Physically Stable Dual-Arm Grasp Generation

A thesis submitted in partial fulfillment
of the requirements for the degree of

*Master of Science in **Computer Science and Engineering** by Research*

by

Md Faizal Karim
2022121004
md.faizal@research.iiit.ac.in

International Institute of Information Technology Hyderabad
Hyderabad — 500032
India

April 2026

International Institute of Information Technology Hyderabad
Hyderabad, India

CERTIFICATE

This is to certify that the work presented in this thesis proposal titled "**Physically Stable Dual-Arm Grasp Generation**" by Md Faizal Karim has been carried out under my supervision and is not submitted elsewhere for a degree.

_____ _____

Date Advisor: Dr. K Madhava Krishna

In the sum of parts, there are only parts.

Acknowledgments

I would like to begin by expressing my sincere gratitude to IIIT Hyderabad for providing an environment that fostered curiosity, resilience, and growth. The past four years have been a journey of learning as well as challenges, shaping both my academic and personal outlook in ways I will always value.

Firstly, I am deeply grateful to my advisor, Prof. Madhava Krishna, for his guidance, support, and constant encouragement. It was he who introduced me to the research project I will speak about in this thesis — the project that has changed my life in ways I still find difficult to fully comprehend. His vision and clarity of thought are things I will carry with me for a long time. I would also like to thank Prof. Nagamanikandan Govindan and Prof. Arun Kumar Singh, who have always been there with advice and direction, helping our team navigate the research landscape.

I am thankful to my co-authors and collaborators – Gaurav, Vignesh, Shreya, and Saad, for their support, patience, and companionship throughout this work. Working together through both successes and setbacks made the process not only productive, but deeply meaningful. Their presence made the difficult phases easier and the good moments more rewarding. I wish all of you the very best for what lies ahead.

Finally, I would like to thank my family for their unwavering support and belief in me. For the past four years, I have stayed away from home, often far removed from the comfort and familiarity it brings. Yet, their support was always constant and always present. The world of research was unfamiliar to them, but they never once held me back from pursuing what I cared about. Their trust, strength, and encouragement have been the ground beneath all of this and whatever lies ahead will stand because of them.

For all of this, I remain deeply grateful.

Abstract

Robotic manipulation of large and complex objects requires not only the ability to generate feasible grasps, but also to ensure stability under multi-contact interactions, particularly in dual-arm settings. While recent advances in data-driven grasp synthesis have shown promising results for single-arm grasping, extending these methods to dual-arm scenarios introduces additional challenges such as coordinated contact placement, force balance, and collision avoidance. This thesis addresses these challenges through a unified progression from generative modeling to analytical understanding, which are subsequently combined into an end-to-end framework for dual-arm grasp generation.

We begin with Constrained Grasp Diffusion Fields (CGDF), a diffusion-based generative model for 6-DoF grasp synthesis on complex objects. CGDF introduces a part-guided diffusion mechanism that enables sample-efficient generation of grasps constrained to specific regions of an object, allowing dense and targeted grasp proposals even on large and geometrically intricate shapes. This provides a strong foundation for generating diverse and region-aware grasp candidates.

Building on this, we recognize that dual-arm grasping requires a principled notion of stability beyond independent single-arm grasps. To address this, we developed DG16M, a large-scale dataset and an analytical framework for evaluating dual-arm grasp stability using an optimization-based force-closure formulation. By explicitly modeling contact forces, friction constraints, and external disturbances, this work provides a physically grounded metric for assessing grasp quality and enables the learning of stability-aware models.

Finally, we integrate these two perspectives — generative modeling and analytical evaluation into DAGDiff, an end-to-end diffusion-based framework for directly generating dual-arm grasp pairs. Unlike prior approaches that rely on heuristic pairing or region selection, DAGDiff formulates grasp generation in the joint $SE(3) \times SE(3)$ space and incorporates guidance from stability and collision-aware classifiers during the diffusion process. This allows the model to generate dual-arm grasps that are not only diverse, but also force-closure stable and collision-free directly from raw input point clouds.

Together, these contributions establish a principled pipeline for dual-arm grasp synthesis: from region-aware generative modeling, to analytical stability evaluation, to fully integrated end-to-end learning. This work advances the state of robotic manipulation by enabling reliable, generalizable, and physically grounded dual-arm grasping on complex real-world objects.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

It is comparatively easy to make computers exhibit adult level performance on intelligence tests or playing checkers, and difficult or impossible to give them the skills of a one-year-old when it comes to perception and mobility.

HANS MORAVEC in *Mind Children*, 1988

Robotic manipulation is the mechanism through which autonomous systems physically act on the world, enabling tasks ranging from assembly and logistics to household assistance and service robotics. Recent advances in humanoid robotics further highlight this vision, where robots are expected to operate in unstructured household environments and perform everyday tasks alongside humans (as shown in Figure 1.1). At the core of this capability lies *grasping* — before a robot can move, reorient, or coordinate an object, it *must* first establish a stable and reliable contact configuration. As emphasized in classical grasp analysis [1–3], a grasp must be able to counteract external disturbances through appropriate contact forces, which makes grasp quality a fundamental prerequisite for manipulation rather than a separate subproblem.

This requirement becomes even more pronounced when the object is large, heavy, or geometrically complex. In such cases, single-arm grasping is often insufficient, and dual-arm manipulation becomes the natural and necessary alternative [4–6]. Dual-arm systems can distribute forces across multiple contacts and support a wider range of object sizes and shapes, but they also introduce a more difficult stability problem: the two grasps must work together as a physically consistent pair, not merely as two individually feasible contacts. This makes dual-arm grasping a problem that sits at the intersection of grasp generation, force balance, and coordinated manipulation [7–9].

In this thesis, we study the problem of generating physically stable dual-arm grasps directly from object geometry. Our goal is to develop methods that can not only propose feasible grasp configurations, but also ensure that these grasps are stable, collision-free, and applicable to complex and large real-world objects. We approach this problem by combining *generative modeling* techniques with physically grounded *stability analysis*, enabling the synthesis of grasp pairs that are both diverse and physically stable.

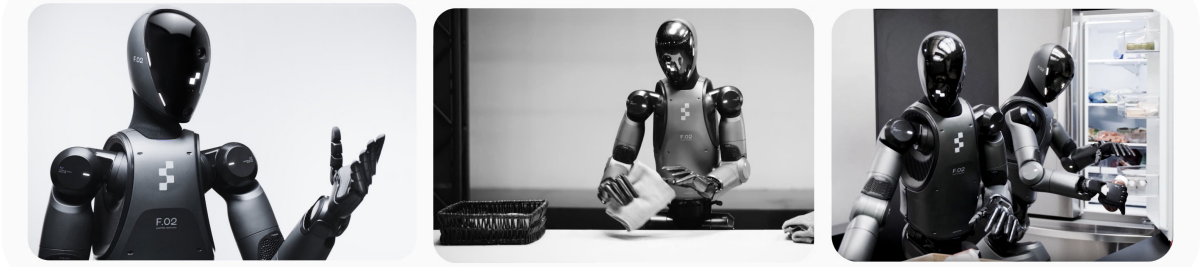

Figure 1.1: Humanoids by *Figure*, doing household work.

1.1 Motivation and Challenges

Consider lifting a large object such as a monitor or carrying a box with both hands. The placement of the hands is not arbitrary, as they must be positioned such that the object remains balanced, does not slip, and does not rotate under its own weight. This intuitive action reflects a fundamental principle in robotic manipulation — before an object can be manipulated, it must first be grasped in a physically stable manner [10]. Classical works in robotic manipulation have formalized this idea through concepts such as *force closure*, which ensures that a set of contact forces can resist arbitrary external disturbances [1].

Grasping, therefore, is not merely a preliminary step but a prerequisite for manipulation. As highlighted in modern surveys, it remains one of the most fundamental and challenging problems in robotics, requiring the integration of perception, planning, and control [5, 6, 11, 12]. While significant progress has been made in generating grasps for single-arm systems, these methods often rely on local geometric reasoning and do not fully capture the complexities of physical interaction.

These challenges become significantly more pronounced in dual-arm manipulation. Unlike single-arm grasping, where stability can often be inferred from local contact geometry, dual-arm grasping requires *coordinated interaction* between multiple contact points. The two grippers must act together as a single system, thereby ensuring consistent force distribution, torque balance, and compatibility with the object's geometry. Hence, such systems introduce additional layers of complexity, including inter-arm coordination, shared control, and coupled physical constraints [4, 6, 13].

A key difficulty arises from the fact that dual-arm grasping cannot be reduced to selecting two independent single-arm grasps. Even if each grasp is individually feasible, their combination may fail to stabilize the object due to poor force balance, unfavorable placement relative to the center of mass, or conflicting contact interactions. This highlights a fundamental gap between geometric feasibility and physical stability, which has long been recognized in grasp analysis literature [1–3, 11]. Furthermore, existing learning-based approaches primarily focus on generating diverse grasp candidates, often treating stability as a post-hoc evaluation step [14–18]. This separation between generation and evaluation leads to inefficiencies and limits the ability to directly produce physically consistent grasp pairs. Bridging this gap requires methods that can jointly reason about geometry and physics dur-

ing the grasp generation process. In this thesis, we address these challenges by developing methods that integrate generative modeling with physically grounded stability analysis, enabling the direct synthesis of stable dual-arm grasps on complex objects.

1.2 Problem Statement

In this thesis, we address the problem of generating dual-arm grasps directly from object geometry (such as pointcloud or RGB-D images). Given an appropriate representation of an object, the goal is to determine a pair of grasp configurations for two robotic arms that can jointly manipulate the object in a reliable and physically consistent manner.

A valid dual-arm grasp must satisfy several key requirements:

- **Physical Stability:** The grasp pair should be able to support the weight of the object while resisting external disturbances and small perturbations. This requires coordinated force and torque distribution between the two grippers to prevent slipping or undesired motion.
- **Collision-Free Configuration:** The grasps must not result in collisions between the grippers and the object or between the two grippers themselves, ensuring that the configuration is physically executable.
- **Geometric Feasibility and Consistency:** Each grasp should align with the local geometry of the object, allowing appropriate contact without penetration or slippage. Additionally, the two grasps must be mutually compatible, forming a coordinated pair that can jointly support and manipulate the object rather than acting as independent contacts.

The objective is to generate such grasp pairs in a joint and scalable manner, without relying on heuristic pairing strategies or post-hoc filtering. This requires jointly reasoning about object geometry and physical constraints within a unified framework.

1.3 Thesis Contributions

The overarching objective of this thesis is to develop a principled framework for dual-arm grasping that combines *grasp generation* with *physical stability assessment*. Rather than treating these as separate problems, we progressively build methods that can generate grasps on large and complex objects, evaluate their stability in a physically grounded manner, and ultimately synthesize dual-arm grasp pairs directly from object pointcloud. The main contributions of this thesis are as follows:

1. **CGDF: Constrained Grasp Diffusion Fields** [19]. We introduce a diffusion-based framework for generating dense 6-DoF grasps on large and geometrically complex objects. CGDF incorporates a part-guided strategy that enables sample-efficient grasp generation on constrained

regions, making it suitable for multi-region and dual-arm manipulation settings. This work establishes the generative foundation of the thesis.

2. **DG16M: A large-scale dataset and analytical framework for dual-arm grasp stability** [20]. We develop an optimization-based force-closure formulation to evaluate dual-arm grasps under friction and actuator constraints, and use it to construct DG16M, a large-scale dataset of physically stable grasp pairs. In addition, we build a simulation-verified benchmark using Isaac Gym to provide more realistic stability labels for dual-arm grasping. This work contributes the stability-centric component of the thesis.

3. **DAGDiff: Dual-Arm Grasp Diffusion** [21]. We present an end-to-end diffusion-based framework that directly generates stable and collision-free dual-arm grasp pairs from object point clouds. DAGDiff models the joint pose space $SE(3) \times SE(3)$ and uses guidance signals derived from stability and collision-aware predictors to steer the diffusion process toward physically feasible grasp pairs. This work unifies grasp generation and stability reasoning in a single framework.

Together, these contributions provide a coherent progression from region-aware grasp generation, to analytical stability evaluation, to unified dual-arm grasp synthesis, advancing the state of physically grounded robotic manipulation.

1.4 Thesis Structure

This thesis is organized to progressively develop the problem of dual-arm grasping from background concepts to generation, stability analysis, and finally a unified framework. *Chapter 2* provides the necessary background on robotic grasping, including representations, grasp generation methods, and stability analysis techniques. *Chapter 3* introduces CGDF, a diffusion-based approach for generating dense grasps on large and complex objects. *Chapter 4* presents DG16M, where we study dual-arm grasp stability through an optimization-based force-closure formulation and construct a large-scale dataset with physically grounded annotations and simulation-based validation. Building on these insights, *Chapter 5* introduces DAGDiff, an end-to-end framework that directly generates stable and collision-free dual-arm grasp pairs by modeling their joint pose space with stability-aware guidance. Finally, *Chapter 6* concludes the thesis with a summary of contributions and directions for future work.

Chapter 2

Background

All models are wrong, but some are useful.

GEORGE BOX in *Science and Statistics*, 1976

2.1 Grasp Representation and Formulation

A *robotic grasp* refers to the configuration and interaction between a robot end-effector and an object that enables stable manipulation. Formally, grasping involves both the geometric relationship between the robot and the object, and the physical constraints governing contact and force transmission. Importantly, while a grasp inherently captures the interaction between the end-effector and the object, the manner in which this interaction is *represented* can vary significantly depending on the gripper embodiment, task requirements, and environmental constraints.

In practice, a grasp can be parameterized in multiple ways, each emphasizing different aspects of the interaction. For instance, it may be represented purely as a pose of the end-effector in $SE(3)$, as a set of contact points and associated surface normals, or as a higher-dimensional configuration encoding gripper geometry, joint states, and contact forces. These representations are often interchangeable under ideal conditions, but differ in their expressiveness, computational efficiency, and suitability for learning-based methods.

