Grasp Candidates. Left panel:** The Grasp Generation Success Rate (S_{gen}) (Sec. III-D.1) for various reconstruction methods. **Right panel:** A Physics-Based Outcome Breakdown (Sec. III-D.3) for grasps planned on these meshes. Note the significant increase in 'Collision' failures for lower-quality models.

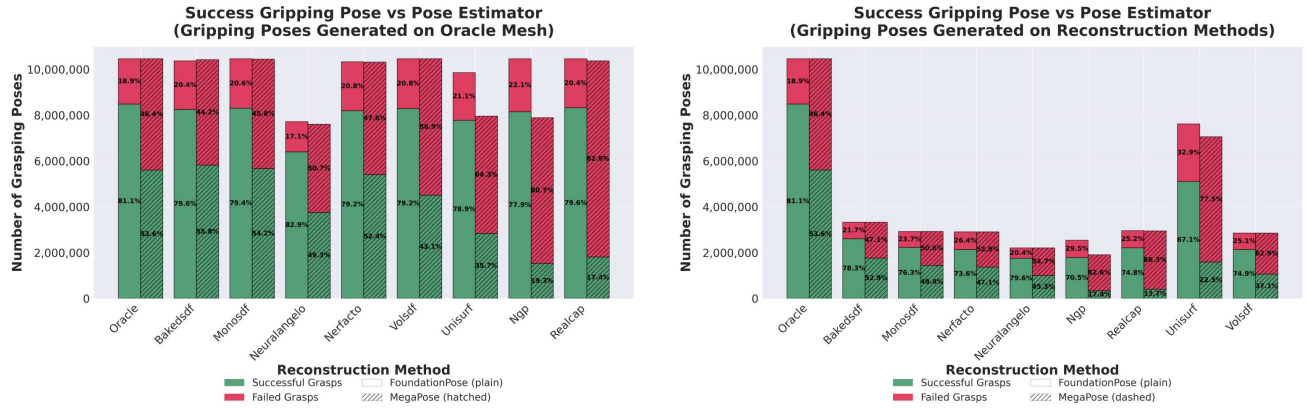

Fig. 5. **Comparative Analysis of Grasping Success under Compounded Errors.** This figure compares the final grasping success, measured by the Estimated Success Rate (S_{est}) (Sec. III-D.2), when combining different sources of geometric and pose uncertainty. **Left:** Performance under the 'Oracle Mesh → Reconstructed mesh' condition. **Right:** Performance under the 'Reconstructed mesh → Reconstructed mesh' condition.

such rotations. For asymmetric objects, however, rotation errors can be equally detrimental. This asymmetry confirms that standard geometric metrics are informative, but only direct functional evaluation exposes the true operational limits of a perception system. Furthermore, using imperfect models as pose references degraded grasping performance, even when grasp candidate generation used a perfect internal model.

Beyond pose accuracy, 3D model fidelity critically impacts grasp sampling. Grasping deteriorates when candidates are generated on imperfect models due to increased geometric collisions. Smooth surface reconstructions like UniSurf offer excellent collision-free candidates, while faster methods like Instant-NGP suffer more geometric failures. However, under compounded scenarios (imperfect meshes for both steps), pose estimation accuracy acts as the ultimate determinant of success. Robust estimators like FoundationPose compensate for moderate mesh flaws, but even perfect poses cannot save grasps compromised by highly flawed shapes.

Finally, our end-to-end analysis clarifies the codependent relationship between model quality and pose accuracy. A high-quality mesh is the foundation for success, as it is required to generate a rich set of viable grasp candidates and enable an accurate 6D pose estimate. However, our results show that the final accuracy of the object's pose is the more direct determinant of grasping success. A state-of-the-art pose estimator can often compensate for moderate geometric inaccuracies in its reference model. Still, even a perfect pose cannot recover a grasp that was miscalculated on a severely flawed mesh. This highlights that while mesh quality is foundational, pose estimation accuracy is the more dependent metric for successful manipulation.

The primary limitations of our work are its reliance on simulation and its strict focus on two-jaw antipodal grasping. Task diversity does not yet extend to other manipulation strategies such as suction or soft grippers that exhibit distinct sensitivities to geometric artifacts. Future work will focus on validating these findings on a physical robotic platform and extending the framework to manipulation primitives beyond grasping, including precision placement and assembly.

Ultimately, this work advocates for a shift toward holistic perception-to-action benchmarks that evaluate functional manipulation efficacy across the entire pipeline.

ACKNOWLEDGMENT

This work was supported by the European Union's RICAIP (Horizon 2020, grant No. 857306) and AIRISE (Horizon Europe, grant No. 101092312) projects, the Grant Agency of the Czech Technical University in Prague (No. SGS23/172/OHK3/3T/13), and the Czech Science Foundation (GACR) EXPRO grant No. 23-07973X (UNI-3D). The authors acknowledge the use of the Gemini AI tool for text and grammar refinement, and the assistance of Claude and Grok in improving the Python code used to generate the plots.

REFERENCES

[1] Y. Labbé, L. Manuelli, A. Mousavian, S. Tyree, S. Birchfield, J. Tremblay, J. Carpentier, M. Aubry, D. Fox, and J. Sivic, "Megapose: 6d pose estimation of novel objects via render & compare," in *Proc. Conf. on Robot Learning (CoRL)*. PMLR, 2022, pp. 715–725. [Online]. Available: https://proceedings.mlr.press/v205/labbe23a.html

[2] B. Wen, W. Yang, J. Kautz, and S. Birchfield, "Foundationpose: Unified 6d pose estimation and tracking of novel objects," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2024, pp. 17 868–17 879. [Online]. Available: https://doi.org/10.1109/CVPR52733.2024.01692

[3] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," in *Proc. European Conf. on Computer Vision (ECCV)*. Springer, 2020, pp. 405–421. [Online]. Available: https://doi.org/10.1007/978-3-030-58452-8_24

[4] P. Wang, L. Liu, Y. Liu, C. Theobalt, T. Komura, and W. Wang, "Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021, pp. 27 171–27 183. [Online]. Available: https://proceedings.neurips.cc/paper/2021/hash/e41e164f7485ec4a28741a2d0ea41c74-Abstract.html

[5] T. Hodan, F. Michel, E. Brachmann, W. Kehl, A. G. Buch, D. Kraft, B. Drost, J. Vidal, S. Ihrke, X. Zabulis, C. Sahin, F. Manhardt, F. Tombari, T. Kim, J. Matas, and C. Rother, "BOP: benchmark for 6d object pose estimation," in *Proc. European Conf. on Computer Vision (ECCV)*. Springer, 2018, pp. 19–35. [Online]. Available: https://doi.org/10.1007/978-3-030-01249-6_2

[6] H. Fang, C. Wang, M. Gou, and C. Lu, "Graspnet-1billion: A large-scale benchmark for general object grasping," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2020, pp. 11 441–11 450. [Online]. Available: https://doi.org/10.1109/CVPR42600.2020.01146

[7] S. Hinterstoisser, V. Lepetit, S. Ilic, S. Holzer, G. R. Bradski, K. Konolige, and N. Navab, "Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes," in *Proc. Asian Conf. on Computer Vision (ACCV)*. Springer, 2012, pp. 548–562. [Online]. Available: https://doi.org/10.1007/978-3-642-37331-2_42

[8] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes," in *Proc. Robotics: Science and Systems (RSS)*, 2018. [Online]. Available: http://www.roboticsproceedings.org/rss14/p19.html

[9] C. Wang, D. Xu, Y. Zhu, R. M. Martin, C. Lu, L. Fei-Fei, and S. Savarese, "Densefusion: 6d object pose estimation by iterative dense fusion," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2019, pp. 3343–3352. [Online]. Available: https://doi.org/10.1109/CVPR.2019.00346

[10] T. Müller, A. Evans, C. Schied, and A. Keller, "Instant neural graphics primitives with a multiresolution hash encoding," *ACM Trans. Graph.*, vol. 41, no. 4, pp. 102:1–102:15, 2022. [Online]. Available: https://doi.org/10.1145/3528223.3530127

[11] L. Yariv, J. Gu, Y. Kasten, and Y. Lipman, "Volume rendering of neural implicit surfaces," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021, pp. 4805–4815. [Online]. Available: https://proceedings.neurips.cc/paper/2021/hash/25e2a30f44898b9f3e978b1786dcd85c-Abstract.html