Consequently, the choice of grasp representation plays a critical role in both modeling and generation, as it determines how grasp candidates are sampled, evaluated, and optimized. In the following subsections, we discuss the most commonly used grasp representations and their corresponding formulations.

2.1.1 Pose-Based Grasp Representation

A widely used representation models a grasp as a rigid transformation in the special Euclidean group $SE(3)$, as shown in Figure 2.1(a). A pose-based grasp configuration H is defined as a homo-

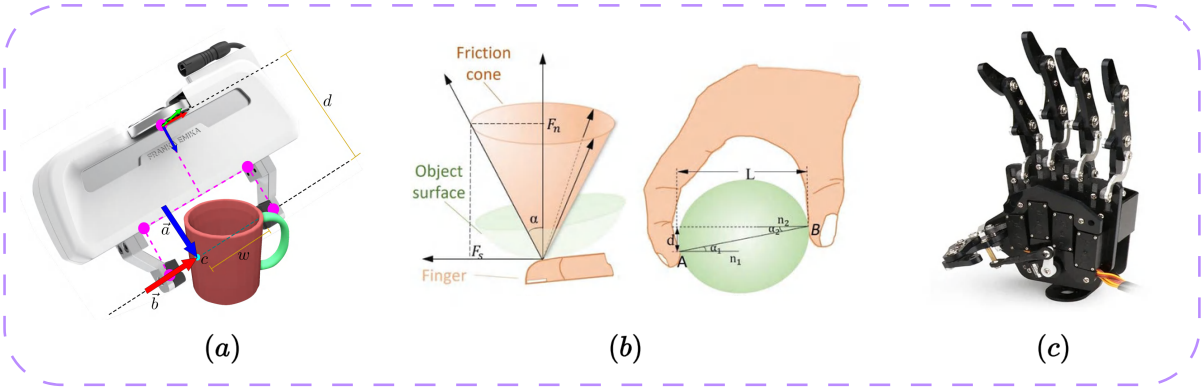

Figure 2.1: Overview of grasp representations at different levels of abstraction, including geometric grasp poses, contact-based friction models, and dexterous multi-fingered grasping.

geneous transformation matrix:

$$H = \begin{bmatrix} R & t \\ 0^T & 1 \end{bmatrix} \in SE(3)$$

where $R \in SO(3)$ represents the rotation matrix describing the orientation of the end-effector, and $t \in \mathbb{R}^3$ denotes the translation vector corresponding to its position in the world frame.

This 6-DoF representation specifies the global pose of the gripper and is commonly used in modern grasp generation methods, particularly in learning-based approaches that operate on point clouds or RGB-D observations [14–18, 22–25]. Pose-based representations are computationally efficient and well-suited for sampling-based and generative methods, including diffusion-based approaches [19, 26, 27]. However, they abstract away the underlying contact interactions, and therefore do not explicitly encode physical feasibility or force transmission properties.

2.1.2 Contact-Based Grasp Representation

To model physical interaction more explicitly, grasps are also represented in terms of contact points between the gripper and the object surface. A grasp can be defined as a set of N contact points:

$$\mathcal{C} = \{(p_i, \hat{n}_i)\}_{i=1}^{N}$$

where $p_i \in \mathbb{R}^3$ denotes the contact location and $\hat{n}_i \in \mathbb{R}^3$ denotes the corresponding surface normal as shown in Figure 2.1(b).

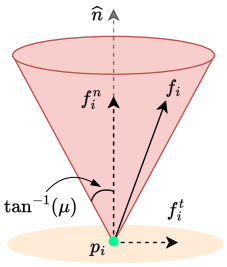

Figure 2.2: Illustration of Friction cone at contact point p_i.

Each contact is associated with a local contact frame and a contact model, such as Point Contact With Friction (PCWF) or Soft-Finger Contact models [10, 11]. The contact forces at each point can be

decomposed into normal (f_i^n) and tangential components (f_i^t) as:

$$f_i = f_i^n \hat{n}_i + f_i^t,$$
$$\|f_i^t\| \leq \mu f_i^n$$

where μ is the friction coefficient of the object. An illustration of this is shown in Figure 2.2.

The effect of these forces on the object is captured through the grasp matrix $G \in \mathbb{R}^{6\times 3}$, which maps contact forces to object wrenches $w = G \cdot f$, where $w \in \mathbb{R}^6$ represents the wrench (force and torque), and $f \in \mathbb{R}^3$ is the contact force vector. This formulation enables analytical evaluation of grasp stability using concepts such as force closure and wrench space analysis. While contact-based representations provide a physically grounded framework, they rely on accurate knowledge of contact locations and surface normals, which are not directly observable in real-world perception and must be estimated from noisy point clouds (or RGB-D data). This makes such formulations difficult to apply reliably and integrate into data-driven generative models.

2.1.3 Higher-Order and Configuration-Space Representations

Beyond pose and contact-based formulations, grasping can also be represented in higher dimensional configuration spaces that explicitly model the robotic gripper kinematics and multi-contact interactions, as shown in Figure 2.1(c). In this setting, a grasp is defined in the joint space of the robotic gripper (or typically robotic hand):

$$q = [q_0, q_1, .. \, q_i, .., q_d] \quad \in \mathbb{R}^d$$

where q_i denotes each joint angle and d is the number of degrees of freedom of the gripper. The forward kinematics f_{kin} map the joint configuration to the end-effector pose as $H = f_{\text{kin}}(q)$, while the contact configuration depends on both q and the object geometry.

Such representations are particularly relevant in dexterous manipulation and dual-arm grasping, where interactions involve multiple coordinated contacts [28–30]. In these settings, a valid grasp must satisfy both geometric constraints (e.g., reachability and collision avoidance) and physical constraints such as force balance and actuator limits.

Notably, the mapping from joint configurations to end-effector poses or contact configurations is, in general, *many-to-one*, i.e., multiple distinct configurations may yield equivalent grasp outcomes. While this redundancy provides additional degrees of freedom that can be exploited to satisfy secondary objectives, it also increases the dimensionality of the search space, thereby complicating the associated planning and optimization problems.

Summary. Overall, pose-based, contact-based, and configuration-space representations offer complementary perspectives on grasping, trading off between computational efficiency, physical fidelity, and representational expressiveness. These distinctions are central to understanding the design of

grasp generation methods and their applicability to complex manipulation scenarios such as dual-arm grasping.

2.2 Grasp Generation Methods

Grasp generation methods aim to predict feasible grasp configurations given object geometry or sensory observations. Broadly, these methods can be categorized into classical analytical approaches and data-driven learning-based approaches.

2.2.1 Classical Methods

Classical grasp generation methods rely on geometric and physical principles to synthesize grasps. A common strategy is to sample candidate grasps using heuristics such as antipodal contact conditions, where opposing contact points satisfy friction constraints [10, 11]. These sampled grasps are then evaluated using analytical metrics derived from Grasp Wrench Space (GWS) analysis, such as Force Closure [1] or the Ferrari-Canny metric [3].

Such methods typically follow a two-stage pipeline: (i) sampling candidate grasps based on object geometry, and (ii) ranking them using stability metrics. Systems such as Dex-Net [22] leverage large-scale simulation to generate grasp candidates and evaluate them using analytic grasp quality measures, enabling robust grasp selection under uncertainty. Simulation frameworks such as GraspIt! [31] further enable physics-based evaluation of grasp stability by modeling contact interactions, friction, and force closure, and have been widely used as tools for analyzing and validating grasp configurations.

While classical methods provide strong physical guarantees, they often suffer from computational inefficiency due to exhaustive sampling on object meshes and limited generalization to complex, real-world geometries.

2.2.2 Learning-Based Methods

Learning-based approaches directly infer grasp configurations from sensory inputs such as single or multi-view RGB-D images or point clouds, leveraging large datasets and neural networks.

2.2.2.1 Predictive Models

Predictive methods formulate grasp generation as a supervised learning problem, where the model predicts grasp parameters or scores candidate grasps. Early works such as [18] and [32] used convolutional neural networks (CNNs) to predict grasp rectangles from images. Subsequent approaches extended this to 6-DoF grasp prediction using depth and point cloud representations, such as [14, 15, 22, 25], which predict dense grasp poses directly from scene observations.

These methods are efficient at inference (as it requires a single forward-pass) and can generalize across object categories. However, they are often limited by dataset bias and typically rely on predefined grasp representations which restrict their ability to capture the multimodal nature of grasp distributions.

2.2.2.2 Generative Models

Generative approaches explicitly model the multimodal distribution over feasible grasps and aim to sample diverse grasp candidates (as illustrated in Figure 2.3). Variational Autoencoders (VAEs) [33] and Conditional VAEs [34] have been used to learn latent representations of grasp distributions, as in 6DoF-GraspNet [23], where grasp poses are sampled and refined using learned evaluators. More recently, Diffusion Models [35] have emerged as a powerful framework for grasp generation.

These methods model grasp synthesis as an iterative denoising process over pose space, enabling the generation of diverse and high-quality grasps. Works such as [26–28] demonstrate that diffusion models can effectively capture the multimodal nature of grasp distributions and improve grasp success rates in both simulation and real-world settings. Additionally, diffusion-based methods can incorporate guidance signals, such as collision avoidance or task constraints, during sampling to improve physical feasibility. Compared to predictive models, generative approaches offer greater diversity and flexibility, but often require additional mechanisms for ensuring physical validity and stability.

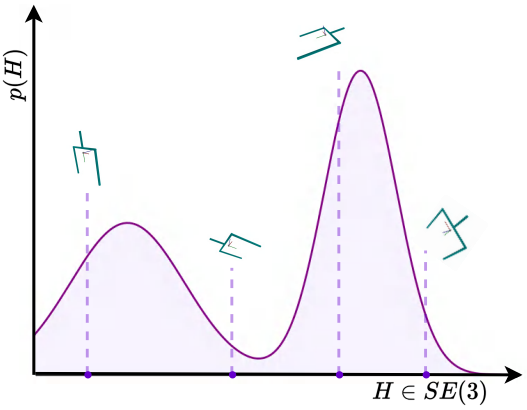

Figure 2.3: Illustration of a multimodal grasp distribution over $SE(3)$, highlighting diverse feasible grasp configurations.

Overall, classical and learning-based approaches offer complementary strengths: analytical methods provide principled guarantees of stability, while learning-based methods enable scalable and generalizable grasp generation. Bridging these paradigms remains a key challenge, particularly in complex settings such as dual-arm grasping.

2.3 Grasp Evaluation Metrics

Evaluating the quality of a robotic grasp is a fundamental component of grasp planning and synthesis. Given that multiple feasible grasp configurations may exist for a single object, grasp evaluation metrics provide a principled way to quantify stability, robustness, and task suitability. These metrics are broadly categorized into analytical, geometric, and empirical approaches, each capturing different aspects of grasp quality [36].

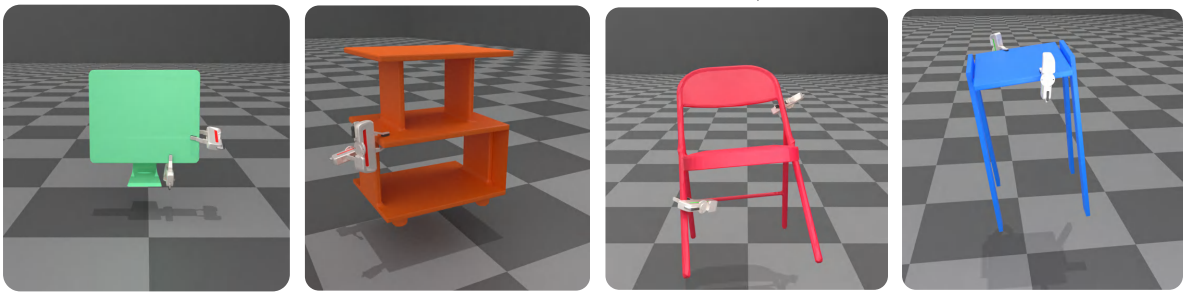

Figure 2.4: Example of grasp evaluation in Isaac Gym [37] simulation environment. The predicted grasps are executed using parallel-jaw grippers on diverse objects to assess stability and success under physics-based interactions.

2.3.1 Analytical Metrics

Analytical metrics are derived from physical models of contact and force transmission. These metrics typically rely on the concept of the *Grasp Wrench Space (GWS)*, which represents the set of all wrenches (forces and torques) that can be applied to an object through contact forces.

2.3.1.1 Force Closure

A grasp is said to achieve *force closure* if it can resist arbitrary external disturbances through appropriate contact forces [10]. Let $G \in \mathbb{R}^{6\times m}$ denote the grasp matrix mapping contact forces to object wrenches, and let $\mathcal{F}$ represent the set of admissible contact forces (e.g., constrained by friction cones). The set of achievable wrenches is given by:

$$\mathcal{W} = \{w \in \mathbb{R}^6 \mid w = Gf,\ f \in \mathcal{F}\} \tag{2.1}$$

The grasp achieves force closure if:

$$0 \in \text{int}(\mathcal{W}) \tag{2.2}$$

i.e., the *origin* lies in the interior of the convex hull of achievable wrenches. Force closure serves as a binary but widely used criterion for grasp stability and is often employed as a necessary condition for a valid grasp.

2.3.1.2 Ferrari – Canny Metric

To obtain a continuous measure of grasp quality, Ferrari and Canny [3] proposed a metric based on the largest disturbance wrench that a grasp can resist in any direction. Let $G \in \mathbb{R}^{6\times m}$ denote the grasp matrix mapping contact forces to object wrenches, $f \in \mathbb{R}^m$ the stacked vector of admissible contact forces, and $\mathcal{F}$ the set of feasible contact forces defined by friction constraints. The set of

achievable wrenches is given by $\mathcal{W} = \{Gf \mid f \in \mathcal{F}\}$. The Ferrari–Canny metric is defined as:

$$\epsilon = \min_{\|w\|=1} \max_{f \in \mathcal{F}} w^\top G f \tag{2.3}$$

where $w \in \mathbb{R}^6$ represents a unit disturbance wrench. This corresponds to the radius of the largest ball centered at the origin that lies within $\mathcal{W}$. Equivalently, when $\mathcal{W}$ is represented as a convex polytope, the metric can be expressed as:

$$\epsilon = \min_i \; d(0, \partial \mathcal{W}_i) \tag{2.4}$$

where $\partial \mathcal{W}_i$ denotes the i-th facet of the wrench space boundary, and $d(0, \partial \mathcal{W}_i)$ is the distance from the origin to that facet. A larger value of ϵ indicates greater robustness which corresponds to grasps that can resist larger external disturbances.

2.3.2 Empirical and Simulation-Based Metrics

Empirical metrics evaluate grasp quality through simulation [37–41] (an example is shown in Figure 2.4) or real-world execution. These include measures such as grasp success rate, number of contact points, robustness to perturbations, and stability under dynamic interactions [17, 22, 30, 42]. For instance, a grasp can be considered successful if the object remains stable after applying gravity or external disturbances, or if it can be reliably lifted and manipulated without slippage.

Simulation-based metrics are widely used in modern grasping pipelines as they implicitly capture complex contact dynamics, frictional effects, and uncertainties that are difficult to model analytically. They also enable large-scale data generation and evaluation across diverse object geometries and task conditions. However, they are computationally expensive, sensitive to simulator fidelity (e.g., contact models and friction parameters), and may suffer from sim-to-real discrepancies when deployed on physical systems.

Summary. Each class of metrics captures different aspects of grasp quality. Analytical metrics provide strong theoretical guarantees but may be conservative and computationally intensive while Empirical metrics offer realistic evaluation but depend on simulation accuracy and computational resources. In practice, many grasp generation systems combine multiple metrics to balance efficiency and robustness, using analytical measures for initial filtering and empirical evaluation for final validation.

Chapter 3

Learning to Grasp Large and Complex objects

The hand is the visible part of the brain.

IMMANUEL KANT

3.1 Introduction

Robotic grasping is a fundamental capability that underpins a wide range of manipulation tasks, from industrial automation to assistive robotics. Significant progress has been made in learning-based grasp generation, where models predict feasible grasp configurations directly from sensory inputs such as point clouds or depth images [14, 15, 17, 22, 23, 25]. These methods have demonstrated strong performance in structured environments involving tabletop objects and relatively simple geometries.

However, real-world manipulation often involves objects that are large, geometrically complex, and structurally diverse, such as furniture, tools, and articulated objects [43–45]. In such scenarios, grasp generation becomes significantly more challenging. Large objects require reasoning over extended spatial extents, while complex geometries demand a detailed understanding of local shape features. Existing methods, which typically rely on global shape representations or uniform grasp sampling strategies, often fail to provide dense and reliable grasp coverage on such objects. As a result, they become sample-inefficient and require generating a large number of candidates to obtain a few valid grasps.

Another important limitation of prior work is the assumption of uniform grasp distributions across the object surface. In practice, many objects require the grasp to be concentrated on specific regions due to physical constraints, functional affordances, or task requirements [46–49]. For instance, grasping a chair may require placing contacts on its arms, while manipulating a tool may require grasping its handle. Uniform grasp generation fails to capture such structured constraints, limiting its applicability in realistic manipulation settings. Recent advances in generative modeling, particularly diffusion-based approaches, have shown promising results in capturing multimodal distributions over grasp poses [26, 27]. These methods improve diversity and coverage, but still struggle to model fine-grained local geometry and generate grasps that are concentrated on specific regions of

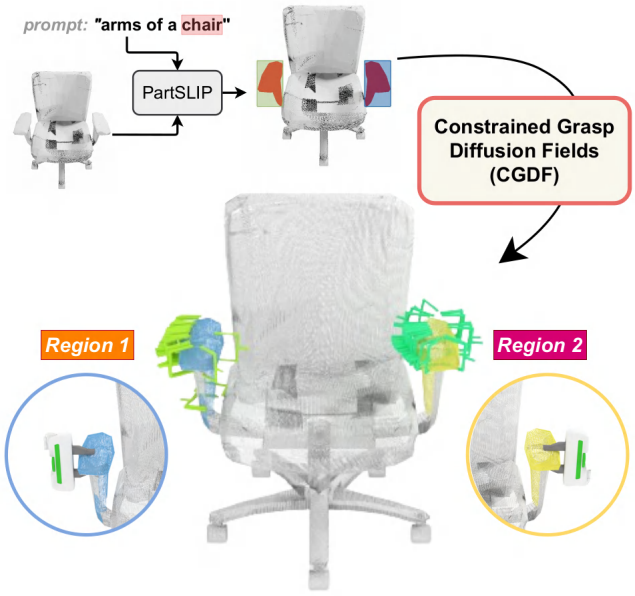

Figure 3.1: CGDF: Constrained Grasp Diffusion Fields produces dense grasp configurations on large, geometrically complex objects (e.g., chairs) in a dual-arm setting. Given target regions (optionally obtained via text prompts using PartSLIP [50]), CGDF employs a part-guided diffusion strategy to efficiently generate grasps localized to these regions, enabling coordinated multi-region and multi-arm grasping.

large objects. This highlights the need for representations and generation strategies that can simultaneously capture global structure and local geometric detail.

In this chapter, we address the problem of learning to generate grasps on large and complex objects. We focus on the setting where grasp generation must be both dense and region-aware, enabling efficient sampling of grasps on specified parts of an object. To this end, we propose **CGDF: Constrained Grasp Diffusion Fields**, a diffusion-based framework that leverages improved shape representations and part-guided generation to produce sample-efficient and geometrically consistent grasp distributions (as shown in Figure 3.1). Our approach enables dense grasp generation on complex objects and naturally extends to multi-arm manipulation settings, where multiple regions of interest must be considered simultaneously. The project page and code can be accessed here: https://constrained-grasp-diffusion.github.io/.

3.2 Related Works

3.2.1 Grasp Generation

Grasp generation for single-arm manipulation, particularly for small objects, has been extensively studied in prior work [14, 15, 17, 22, 23]. Most existing methods learn a mapping from object observa-

tions, such as point clouds or images, to a set of feasible grasp poses. These approaches often include an additional evaluation stage to filter out collisions or unstable grasps.

Several works explore alternative formulations for grasp generation. For instance, differentiable sampling-based methods optimize grasp configurations using task-specific objectives [51, 52], while energy-based models learn implicit representations of grasp quality [53]. Other approaches leverage keypoint-based representations to predict grasp poses directly from visual inputs [54–57]. Related task-oriented grasping methods also use keypoints to bridge point clouds and action representations, and some recent VLM-based approaches inject semantic knowledge or affordance cues to guide grasp selection and improve open-vocabulary grasping [58–64]. Despite these advances, the majority of these methods still generate grasps uniformly across the object or scene, which limits their effectiveness in scenarios where grasps need to be localized to specific regions.

While some works extend grasping to dual-arm settings, they often rely on multi-stage pipelines that separately predict grasp points and semantic regions. Such approaches do not explicitly model dense grasp distributions and are not well-suited for generating coordinated grasps across multiple regions of large objects.

3.2.2 Constrained Generative Grasp Sampling

Constrained grasp generation has been explored in both single-arm and dual-arm settings [23, 44, 65]. Many of these methods focus on task-oriented grasping, where grasps are conditioned on specific manipulation objectives. However, such approaches are often limited by the tasks they are trained on and do not generalize well to arbitrary regions.

A more general formulation of constrained grasping is introduced in VCGS [66], which models grasp generation as a conditional generative process [33] over specified target regions. VCGS develops and uses the CONG dataset, which augments large-scale grasp datasets with region annotations and corresponding grasps. While this approach enables dense grasp generation on selected regions, it requires explicitly labeled constrained datasets for training.

In contrast, our work extends this setting to large objects and dual-arm manipulation without relying on conditionally labeled data. By leveraging richer shape representations and a part-guided diffusion strategy, our method is able to generate constrained grasps in a sample-efficient manner while maintaining generalization to complex object geometries.

3.3 Preliminaries

In this section, we briefly review the key representations and ideas that form the foundation of our approach, including implicit shape representations, neural descriptor fields, and diffusion-based grasp generation.

3.3.1 Implicit Shape Representations

Implicit representations model a 3D shape as a continuous function defined over space. A commonly used formulation is the signed distance function (SDF), which maps a 3D point $x \in \mathbb{R}^3$ to its distance from the object surface:

$$\text{SDF}(x) = f_\theta(x),$$

where f_θ is a neural network parameterized by θ. The object surface is implicitly defined as the zero level-set of this function: $\{x \in \mathbb{R}^3 \mid f_\theta(x) = 0\}$.

Convolutional Occupancy Networks extend this idea by encoding shape information into structured feature planes [67]. Given an input point cloud P, per-point features are first extracted using a point cloud encoder (e.g., PointNet [68] or PointTransformer [69]). These features are then projected onto three canonical planes (XY, XZ, YZ), forming three feature maps of size $H \times W \times d$. A shared 2D convolutional network [70] processes these planes to capture local geometric structure.

For a query point x, features are obtained by bilinear interpolation from the feature planes and passed through a decoder:

$$f_\theta(x) = \text{MLP}_\theta(z_x \oplus x),$$

where z_x is the interpolated feature and $\oplus$ denotes concatenation operation. This representation enables efficient encoding of fine-grained local geometry, which is critical for grasp generation on complex objects.

3.3.2 Neural Descriptor Fields

Neural Descriptor Fields represent object geometry through learned feature descriptors associated with spatial locations [71]. Formally, a neural field $f_\theta : \mathbb{R}^3 \to \mathbb{R}^d$ maps any spatial location to a latent descriptor, enabling dense correspondence across object instances. A grasp pose can be represented by transforming a fixed query point cloud P_q using a rigid transformation $H \in SE(3)$ as $P_H = H \cdot P_q$. The descriptor of a grasp pose is then obtained by aggregating the features of transformed query points:

$$f_H = \{f_\theta(x_i) \mid x_i \in P_H\}$$

A key property of NDFs is that they encode not only local geometry but also relative pose relationships between the object and the query. This allows grasp poses to be represented as descriptor sets that are comparable across different object instances. Importantly, NDFs are designed to be $SE(3)$-equivariant [72, 73], meaning that transformations of the input correspond to predictable transformations in the descriptor space. This property enables generalization across arbitrary object poses and spatial configurations. Additionally, the model is typically trained in a self-supervised manner via 3D reconstruction or autoencoding objectives, removing the need for explicit keypoint or part annotations.

This formulation allows grasps to be represented in terms of local geometric features, enabling similarity between grasp configurations and corresponding object regions. As a result, it provides a natural way to relate grasp poses to underlying object geometry.

3.3.3 SE(3) Diffusion Models for Grasp Generation

Recent work formulates grasp generation as a diffusion process over the space of rigid transformations $SE(3)$ [26]. Since $SE(3)$ is a Lie group, diffusion is performed in its associated Lie algebra $\mathfrak{se}(3) \cong \mathbb{R}^6$ using logarithmic and exponential maps:

$$\begin{aligned} \text{Logmap}: \ & SE(3) \to \mathbb{R}^6 \\ \text{Expmap}: \ & \mathbb{R}^6 \to SE(3) \end{aligned}$$

Given a ground-truth grasp pose H, the forward diffusion process perturbs it by adding Gaussian noise in $\mathbb{R}^6$:

$$H_k = \text{Expmap}\left(\text{Logmap}(H) + \epsilon_k\right) \tag{3.1}$$

where $\epsilon_k \sim \mathcal{N}(0, \sigma_k^2 I)$. The model learns a denoising function $\epsilon_\theta(H_k, k, P)$ that predicts the noise added at timestep k, conditioned on the object point cloud P. The training objective minimizes the discrepancy between predicted and true noise:

$$\mathcal{L} = \mathbb{E}_{H,\epsilon,k}\left[\|\epsilon - \epsilon_\theta(H_k, k, P)\|^2\right] \tag{3.2}$$

This formulation represents the general denoising diffusion framework. $SE(3)$-diffusion [26] implements the above as an equivalent score-based formulation [74] that estimates the gradient of the data distribution instead of directly predicting noise. During inference, grasp poses are generated via an iterative denoising process:

$$H_{k-1} = \text{Expmap}\left(\text{Logmap}(H_k) - \alpha_k\, \epsilon_\theta(H_k, k, P) + \sigma_k \epsilon\right) \tag{3.3}$$

where $\epsilon \sim \mathcal{N}(0, I)$, and α_k, σ_k are schedule-dependent coefficients.

This formulation enables the generation of diverse and high-quality grasp configurations by modeling the underlying distribution over feasible grasps. Diffusion-based methods are particularly effective in capturing multimodal distributions and improving sample efficiency in grasp generation tasks.

3.4 CGDF: Constrained Grasp Diffusion Fields

3.4.1 Method

We present **Constrained Grasp Diffusion Fields (CGDF)**, a diffusion-based framework for generating 6-DoF grasps on complex objects under spatial constraints. Given an object point cloud (with n points) $P \in \mathbb{R}^{n\times 3}$ and a target region $P_{\text{target}} \subseteq P$, the goal is to generate a set of grasp poses $\{H_i\}_{i=1}^{M}$, where $H_i \in SE(3)$ and M is user-specified number, such that the grasps are both *stable* and *localized* within the target region P_{target}. We do not impose any constraints on the how detailed the target region should be, allowing P_{target} to vary from a small localized subset to the entire point cloud P. In the case of an n-arm manipulation setup involving large objects, this can be naturally extended to multiple regions $P_{\text{target}} = \{P^1_{\text{target}}, P^2_{\text{target}}, \ldots, P^n_{\text{target}}\}$, each corresponding to a different manipulator.