[12] L. Yariv, P. Hedman, C. Reiser, D. Verbin, P. P. Srinivasan, R. Szeliski, J. T. Barron, and B. Mildenhall, "Bakedsdf: Meshing neural sdfs for real-time view synthesis," in *Proc. ACM SIGGRAPH*. ACM, 2023, pp. 46:1–46:9. [Online]. Available: https://doi.org/10.1145/3588432.3591536

[13] T. Schöps, J. L. Schönberger, S. Galliani, T. Sattler, K. Schindler, M. Pollefeys, and A. Geiger, "A multi-view stereo benchmark with high-resolution images and multi-camera videos," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 2538–2547. [Online]. Available: https://doi.org/10.1109/CVPR.2017.272

[14] A. Knapitsch, J. Park, Q. Zhou, and V. Koltun, "Tanks and temples: benchmarking large-scale scene reconstruction," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 78:1–78:13, 2017. [Online]. Available: https://doi.org/10.1145/3072959.3073599

[15] V. Burde, A. Benbihi, P. Burget, and T. Sattler, "Comparative evaluation of 3d reconstruction methods for object pose estimation," in *Proc. IEEE/CVF Winter Conf. on Applications of Computer Vision (WACV)*. IEEE, 2025, pp. 7669–7681. [Online]. Available: https://doi.org/10.1109/WACV61041.2025.00745

[16] A. Bicchi and V. Kumar, "Robotic grasping and contact: A review," in *Proc. IEEE Int. Conf. on Robotics and Automation (ICRA)*. IEEE, 2000, pp. 348–353. [Online]. Available: https://doi.org/10.1109/ROBOT.2000.844081

[17] L. Pinto and A. Gupta, "Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours," in *Proc. IEEE Int. Conf. on Robotics and Automation (ICRA)*. IEEE, 2016, pp. 3406–3413. [Online]. Available: https://doi.org/10.1109/ICRA.2016.7487517

[18] J. Mahler, J. Liang, S. Niyaz, M. Laskey, R. Doan, X. Liu, J. A. Ojea, and K. Goldberg, "Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics," in *Proc. Robotics: Science and Systems (RSS)*, 2017. [Online]. Available: http://www.roboticsproceedings.org/rss13/p58.html

[19] B. Çalli, A. Singh, A. Walsman, S. S. Srinivasa, P. Abbeel, and A. M. Dollar, "The YCB object and model set: Towards common benchmarks for manipulation research," in *Proc. Int. Conf. on Advanced Robotics (ICAR)*. IEEE, 2015, pp. 510–517. [Online]. Available: https://doi.org/10.1109/ICAR.2015.7251504

[20] E. Coumans and Y. Bai, "Pybullet, a python module for physics simulation for games, robotics and machine learning," 2016.

[21] M. Rudorfer, "Burg toolkit, a python module for benchmarking and understanding robotic grasping," https://github.com/mrudorfer/burg-toolkit, 2022.

[22] M. Tancik, E. Weber, E. Ng, R. Li, B. Yi, T. Wang, A. Kristoffersen, J. Austin, K. Salahi, A. Ahuja, D. McAllister, J. Kerr, and A. Kanazawa, "Nerfstudio: A modular framework for neural radiance field development," in *Proc. ACM SIGGRAPH*. ACM, 2023, pp. 72:1–72:12. [Online]. Available: https://doi.org/10.1145/3588432.3591516

[23] Z. Li, T. Müller, A. Evans, R. H. Taylor, M. Unberath, M. Liu, and C. Lin, "Neuralangelo: High-fidelity neural surface reconstruction," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2023, pp. 8456–8465. [Online]. Available: https://doi.org/10.1109/CVPR52729.2023.00817

[24] M. Oechsle, S. Peng, and A. Geiger, "UNISURF: unifying neural implicit surfaces and radiance fields for multi-view reconstruction," in *Proc. IEEE/CVF Int. Conf. on Computer Vision (ICCV)*. IEEE, 2021, pp. 5569–5579. [Online]. Available: https://doi.org/10.1109/ICCV48922.2021.00554

[25] Z. Yu, S. Peng, M. Niemeyer, T. Sattler, and A. Geiger, "Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. [Online]. Available: http://papers.nips.cc/paper_files/paper/2022/hash/9f0b1220028dfa2ee82ca0a0e0fc52d1-Abstract-Conference.html

[26] RealityCapture, "RealityCapture," https://www.capturingreality.com/, 2023.