This setting necessitates learning a robust and expressive shape representation that can generalize across arbitrary object geometries and varying region specifications. Additionally, the model must be capable of producing a diverse set of grasps with sufficient coverage over the specified target regions. Diffusion models are particularly well-suited for this purpose, as they effectively capture complex data distributions while enabling diverse sample generation. Therefore, we adopt a diffusion-based formulation as the backbone of our approach.

Our method builds upon $SE(3)$-equivariant diffusion models and introduces a *part-guided diffusion* mechanism to enable constrained grasp generation without requiring large-scale constraint-annotated training data. The overall formulation consists of three components: (i) a $SE(3)$ diffusion model, (ii) an energy-based scoring function, and (iii) a constrained sampling strategy.

In the next sections, we first (Section 3.4.1.1) explain the diffusion formulation in $SE(3)$, then (Section 3.4.1.2) formulate the *part-guided* strategy to generate grasps on constrained regions and finally (Section 3.4.1.3) outline the CGDF architecture to generate grasps with and without constrained regions.

3.4.1.1 SE(3) Diffusion Formulation

We adopt the formulation of diffusion models on the special Euclidean group $SE(3)$, following prior work [26]. Since a grasp pose H lies on the Lie group $SE(3)$, it does not admit standard Euclidean operations such as addition and subtraction. To enable diffusion in a vector space, we map H to its corresponding Lie algebra $\mathfrak{se}(3) \cong \mathbb{R}^6$ using the logarithmic map:

$$h = \text{Logmap}(H) \in \mathbb{R}^6 \tag{3.4}$$

This mapping allows us to perform standard perturbations in $\mathbb{R}^6$, after which we can project back to $SE(3)$ using the exponential map. The forward diffusion process perturbs grasp poses by injecting

Gaussian noise in the Lie algebra space:

$$\hat{H}_k = \text{Expmap}\left(\text{Logmap}(H) + \epsilon_k\right), \quad \epsilon_k \sim \mathcal{N}(0, \sigma_k^2 I) \tag{3.5}$$

where k denotes the diffusion timestep and σ_k controls the noise magnitude.

In standard diffusion models, the objective is to directly predict the noise ϵ_k added at each timestep [35, 75]. However, in continuous domains such as $SE(3)$, it is often more stable to instead learn the *score function*, i.e., the gradient of the log-density of the data distribution [74].

Following [26], we parameterize this score implicitly using an **Energy-based Model** (EBM) E_θ, which assigns a scalar energy to a grasp pose conditioned on the point cloud:

$$e_k = E_\theta(H_k, k, P) \tag{3.6}$$

The score function is then obtained as the negative gradient of this energy with respect to the pose:

$$s_\theta(H_k, k, P) = -\frac{\partial E_\theta(H_k, k, P)}{\partial H_k} \tag{3.7}$$

This formulation provides two advantages: (i) it naturally induces a scoring mechanism over grasps, where lower energy corresponds to higher-quality grasps, and (ii) it ensures a smoother and more stable vector field over $SE(3)$ compared to directly regressing noise.

The model is trained using a denoising objective, where the predicted score is matched to the ground-truth normalized noise:

$$\mathcal{L}_{\text{diff}} = \left\| s_\theta(H_k, k, P) - \frac{\epsilon_k}{\sigma_k} \right\|_1 \tag{3.8}$$

During inference, grasp poses are iteratively updated over timesteps $k = \{K, K-1, \ldots, 0\}$ using Langevin dynamics [76]:

$$H_{k-1} = \text{Expmap}\left(\frac{\alpha_k}{2} s_\theta(H_k, k, P) + \alpha_k \epsilon\right) H_k \tag{3.9}$$

where $\epsilon \sim \mathcal{N}(0, I)$ and α_k is a step-dependent coefficient. This process gradually moves noisy samples toward low-energy regions corresponding to valid grasp configurations.

3.4.1.2 Part-Guided Strategy

The diffusion formulation described above generates grasps over the entire object in an unconstrained manner. However, our goal is to generate grasps that are restricted to a specified target region $P_{\text{target}} \subseteq P$. A key observation is that the learned energy function E_θ already encodes grasp quality. In particular,

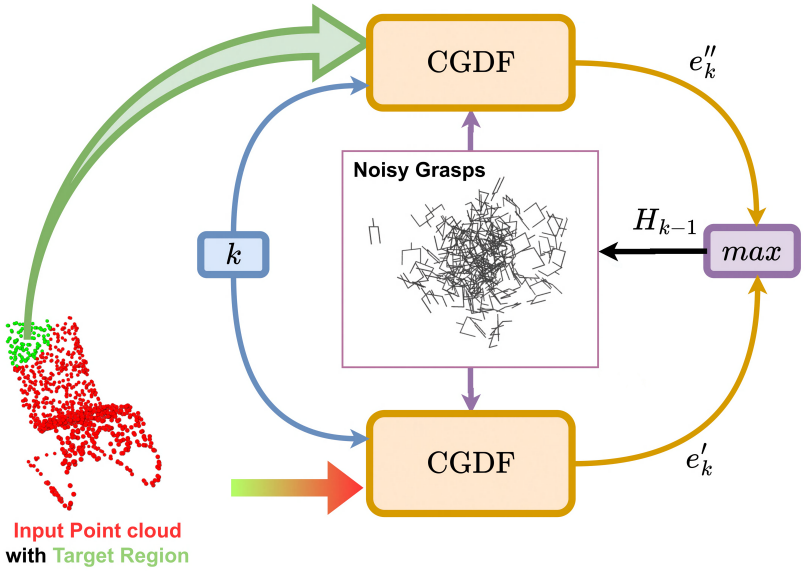

Figure 3.2: Part-guided diffusion strategy: We demonstrate the approach using two CGDF instances, one conditioned on the full point cloud and the other on target region.

> *grasps that are unstable or in collision correspond to high-energy configurations, whereas stable grasps lie in low-energy regions*

We use this property to enable constrained grasp generation *without retraining* the model. Given a candidate grasp H, we evaluate its quality under two contexts: the full object point cloud P and the target region P_t:

$$e' = E_\theta(H, 0, P), \tag{3.10}$$

$$e'' = E_\theta(H, 0, P_{\text{target}}) \tag{3.11}$$

Here, e' measures whether the grasp is globally stable with respect to the full object P, while e'' measures whether the grasp is consistent with the local geometry of the target region P_{target}. A grasp is considered a valid *constrained grasp* only if it satisfies both criteria:

$$e' < \delta \quad \text{and} \quad e'' < \delta \tag{3.12}$$

where δ is a user-defined threshold. Intuitively, this ensures that the grasp is both physically feasible on the object and localized within the desired region. While the above formulation provides a filtering criterion, directly sampling grasps on P_{target} and thresholding based on energy is sample-inefficient. To address this, we introduce a **part-guided diffusion** strategy that integrates the constraint directly into the sampling process.

As shown in Figure 3.2, we guide the diffusion process using two instances of the energy model — one conditioned on the full point cloud P and the other on the target region P_{target}. At each diffusion

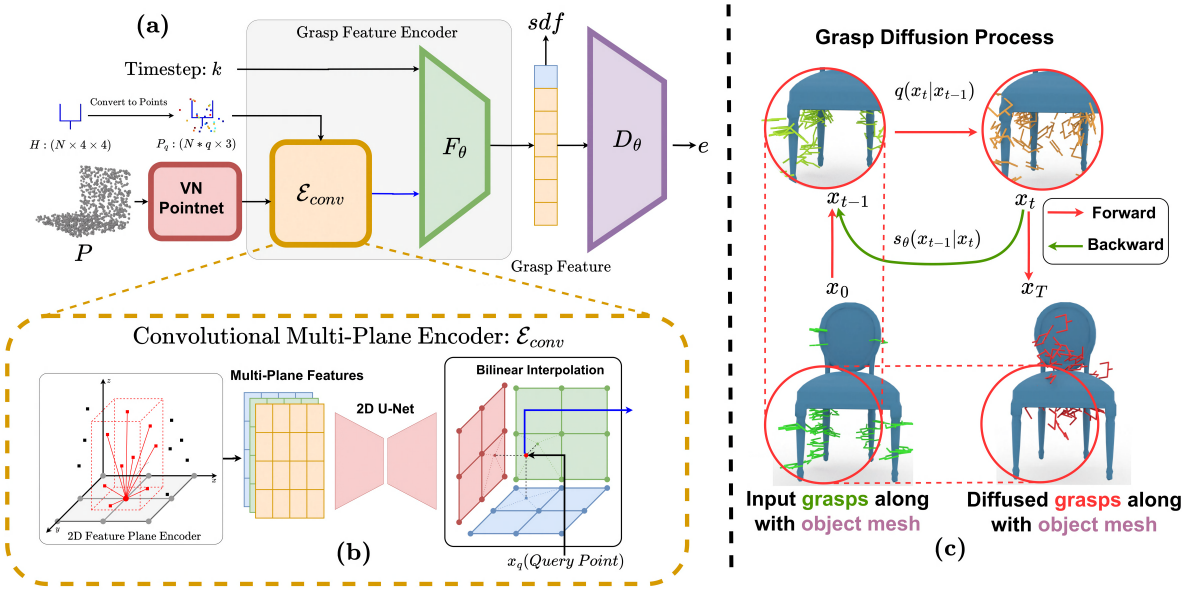

Figure 3.3: CGDF: Constrained Grasp Diffusion Fields, enables the generation of dense grasp configurations on large objects with complex geometries, such as chairs, within a dual-arm manipulation setting. Given target regions, the approach employs a part-guided diffusion strategy to efficiently generate grasps localized to these regions. This facilitates coordinated grasping across multiple regions, making it well-suited for multi-arm manipulation scenarios.

step k, we define a combined energy:

$$e_k = \max \left\{ E_\theta(H_k, k, P),\ E_\theta(H_k, k, P_{\text{target}}) \right\} \tag{3.13}$$

The corresponding score function is given by:

$$s_\theta(H_k, k, \{P, P_t\}) = \begin{cases} s_\theta(H_k, k, P) & \text{if } e'_k \geq e''_k, \\ s_\theta(H_k, k, P_{\text{target}}) & \text{otherwise} \end{cases} \tag{3.14}$$

where $e'_k = E_\theta(H_k, k, P)$ and $e''_k = E_\theta(H_k, k, P_{\text{target}})$.

This formulation provides an adaptive guidance mechanism during diffusion. If a grasp is near the target region but unstable with respect to the full object P (high e'_k), the model uses the global context to improve stability. Conversely, if the grasp is stable but far from the target region P_{target} (high e''_k), the model uses the local context to guide it toward the desired region. Through this iterative process, grasp samples are progressively refined toward configurations that simultaneously minimize both energies. As a result, the diffusion converges to grasps that are both stable and localized within the target region, while maintaining high sample efficiency.

3.4.1.3 CGDF Architecture

A critical component of CGDF is a shape representation that can capture fine-grained local geometry, which is essential for generating grasps on large objects with complex structures. Prior works such as Neural Descriptor Fields (NDF) [71] and GIGA [24] demonstrate that jointly learning object geometry and grasping leads to expressive grasp descriptors. However, these methods rely on a *global* point cloud embedding as the conditioning signal, which is insufficient for large objects where grasp feasibility depends heavily on local geometric variations.

To address this limitation, we incorporate **convolutional multi-plane features** inspired by Convolutional Occupancy Networks [67], enabling the model to encode spatially localized geometry in a translation-equivariant manner. As illustrated in Figure 3.3, the architecture consists of three stages: (i) point cloud encoding, (ii) multi-plane feature construction, and (iii) query-based grasp descriptor extraction.

Point Cloud Encoding and Multi-Plane Features. Given the input point cloud P, we first extract per-point features using a VN-PointNet encoder [77], which preserves $SO(3)$-equivariance. These features are then projected onto three canonical planes aligned with the coordinate axes (XY, XZ, YZ). For each plane, features are aggregated onto a regular grid via projection and pooling, producing three feature maps of size $H \times W \times d$. These planar feature maps are further processed using a shared 2D U-Net [70], which enables multi-scale feature aggregation and enforces translation equivariance. This representation allows efficient storage of dense local geometric information while maintaining scalability to large scenes.

Query-Based Feature Extraction. To evaluate a grasp pose $H \in SE(3)$, we follow a query-based formulation inspired by NDF [71]. A fixed set q of query points $P_q \in \mathbb{R}^{q\times 3}$ is transformed using the grasp pose as $P_H = H \cdot P_q$. For each transformed query point $x_q \in P_H$, features are obtained by bilinearly interpolating from the three feature planes and passing through a feature encoder:

$$f^k_{x_q} = F_\theta(z_{x_q} \oplus x_q, k) \tag{3.15}$$

where z_{x_q} denotes the interpolated multi-plane feature at x_q, $\oplus$ denotes concatenation, and k is the diffusion timestep. Importantly, for each $x_q \in P_H$, F_θ jointly predicts (i) a signed distance function (SDF) value and (ii) a latent feature vector:

$$f^k_{x_q} = \{\text{sdf}(x_q), F_\theta(z_{x_q} \oplus x_q, k)\} \tag{3.16}$$

thereby coupling surface reconstruction with grasp representation learning. This joint formulation encourages the learned features to encode both geometric structure and grasp-relevant semantics which is used by the diffusion module to carry out the grasp generation process.

Grasp Descriptor and Energy Prediction. The grasp descriptor is constructed by aggregating features over all query points:

$$f_H^k = \bigcup_{x_i \in P_H} f_{x_i}^k \tag{3.17}$$

This descriptor effectively captures the interaction between the gripper geometry and the local object surface. The aggregated feature is then passed through a decoder network D_θ (implemented as an MLP) to predict the energy $e_k \in \mathbb{R}$ of the grasp:

$$e_k = D_\theta(f_H^k) \tag{3.18}$$

Finally, the model is trained using the objective described in Equation 3.8 along with an auxiliary L_1 loss objective to supervise the SDF prediction.

3.5 Experiments and Results

We evaluate CGDF on both **unconstrained** and **constrained** grasp generation tasks, with a focus on *large objects with complex geometries* in a dual-arm setting. The primary goal is to assess (i) the quality of generated grasps, (ii) their alignment with specified target regions, and (iii) the overall sample efficiency of the method.

While grasp quality can be evaluated in a single-arm setting, large objects often require coordinated multi-arm manipulation. Therefore, we perform evaluation in a dual-arm setup, which provides a more realistic and stringent test of grasp stability and feasibility.

For all experiments, object point clouds are uniformly sampled with 1000 points. In the constrained setting, target regions are generated by selecting seed points using farthest point sampling and extracting their local neighborhoods via k-nearest neighbors (with $k = 100$). This results in regions of varying geometric complexity, ranging from small planar patches to elongated structures such as chair legs or handles.

3.5.1 Datasets

We train and evaluate our model on the DA^2 dataset [43], which is specifically designed for dual-arm grasping. This dataset contains grasp annotations on large-scale objects with diverse and complex geometries, such as chairs, lamps, and instruments. Unlike tabletop datasets, objects in DA^2 are kept at their original scale, making shape understanding significantly more challenging. The underlying object meshes in DA^2 are sourced from ShapeNetSem [78], which includes a wide variety of object categories with limited instances per class. This increases the difficulty of learning robust shape priors, particularly for methods relying on global representations.

For comparison with prior constrained grasping methods, we augment the DA^2 dataset to simulate target regions and associated grasps, following the protocol used in prior work [23]. However, it

is important to note that our model is trained only on the original (unconstrained) dataset, and does not require constraint-specific annotations.

To evaluate generalization, we additionally use PartNet [45] objects with region proposals generated using a language-conditioned segmentation model (e.g., text prompts such as "handle of a microwave"). This allows us to test CGDF in a realistic pipeline where constraints are obtained from high-level semantic inputs.

3.5.2 Baselines

We compare CGDF against two representative baselines $SE(3)$-DiF [26] and VCGS [66]. These baselines provide a strong comparison across both unconstrained and constrained settings.

1. $SE(3)$-Diffusion Fields ($SE(3)$-DiF) [26]: This serves as the primary baseline for unconstrained grasp generation. It models grasp synthesis as a diffusion process over $SE(3)$ using an energy-based formulation. However, it relies on global shape embeddings, which limits its ability to capture fine-grained local geometry.

2. Variational Constrained Grasp Sampling (VCGS) [66]: This is a state-of-the-art method for *constrained* grasp generation. It consists of a sampler that generates candidate grasps on target regions and an evaluator that ranks them. Unlike our approach, VCGS requires training on a dataset augmented with constraint-specific annotations.

3.5.3 Metrics

We evaluate grasp quality using three complementary metrics:

1. Force Closure (FC). This is a strict analytical metric that evaluates whether a grasp can resist arbitrary external wrenches. It considers both contact geometry and collisions between the gripper and the object. In our setup, dual-arm grasps involve four contact points (two per gripper), and a grasp is considered valid if it satisfies force closure constraints.

   To compute contact points, we simulate the gripper geometry at the predicted pose and reject grasps that intersect with the object mesh. We then cast rays from the gripper fingers to estimate contact locations and normals. FC is reported as the percentage of grasps satisfying these conditions out of all generated grasps.

2. Grasp Success Rate (GSR). This metric evaluates grasp performance in simulation. We use two free-floating grippers in an Isaac Gym environment [37] to execute the predicted grasps. The grippers close simultaneously, and the object is lifted under gravity. A grasp is considered successful if the object remains securely held after lifting.

   Compared to FC, GSR provides a more practical measure of grasp robustness under dynamic conditions and minor modeling inaccuracies.

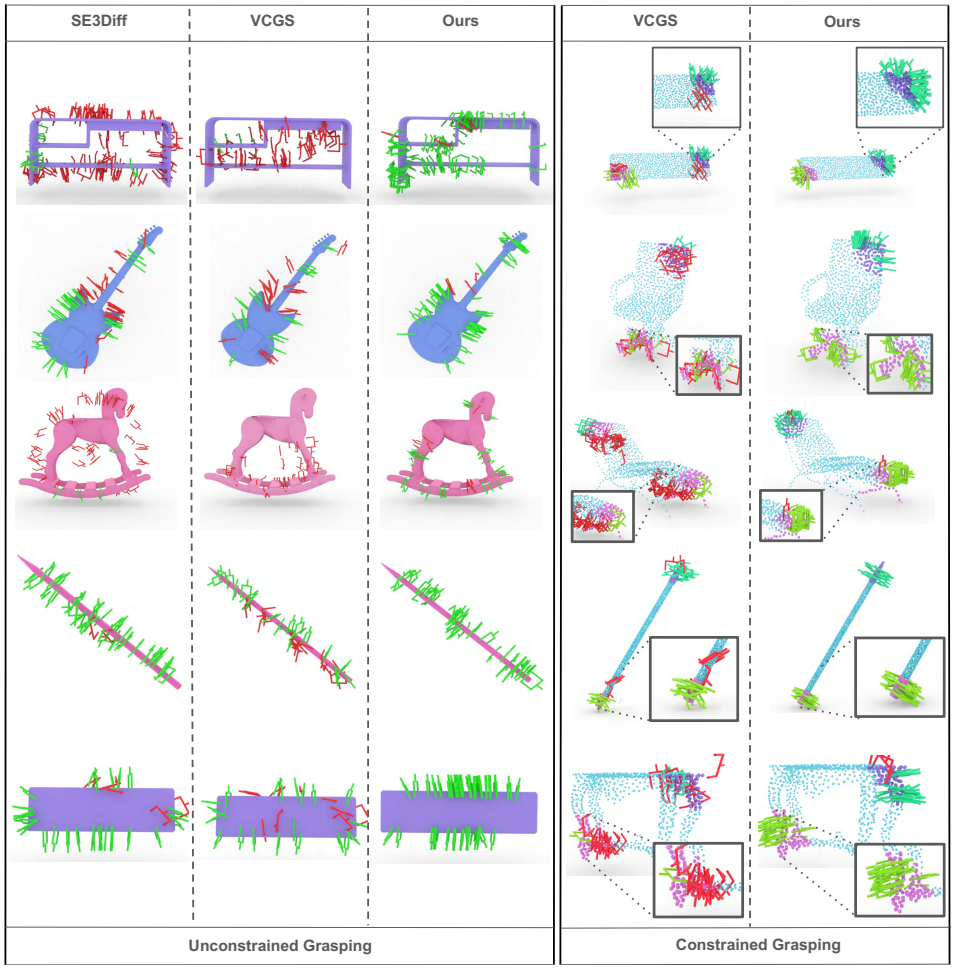

Figure 3.4: Qualitative comparison of unconstrained and constrained dual-arm grasp generation using CGDF and baseline methods (VCGS, $SE(3)$-DiF). While all methods perform well on simple geometries (e.g., planar or elongated objects), CGDF produces dense and feasible grasps for more complex objects (e.g., chairs, instruments), where the baselines struggle. Green grasps indicate non-colliding configurations, while red grasps denote collisions. Zoom in for details.

3. Target Grasp Rate (TG). This metric measures how well the generated grasps adhere to the specified target regions. For each grasp, we compute the minimum distance between the grasp center and the points in the target region. Grasps that lie beyond a fixed threshold (6 cm in our setup) are discarded. TG is defined as the percentage of grasps that lie within the target region, and is particularly important for evaluating constrained grasp generation.

Together, these metrics capture complementary aspects of performance: **FC** evaluates physical feasibility, **GSR** measures execution success, and **TG** quantifies adherence to constraints.

3.5.4 Results

Qualitative Results. We present qualitative comparisons of CGDF against $SE(3)$-DiF and VCGS in both unconstrained and constrained grasp generation settings.

Metric	Methods	Unconstrained	Constrained
FC (%)↑	VCGS	3.28	3.96
	$SE(3)$-DiF	14.22	–
	CGDF (Ours)	**43.51**	**44.8**
GSR (%)↑	VCGS	41.01	43.36
	$SE(3)$-DiF	46.2	–
	CGDF (Ours)	**60.3**	**60.88**
TG (%)↑	VCGS	100	76.2
	CGDF (Ours)	100	**91.86**

Table 3.1: Quantitative comparison of CGDF with $SE(3)$-DiF [26] and VCGS [66] under both unconstrained and constrained settings. Force Closure (FC) measures the ability of a grasp to resist external disturbances, Grasp Success Rate (GSR) denotes the percentage of successful executions, and Target Grasp (TG)% indicates the proportion of grasps on the specified target regions.

In the **unconstrained setting**, all methods perform reasonably well on simple geometries such as elongated or planar objects. However, for complex objects with intricate structures (e.g., table, guitars, and rocking horses), clear differences emerge. As observed in Figure 3.4, $SE(3)$-DiF often produces sparse grasp distributions with limited coverage, while VCGS generates grasps that are more localized but frequently include collisions (indicated by red grasps). In contrast, CGDF produces *dense and well-distributed grasps* across the object surface, with a significantly higher proportion of collision-free grasps (green), demonstrating better modeling of local geometry.

In the **constrained setting**, the differences are more pronounced. VCGS is able to generate grasps near the target region, but fails to accurately capture fine geometric details, resulting in many grasps that partially intersect the object. This is especially evident in thin structures such as handles or legs, where slight geometric inaccuracies lead to collisions. On the other hand, CGDF generates grasps that are both *localized to the target region* and *physically feasible*, showing strong alignment with the constraint while maintaining stability.

These observations highlight two key advantages of CGDF: (i) improved *local geometric understanding*, and (ii) higher *sample efficiency*, as fewer invalid grasps are produced. Overall, CGDF demonstrates superior performance in generating dense, stable, and region-consistent grasps on complex objects, which is critical for dual-arm manipulation scenarios.

Quantitative Results. We quantitatively evaluate CGDF against $SE(3)$-DiF and VCGS using Force Closure (FC), Grasp Success Rate (GSR), and Target Grasp (TG) metrics. The results are summarized in Table 3.1.

CGDF significantly outperforms both baselines across all metrics. In the **unconstrained setting**, CGDF achieves an FC of 43.51%, compared to 14.22% for $SE(3)$-DiF and 3.28% for VCGS. This indicates a substantial improvement in generating physically stable and collision-free grasps. Similarly,

CGDF attains the highest GSR of 60.3%, outperforming both $SE(3)$-DiF (46.2%) and VCGS (41.01%), demonstrating better real-world execution robustness.

In the **constrained setting**, CGDF maintains its advantage with an FC of 44.8% and GSR of 60.88%, significantly higher than VCGS. Notably, CGDF also achieves a TG score of 91.86%, indicating strong adherence to the specified target regions while preserving grasp quality.

The relatively poor FC performance of VCGS, despite reasonable TG scores, suggests that while it can localize grasps to the target region, it struggles to model fine geometric details, resulting in collisions. $SE(3)$-DiF, although better in avoiding collisions, lacks the ability to localize grasps effectively due to its global representation.

Overall, these results demonstrate that CGDF achieves a better balance between *stability*, *execution success*, and *constraint satisfaction*, making it well-suited for constrained grasp generation on complex objects.

Part-Based Grasping with PartSLIP. To evaluate grasp generation on semantically meaningful object regions, we integrate CGDF with PartSLIP [50], a language-guided part segmentation model that extracts target regions from text prompts. Given a prompt (e.g., "handle of a microwave" or "legs of a table"), PartSLIP produces a segmented point cloud corresponding to the desired functional part, which is then used as the constraint region P_t. As shown in Figure 3.5, CGDF successfully generates dense and stable grasps localized to these segmented regions, even on complex objects with fine structures. This demonstrates the ability of our method to leverage high-level semantic cues for grasp generation without requiring explicit training on part-annotated datasets. In contrast to traditional pipelines that rely on post-hoc filtering of uniformly sampled grasps, our approach directly incorporates part-level information into the generation process, resulting in improved sample efficiency and better alignment with task-relevant regions.

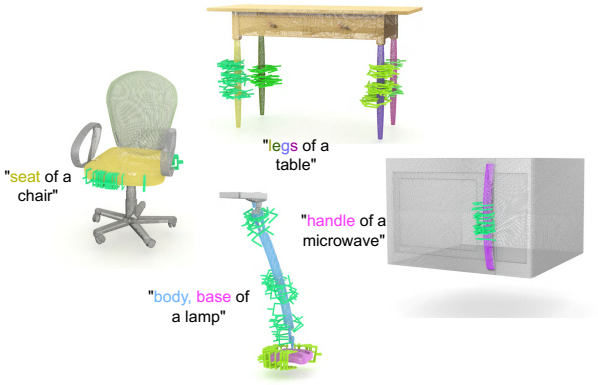

Figure 3.5: Qualitative results illustrating region segmentation via PartSLIP from textual prompts, followed by CGDF generating grasps on the segmented regions.

3.5.5 Ablations

We analyze the contribution of the two key components of CGDF: (i) convolutional plane features, and (ii) part-guided diffusion.

(i) Effect of Convolutional Plane Features: To evaluate the role of local geometric features, we compare CGDF with a variant that relies on global features (i.e., without convolutional plane encoding).

Since both models jointly learn an SDF representation, we assess reconstruction quality using Chamfer Distance (CD) between predicted and ground-truth point clouds. As shown in Table 3.2, CGDF achieves a significantly lower CD (14.04) compared to the variant without convolutional features (60.21), indicating a substantial improvement in capturing fine-grained geometry. This directly translates to better grasp generation, as accurate local geometry is critical for identifying stable contact regions.

Modification	**Metric**	**Ours**	**Modified**
w/o Conv	CD ↓	**14.04**	60.21
w/o PD	SE (%)↑	**93.6**	37.4

Table 3.2: Ablation study showing the contribution of convolutional features and part-guided diffusion.

(ii) Effect of Part-Guided Diffusion: We evaluate the importance of the part-guided strategy by comparing against a naive approach where grasps are generated on the target region and filtered using energy thresholds. We measure *sample efficiency* as the percentage of grasps retained after filtering. The results show that part-guided diffusion improves sample efficiency from 37.4% to 93.6%, representing a $2.5\times$ improvement. This demonstrates that integrating constraints directly into the diffusion process is significantly more efficient than post-hoc filtering.

Together, these ablations confirm that both design choices are critical: convolutional plane features enable accurate geometric understanding, while part-guided diffusion ensures efficient and effective constrained grasp generation.

3.6 Chapter Summary

In this chapter we introduced **CGDF**, a diffusion-based framework for generating dense, region-aware grasps on large and complex objects, addressing key limitations of prior methods that rely on uniform sampling and global representations. By combining $SE(3)$ diffusion with an energy-based formulation and a novel part-guided strategy, CGDF efficiently produces grasps that are both physically stable and localized to specified regions, without requiring constraint-labeled data. The use of multi-plane feature representations further enables accurate modeling of fine-grained local geometry, leading to significant improvements in grasp quality, success rate, and sample efficiency in both unconstrained and constrained settings. In our dual-arm setup, grasps are constructed by pairing single-arm grasps from the two farthest regions, which provides a simple way to enable bimanual interaction but does not provide a principled formulation of dual-arm grasping. This naturally motivates the next chapter, where we shift focus from grasp generation to a deeper analysis of dual-arm grasp stability, introducing metrics and formulations to rigorously evaluate the robustness of generated grasps.

Chapter 4

Evaluating the Stability of Dual-Arm Grasps

Lots of machine learning researchers from other fields don't understand how hard robotics benchmarking is .... In the end, robotics benchmarking will be solved by having lots and lots of robots, and models that actually work across most of them.

CHRIS PAXTON in *How to Fake A Robotics Result*, 2026

4.1 Introduction

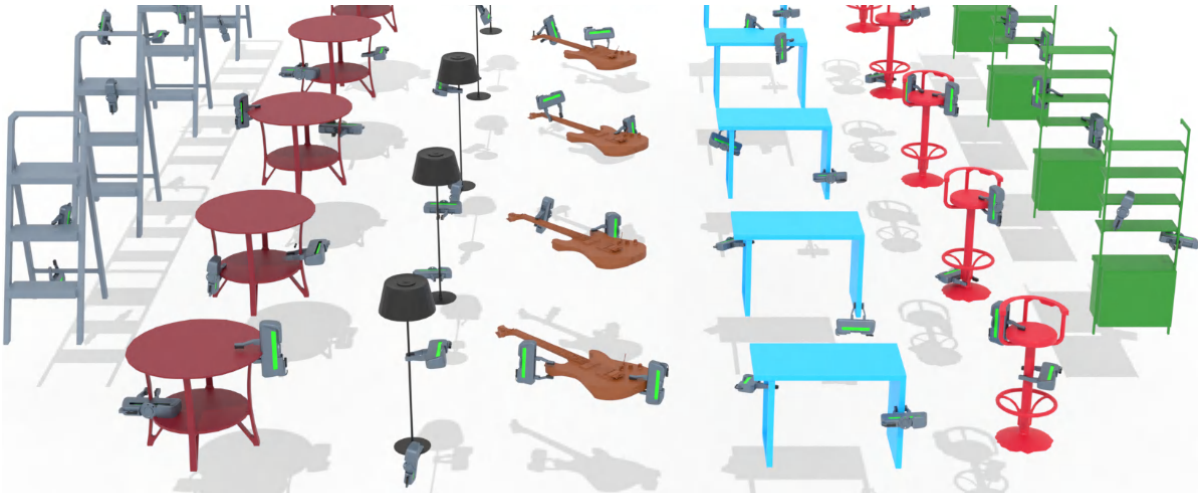

Figure 4.1: DG16M is a large-scale dataset of dual-arm grasps constructed using an optimization-based force closure formulation to ensure physically stable and feasible grasp configurations.

Consider lifting a computer monitor using both hands. Intuitively, we grasp it from the sides rather than the top or bottom, which ensures that the object remains balanced and does not slip or rotate. This choice is not arbitrary and it reflects our implicit understanding of force balance, torque equilibrium, and stable contact. Translating such intuition to robotic systems, however, is non-trivial. Evaluating whether a pair of grasps can jointly stabilize an object under real-world con-

ditions remains a challenging problem in robotic manipulation [36]. While recent learning-based methods can generate diverse and feasible grasp candidates [14, 15, 17, 23, 66], they often lack mechanisms to reliably assess the physical stability of these grasps. Unlike single-arm settings, dual-arm grasping involves coordinated interaction between multiple contact points, requiring a global analysis of force balance, torque equilibrium, and inter-gripper consistency [4, 5, 7]. This makes stability evaluation significantly more complex and highlights the need for physically grounded methods that go beyond purely geometric reasoning.

A central concept in grasp evaluation is *force closure*, which determines whether a set of contact forces can resist arbitrary external disturbances [1]. Traditional analytical approaches assess this property using friction cone approximations and wrench space analysis, often relying on metrics such as the Ferrari-Canny epsilon [3]. However, these formulations are typically developed for single-arm or simplified contact scenarios and do not fully capture the complexities introduced by dual-arm grasping — such as inter-gripper coordination, force distribution, and actuator constraints.

Existing dual-arm grasp evaluation method, DA^2 [43] relies on simplified analytical assumptions, which introduces a mismatch between theoretically valid grasps and those that remain stable in practice. In many cases, grasp pairs that satisfy geometric criteria, such as antipodal configurations fail to maintain stability under realistic conditions. These failures typically arise from (i) inadequate force balance, (ii) suboptimal placement with respect to the object's center of mass, or (iii) violations of actuator and gripper force limits. Such observations highlight the limitations of purely geometric or kinematic reasoning for grasp evaluation and motivate the need for approaches that explicitly account for the underlying physical constraints.

In this chapter, we present a physically grounded framework for evaluating the stability of dual-arm grasps based on an optimization-based force closure formulation. Our method explicitly models contact forces and enforces frictional and actuator constraints to determine whether a grasp can achieve equilibrium under external disturbances. Using this evaluation framework, we construct a large-scale dual-arm grasp dataset called **DG16M**, comprising over 16 million grasp pairs across 4,143 objects (Figure 4.1 visualizes a small subset). Candidate grasps are systematically generated and subsequently filtered using the proposed force closure criterion, ensuring that only physically feasible and stable configurations are retained. This process results in a new high-quality dataset with well-defined grasp labels, enabling more reliable training and benchmarking of dual-arm grasping systems. The project page and code can be accessed here: https://dg16m.github.io/DG-16M/.

4.2 Related Work

4.2.1 Grasp Sampling Methods

Grasp sampling methods aim to explore the space of possible grasp configurations to identify stable and feasible grasps. Classical approaches rely on geometric heuristics and analytical formulations, while more recent methods leverage large-scale datasets and learning-based models. A widely

Dataset	Grasps per Obj	Grasp Label	Total Obj	Total Grasps	Gripper Type
ACRONYM [42]	2000	6-DoF	8872	17.7M	Single
DA^2 [43]	up to 2001	6-DoF	6327	9M	Dual
DG16M (ours)	up to 4000	6-DoF	4132	16M	Dual

Table 4.1: Comparison of grasp datasets based on grasp density, labeling method, object count, total grasps, and gripper types supported.

used strategy is **approach-based sampling** [79–82], where the gripper approach direction is aligned with the local surface normal of the object. This improves contact stability by encouraging perpendicular interactions, but can introduce bias by restricting the diversity of feasible grasps and potentially missing valid configurations that deviate from surface normals.

Another dominant paradigm is **antipodal grasp sampling** [83], which selects pairs of contact points that satisfy force closure conditions under friction constraints. This method has been extensively used in grasp planning systems such as Dex-Net 2.0 [22] and datasets like ACRONYM [42], REGRAD [84] and DA^2 [43], owing to its simplicity and effectiveness in generating stable grasps for parallel-jaw grippers. However, while antipodal sampling works well for single-arm grasping, extending it to dual-arm settings is non-trivial, as it does not inherently account for the joint stability between both the grippers. As a result, naively pairing single-arm grasps often leads to unstable configurations that fail to satisfy global force and torque balance.

4.2.2 Force Closure

Force closure is a fundamental concept in robotic grasping, ensuring that a grasp can resist arbitrary external disturbances through appropriate contact forces. Classical formulations compute force closure by constructing a convex hull over discretized friction cones in wrench space, under assumptions such as normalized contact forces [1, 10]. Metrics such as the Ferrari-Canny epsilon are commonly used to quantify grasp robustness [3].

Recent work has explored differentiable and optimization-based formulations of force closure to better integrate with learning frameworks. For instance, [85] propose a differentiable force closure estimator by introducing relaxations such as equal force magnitudes and simplified friction models. While such relaxations improve tractability, they may not accurately capture the constraints of real robotic grippers, particularly in multi-contact or dual-arm scenarios. Additionally, optimization-based approaches provide a more physically grounded alternative by explicitly solving for contact forces that satisfy equilibrium under friction and actuator constraints. PhyGrasp [86] adopts such a formulation to compute feasible force distributions; however, it samples contact points without strong geometric constraints, which can lead to physically infeasible grasps.

Extending these formulations to dual-arm settings introduces additional challenges, including coordination between multiple contact points, force distribution across grippers, and adherence to

hardware limits. These considerations necessitate more constrained and physically consistent formulations for reliable grasp evaluation.

4.2.3 Benchmarking Datasets for Grasping

Large-scale datasets have played a critical role in advancing learning-based grasping. Single-arm grasp datasets such as Dex-Net [22], GraspNet-1Billion [25], Jacquard [87], and REGRAD [84], CMU Grasp Dataset [88], EGAD [89] provide extensive annotations in the form of grasp poses or image-based labels. These datasets typically focus on small to medium-sized objects and are often constructed using synthetic data or physics simulations to ensure grasp validity. Additionally, datasets like ACRONYM [42] leverage analytic metrics combined with simulation to produce large-scale, high-quality grasp annotations.

In contrast, datasets for dual-arm grasping remain relatively limited. The DA^2 dataset [43] represents the first large-scale efforts in this direction, containing over 9 million of dual-arm grasps across diverse objects from the ShapeNet dataset [78]. However, its grasp evaluation relies on simplified force closure assumptions that do not explicitly account for actuator constraints or optimal force distributions. As a result, many grasps in the dataset may be analytically valid but fail under realistic physical conditions, reducing their effectiveness for training robust models.

4.3 DG16M: A Large-Scale Dataset for Dual-Arm Grasping with Force-Optimized Grasps

This section presents the methodology for constructing DG16M dataset. We begin by introducing an optimization-based force-closure framework (Section 4.3.1) that evaluates the stability of dual-arm grasps by explicitly solving for contact forces under frictional and actuator constraints. Building on this formulation, we then describe the dataset generation pipeline (Section 4.3.2), where candidate grasp pairs are systematically sampled and filtered using the proposed optimizer to retain only physically stable configurations.

4.3.1 Improved Force-Closure Optimizer

Evaluating the stability of dual-arm grasps requires reasoning about the ability of multiple contact points to collectively resist external disturbances. Unlike single-arm grasping, where stability is largely governed by local contact properties, dual-arm grasping involves coordinated force interactions across multiple contacts. To capture this, we formulate grasp evaluation as a **QP Optimization Problem with SOCP constraints** that explicitly solves for feasible contact forces under physical constraints.

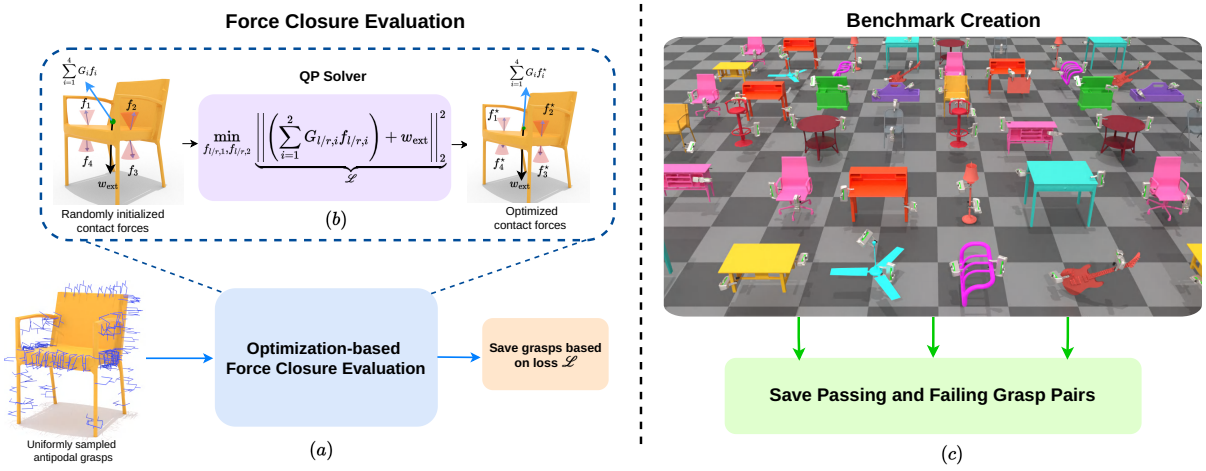

Figure 4.2: Overview of the proposed pipeline: (a) Antipodal grasps are sampled and combined into dual-arm grasp pairs, followed by distance-based pruning. (b) Each pair is evaluated using an optimization-based force-closure framework that enforces friction and actuator constraints to identify stable grasps. (c) A subset of grasps is further validated in simulation to construct a benchmark dataset with ground-truth stability labels.

4.3.1.1 Contact Model and Grasp Representation

We consider a dual-arm setup with two parallel-jaw grippers, resulting in four contact points on the object. Each contact follows the Point Contact with Friction (PCWF) model [10]. Let the position and orientation of the i^{th} contact of the left/right gripper be denoted by $\mathbf{p}_{l/r,i} \in \mathbb{R}^3$ and $\mathbf{R}_{l/r,i} \in SO(3)$, where $i \in \{1, 2\}$. The contact force at each point is decomposed into normal ($f^n_{l/r,i}$) and tangential components ($f^t_{l/r,i}$) :

$$\mathbf{f}_{l/r,i} = \mathbf{f}^t_{l/r,i} + \mathbf{f}^n_{l/r,i} \tag{4.1}$$

Each contact contributes to the object wrench through a grasp matrix $\mathbf{G}_{l/r,i}$:

$$\mathbf{G}_{l/r,i} = \begin{bmatrix} \mathbf{R}_{l/r,i} & \mathbf{0} \\ [\mathbf{p}_{l/r,i}]_\times \mathbf{R}_{l/r,i} & \mathbf{R}_{l/r,i} \end{bmatrix} \cdot \mathbf{B} \tag{4.2}$$

where $[\mathbf{p}]_\times \in \mathbb{R}^{3\times3}$ denotes the skew-symmetric matrix and $\mathbf{B} = [\mathbf{I}_{3\times3} \;\; \mathbf{0}_{3\times3}]^T \in \mathbb{R}^{6\times3}$ is the contact basis for the PCWF model. The full grasp matrix $\mathbf{G}$ is obtained by concatenating all contact matrices.

Feasibility via Rank Condition. Before evaluating force closure, we ensure that the grasp configuration is capable of generating arbitrary wrenches. This is verified through a rank condition on the grasp matrix [10]:

$$\boxed{\text{rank}(\mathbf{G}) = 6} \tag{4.3}$$

This condition ensures that the set of contact wrenches spans the entire 6D wrench space, enabling the grasp to generate arbitrary combinations of forces and torques on the object. If the grasp matrix is rank-deficient, the corresponding contact configuration lacks the ability to counteract certain disturbance directions, making the grasp inherently unstable regardless of the chosen contact

forces. Therefore, the rank condition serves as a necessary prerequisite for force closure, filtering out geometrically valid but physically incapable grasp configurations before proceeding to force optimization.

4.3.1.2 Optimization-based Force Closure

To evaluate grasp stability, we seek a set of contact forces that can counteract an external wrench $\mathbf{w}_{\text{ext}}$ acting at the object's center of mass (for example, the weight of the object). Intuitively, a grasp is stable if the contact wrenches can collectively balance external disturbances, resulting in a net zero wrench on the object. This is equivalent to requiring that the external wrench lies within the span of the contact wrench space generated by the grasp matrix.

We formulate this as an optimization problem that minimizes the L_2 residual wrench:

$$\mathcal{L} = \left\| \sum_i \mathbf{G}_{l/r,i} \mathbf{f}_{l/r,i} + \mathbf{w}_{\text{ext}} \right\|_2^2 \tag{4.4}$$

The optimization problem is formulated as:

$$\min_{\{\mathbf{f}_{l,i},\, \mathbf{f}_{r,i}\}} \quad \mathcal{L} \tag{4.5}$$

such that,

$$\|\mathbf{f}_{l/r,i}\|_2 \leq f_{\max}, \quad \forall i \in \{1,2\}, \tag{4.6}$$

$$\|\mathbf{f}^t_{l/r,i}\|_2 \leq \mu\, \mathbf{f}^n_{l/r,i}, \quad \forall i \in \{1,2\} \tag{4.7}$$

where $f_{\max}$ denotes the actuator force limit and μ is the friction coefficient. The first constraint (4.6) enforces hardware limits on the gripper forces, while the second (4.7) ensures that the contact forces lie within the friction cone.

Grasp Stability Criterion. The above problem is a Second-Order Cone Program (SOCP) and is solved using a convex optimization solver such as CVXPY [90]. A grasp is considered stable if the optimal residual satisfies:

$$\boxed{\mathcal{L} \leq \delta} \tag{4.8}$$

where δ is a small user-defined threshold ensuring near-zero net wrench on the object. For all our experiments, we choose $\delta = 10^{-5}$.

Unlike traditional force closure formulations that assume fixed or normalized contact forces, this approach explicitly computes feasible force distributions while respecting frictional and actuator constraints. This leads to a more physically accurate evaluation, particularly in dual-arm settings where coordinated force interactions are essential. Using this evaluation framework, we next describe the dataset creation pipeline, where candidate dual-arm grasps are generated and systemati-

cally filtered using the proposed optimizer to construct a large-scale dataset of physically stable grasp configurations.

4.3.2 Dataset Generation Pipeline

The dataset generation pipeline consists of three main stages: (i) sampling candidate dual-arm grasps, (ii) evaluating their physical feasibility using the proposed force-closure optimizer, and (iii) selecting stable grasp pairs to construct the final dataset. An overview of this pipeline is shown in Figure 4.2.

4.3.2.1 Grasp Sampling

To generate candidate grasps, we adopt the **antipodal grasp sampling** strategy [22] (also shown in Figure 4.3), a widely used method for generating stable contact configurations. The process begins by randomly sampling a point on the object surface along with its corresponding surface normal. A second contact point is then identified by tracing a ray within a constrained region defined around the normal direction, ensuring that the resulting pair satisfies antipodal conditions.

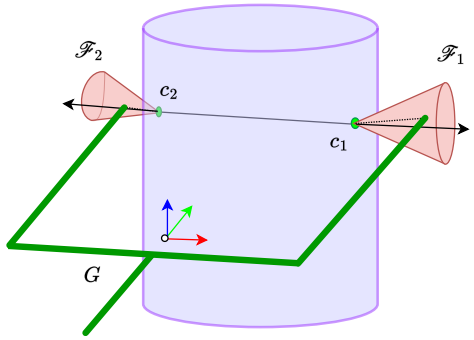

Figure 4.3: Antipodal grasp representation showing contact points c_1, c_2, corresponding friction cones, and the resulting grasp G.

The allowable region for the second contact is determined by a friction-dependent cone, where the threshold is given by $\gamma = \tan(\alpha/2)$ and cone angle $\alpha = \tan^{-1}(\mu)$, with μ representing the friction coefficient [1]. This ensures that the sampled contact pair lies within the friction cone and is capable of producing stable contacts.

To ensure physical validity, collision checks are performed to discard grasp configurations that result in interpenetration between the gripper and the object. For each object, we initially sample a fixed number of single-arm grasps, which are then combined exhaustively to form candidate dual-arm grasp pairs. To improve diversity and avoid redundant configurations, a distance-based pruning step is applied to remove grasp pairs that are spatially too close. This results in a large set of candidate dual-arm grasps for each object.

4.3.2.2 Force-Closure Evaluation

For each object, we first generate candidate dual-arm grasp pairs by sampling 500 single-arm grasps and exhaustively combining them (that is, $\binom{500}{2}$). To improve diversity and avoid redundant configurations, a distance-based pruning step is applied, resulting in approximately $30,000$ to $80,000$ candidate grasp pairs per object.

[1] We use a constant $\mu = 0.4$ throughout our experiments.

These candidate grasps are then evaluated using the proposed optimization-based force-closure formulation (Section 4.3.1). For each grasp pair, the optimizer attempts to compute a set of contact forces that can balance an external wrench applied at the object's center of mass while satisfying friction cone and actuator constraints.

The quality of a grasp is quantified by the residual wrench, as defined in Equation 4.4. A grasp is considered stable if the optimizer is able to find a feasible force vectors that result in a zero residual (or near-zero), indicating that the grasp can maintain equilibrium under external disturbances. This step effectively filters out grasp pairs that are geometrically valid but physically unstable.

4.3.2.3 Grasp Selection and Dataset Construction

Based on the optimization outcome, grasp pairs are classified as stable or unstable using a threshold on the optimization loss. Only grasp pairs that satisfy both the rank condition (4.3) and the force-closure criterion (4.8) are retained, ensuring that the final dataset consists of grasps that are not only geometrically valid but also physically stable.

Applying this pipeline across a large set of objects results in a dataset comprising millions of dual-arm grasp pairs with reliable stability annotations. The use of optimization-based filtering leads to well-separated positive and negative grasp samples, which is crucial for training learning-based grasp evaluation and generation models.

To further validate the physical feasibility of the generated grasps, we construct a simulation-based benchmark using Isaac Gym [37]. The grasp pairs are evaluated using floating grippers initialized at the predicted grasp poses with gravity disabled. Once the grippers close fully [2], gravity is enabled, and the ability of the grasp to hold the object is assessed. A grasp is considered successful if the object remains stable without slipping or dropping under these conditions. This simulation pipeline provides an additional layer of validation beyond analytical force-closure, capturing dynamic effects and contact interactions that are not modeled in the optimization framework, thereby yielding more realistic and reliable grasp stability labels.

4.4 Experiments and Results

In this section, we evaluate the quality of the proposed DG16M dataset and its effectiveness in training dual-arm grasp classifiers. We compare against existing datasets and analyze performance using both analytical and simulation-based metrics.

[2] A force of 70N is applied, based on the data provided in Franka Emika Robot's Instruction Handbook

4.4.1 Baselines

To assess the quality of the generated grasp dataset, we benchmark learning-based grasp classifiers trained on DG16M against those trained on the DA^2 dataset [43]. We consider two representative baselines:

(i) Dual-PointNetGPD. We adapt PointNetGPD [91], a widely used grasp evaluation network, to the dual-arm setting following [43]. Given a dual-arm grasp configuration, the gripper templates are transformed to their respective grasp poses using a rigid transformation matrix $H \in SE(3)$. To capture local geometric context, the nearest 512 points from the object point cloud P_0 are sampled around each transformed gripper pose, forming two local point cloud regions P_1 and P_2. These regions correspond to the contact neighborhoods of the two grippers. The combined set $P = P_1 \cup P_2$ serves as a local representation of the dual-arm grasp. This combined point set is passed through the PointNetGPD architecture to extract a feature vector, which is then processed by a MLP to classify whether the grasp is successful or not.

(ii) CGDF-Classifier. We build upon recent $SE(3)$-aware grasp generation approaches [19, 26] by integrating a classification head on top of learned grasp descriptors derived from neural descriptor fields [71]. The network takes as input the object point cloud $P_o \in \mathbb{R}^{M\times 3}$, consisting of M points, and the gripper template point cloud $P_g \in \mathbb{R}^{N\times 3}$, consisting of N points. The gripper templates are transformed into the object coordinate frame using a rigid transformation $H \in SE(3)$, enabling the model to encode the relative pose between the gripper and the object. The transformed gripper representation, together with the object point cloud, is processed through a vision encoder and a feature encoder to obtain a joint feature representation. A classification head is then applied to these features to predict whether the given dual-arm grasp is successful. This formulation allows the model to jointly reason about object geometry and grasp configuration in a unified representation space.

4.4.2 Metrics

Evaluating dual-arm grasps requires assessing both analytical feasibility and practical execution. To this end, we use two complementary metrics:

(i) Force Closure Evaluation (FCE). FCE is an analytical metric that evaluates whether a grasp satisfies force closure conditions. Using the optimization-based formulation described in Section 4.3.1, we compute the optimal contact forces and verify whether the grasp can balance external wrenches while satisfying friction cone and actuator constraints. A grasp is considered successful if a feasible force solution exists that results in near-zero residual wrench.

(ii) Grasp Success Rate (GSR). GSR is a simulation-based metric that evaluates the practical stability of grasps. Following the benchmark setup, grasps are executed in the Isaac Gym simulator [37], where grippers are initialized at the grasp pose with gravity disabled. After closing the grippers, gravity is enabled, and the grasp is deemed successful if the object remains stable under these conditions. This metric captures real-world effects such as contact dynamics and unmodeled interactions that are not accounted for in analytical evaluation.

Model	DG16M		DA^2 Dataset	
	FCE(%)↑	GSR(%)↑	FCE(%)↑	GSR(%)↑
Dual-PointNetGPD [91]	69.5	61.22	61.8	53.26
CGDF-Classifier [19]	**88.73**	**76.34**	**65.25**	**54.33**

Table 4.2: Performance comparison of different models on our dataset and DA^2 dataset.

Together, FCE and GSR provide a comprehensive evaluation framework: FCE measures theoretical feasibility under physical constraints, while GSR evaluates robustness under dynamic execution. The gap between these metrics further highlights the challenges in translating analytical grasp stability into practical performance.

4.4.3 Results

Quantitative Results Table 4.2 presents the quantiative performance of Dual-PointNetGPD and CGDF-Classifier trained on DG16M and DA^2, evaluated using Force Closure Evaluation (FCE) and Grasp Success Rate (GSR). Models trained on DG16M consistently outperform those trained on DA^2 across both metrics. In particular, CGDF-Classifier achieves 88.73% FCE and 76.34% GSR on DG16M, compared to 65.25% FCE and 54.33% GSR on DA^2. A similar trend is observed for Dual-PointNetGPD, indicating that the improvements are not model-specific but stem from the quality of the dataset.

A key observation is that models trained on DA^2 exhibit near-random performance in simulation, as reflected by GSR values close to 50%. This suggests that these models struggle to reliably distinguish between successful and failed grasps. Although DA^2 employs a structured grasp evaluation criterion, its grasp distribution does not translate effectively to simulation-based evaluation, limiting generalization.

In contrast, models trained on DG16M achieve significantly higher performance in both FCE and GSR, demonstrating improved theoretical feasibility as well as practical stability. This indicates that the optimization-based force-closure formulation used in DG16M results in better-separated positive and negative grasp samples, enabling more effective learning.

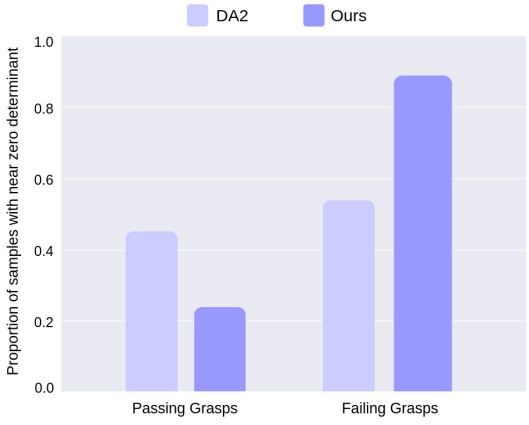

Figure 4.4: Grasp Stability Analysis with $Q(G) = \sqrt{\det(GG^T)}$.

We also observe a gap between FCE and GSR across all models, which highlights the challenge of translating analytical grasp stability into practical execution. This discrepancy arises due to factors such as contact dynamics, unmodeled physical effects, and gripper-object interactions in simulation.

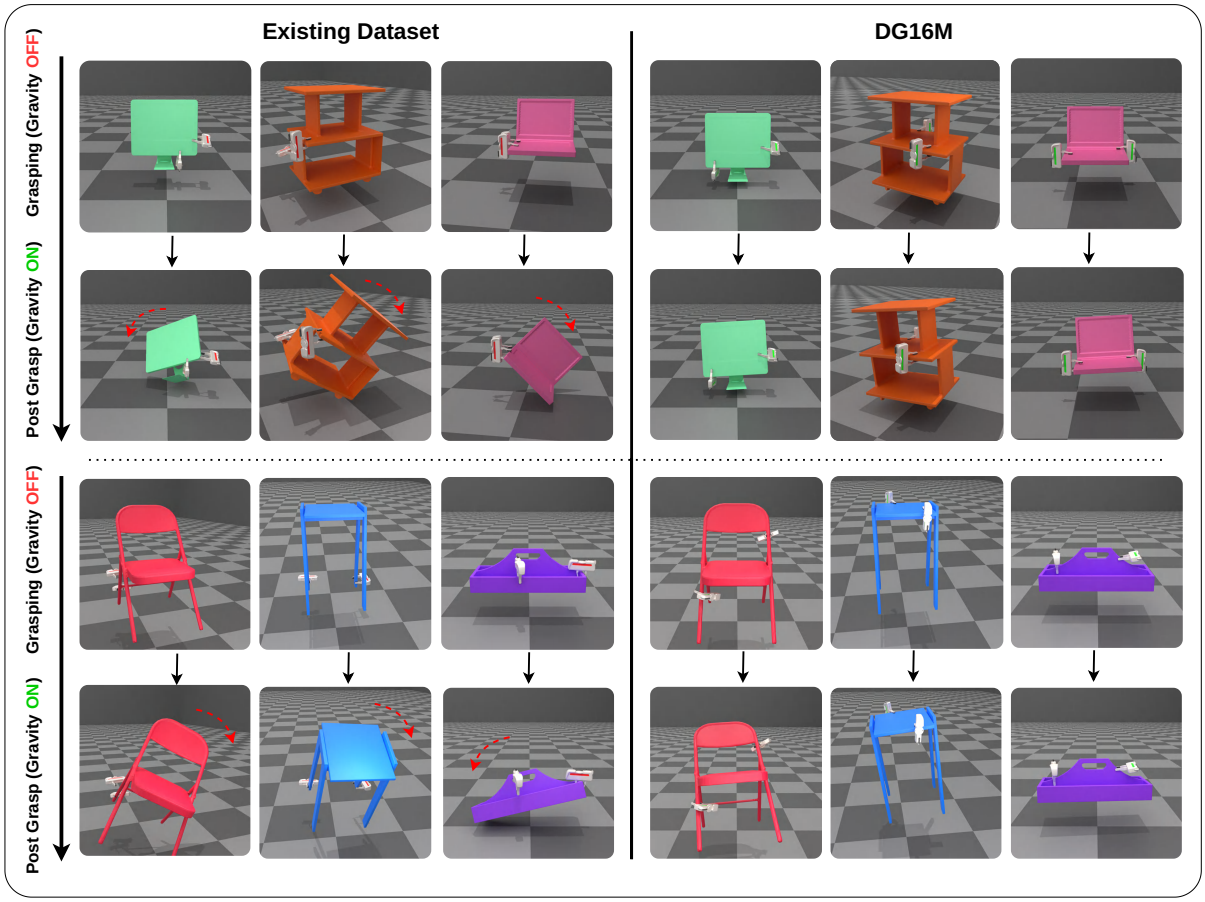

Figure 4.5: Qualitative comparison between DA^2 and DG16M. Grasps are evaluated in simulation using floating grippers (white), initialized with gravity off and tested after enabling gravity. DA^2 grasps often fail due to poor force balance or placement, while DG16M produces stable and physically consistent grasps.

To further analyze dataset quality, we evaluate grasp stability using the quality metric $Q(G) = \sqrt{\det(GG^T)}$ [92], where G is the grasp matrix. This metric quantifies the *volume* of the grasp wrench space, reflecting the ability of a grasp to generate stabilizing wrenches. A higher value of $Q(G)$ indicates a robust and well-conditioned grasp, while near-zero values correspond to unstable or near-singular configurations. As shown in Figure 4.4, DG16M exhibits a clear separation between successful and failed grasps, with most failures concentrated near zero and successful grasps having significantly higher values. In contrast, DA^2 shows poor separation, with both successful and failed grasps exhibiting similar near-zero values. This highlights that DG16M provides more physically meaningful and discriminative grasp annotations, which directly contributes to improved learning performance.

Qualitative Results Figure 4.5 presents a qualitative comparison between grasps generated by models trained on DG16M and those trained on the DA^2 dataset. For large objects with complex geometries, models trained on DA^2 frequently predict grasp configurations that fail to satisfy force-

Generation Technique	Evaluation Metrics	
	FCE(%) ↑	GSR(%) ↑
Random Dual-Arm Pairs	16.33	23.66
Farthest-Grasp Pairs	26.16	38.45
DG16M (Ours)	**100.0**	**76.33**

Table 4.3: Ablation study comparing different dual-arm grasp generation techniques.

closure or dynamic stability conditions. As shown in the figure, objects such as chairs, monitors, and irregular structures often slip, rotate, or lose contact once gravity is enabled. These failures arise due to grasp pairs being too close together, positioned on the same side of the object's center of mass, or lacking proper force balance. In contrast, models trained on DG16M consistently produce more stable and well-distributed grasp configurations. For instance, grasps on objects like chairs and tables are placed across structurally supportive regions, enabling balanced force application. As a result, the objects remain stable even after gravity is enabled, indicating successful grasp execution.

Overall, these qualitative results show that the improved force-closure formulation in DG16M leads to physically meaningful grasp predictions that translate effectively from analytical evaluation to practical stability.

4.4.4 Ablations

To evaluate the effectiveness of the proposed optimization-based force-closure formulation, we conduct ablation studies by comparing our grasp selection strategy with alternative dual-arm grasp pairing methods.

(i) Random Pairing. In this setting, single-arm grasps generated via antipodal sampling [22] are randomly paired to form dual-arm grasps. This naive strategy yields poor performance, achieving only 16.33% FCE and 23.66% GSR. The resulting grasp pairs often lack proper force balance and fail to satisfy stability constraints, leading to frequent failures in both analytical and simulation evaluations.

(ii) Farthest-Grasp Pairing. Here, single-arm grasps are paired based on maximum spatial separation, ensuring that grasp pairs are geometrically opposite. While this improves performance to 26.16% FCE and 38.45% GSR, it still fails to ensure stability. This method assumes that geometric opposition implies stability, but does not account for force and torque equilibrium, resulting in many unstable grasps.

(iii) Force-Closure-Based Pairing (Ours). Our method explicitly enforces force-closure constraints during grasp selection, achieving 100% FCE and 76.33% GSR. By ensuring that grasp pairs satisfy both geometric and physical constraints, this approach produces grasps that are not only theoretically valid but also robust in simulation.

These results demonstrate that antipodal sampling alone is insufficient for generating stable dual-arm grasps. While geometric heuristics such as spatial separation provide marginal improvements, incorporating physical constraints through force-closure optimization is essential for achieving reliable and stable grasp configurations.

4.5 Chapter Summary

This chapter focuses on the problem of evaluating dual-arm grasp stability, where we introduce a physically grounded framework based on an optimization-based force closure formulation that explicitly models contact forces, friction, and actuator constraints. Building on this, we construct **DG16M**, a large-scale dataset of dual-arm grasps by systematically sampling candidate grasp pairs and filtering them using the proposed optimizer, resulting in physically consistent and well-separated stability labels. Our experiments show that DG16M significantly improves both analytical feasibility and simulation performance compared to prior datasets, highlighting the importance of incorporating physical constraints rather than relying on geometric heuristics. We further observe that common strategies such as random or farthest-grasp pairing are insufficient for generating stable dual-arm grasps, reinforcing the need for principled formulations that jointly consider geometry and force interactions. This naturally motivates the next chapter, where we move beyond evaluation and leverage both the generative modeling framework of *Chapter 3* and the analytical insights from this chapter to directly generate stable dual-arm grasps on object point clouds.

Chapter 5

Learning to Generate Dual-Arm Grasps

Deep learning has given us machines with truly impressive abilities but no intelligence. The difference is profound and lies in the absence of a model of reality.

JUDEA PEARL in *The Book of Why*, 2018

5.1 Introduction

Generating dual-arm grasps requires addressing two fundamental challenges. First, the system must be capable of generating grasps on large and geometrically complex objects, where single-arm strategies are often insufficient. Second, it must understand whether a *pair* of grasps can jointly stabilize the weight of the object (or external disturbances). While these two aspects — grasp generation and grasp stability — are individually well-studied, integrating them into a unified framework remains a challenging problem, raising the question of how to jointly model grasp generation and physical stability.

In *Chapter* 3, we introduced CGDF [19], a learning-based framework for generating dense and high-quality grasps on large objects with complex geometries. This provided a scalable solution for exploring the grasp space, particularly in scenarios where multiple grasp configurations are required. In *Chapter* 4, we addressed the complementary problem of evaluating dual-arm grasp stability through an optimization-based framework that rigorously models force closure and physical feasibility of grasp pairs and realsed the DG16M [20] dataset. Together, these contributions establish the two key ingredients necessary for dual-arm manipulation — the ability to *generate grasps*, and the ability to *assess dual-arm grasp stability*.

However, these components remain disjoint. CGDF generates grasps without explicit awareness of how two grasps interact, while DG16M evaluates stability only after candidate grasp pairs are formed, and operates on object meshes since it requires access to detailed geometric information that is typically unavailable in real-world perception settings such as RGB-D observations. As a result, forming stable dual-arm grasps requires additional heuristics, such as pairing grasps from distant

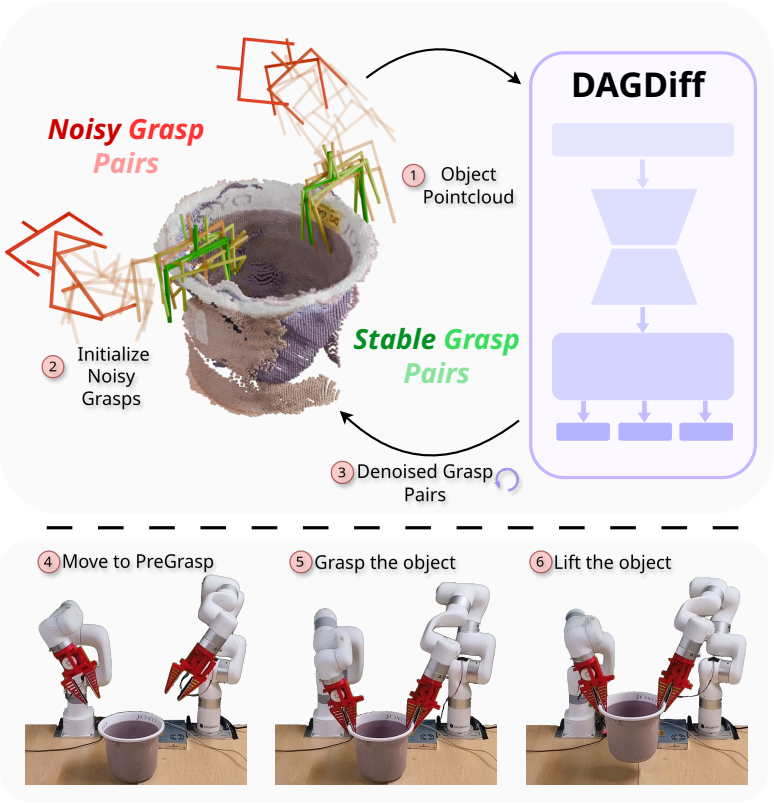

Figure 5.1: We introduce DAGDiff, a diffusion-based framework in $SE(3) \times SE(3)$ that transforms noisy dual-arm grasp pairs into stable, collision-free configurations conditioned on object point clouds. Guided by multi-head outputs, the model generates physically valid grasps that transfer effectively to real-world dual-arm execution.

regions or relying on semantic cues, which do not guarantee physical consistency. This separation leads to inefficiencies and often produces grasp pairs that are geometrically valid but physically unstable.

This limitation is also reflected in the broader class of dual-arm grasping approaches, which largely treat the problem as a combination of two independent single-arm grasps. In CGDF, for instance, we explore on using the part-guided grasp generation for a dual-arm setting by sampling grasps in spatially distant regions and pairing them heuristically. Although this provides a simple and scalable strategy, it does not guaranty joint stability. The key challenge in such approaches is region selection: heuristic strategies such as choosing farthest regions can fail when those regions are physically incompatible for stable grasping. To address this, methods like UniDiffGrasp [61] leverage Vision-Language Models (VLMs) to identify graspable object parts. However, VLM-based reasoning remains limited in capturing 3D geometry and physical interactions [93–97], and is further constrained by the fact that graspable regions rarely have clear semantic grounding. Consequently, these approaches rely heavily on region priors or semantic cues without explicitly enforcing physical constraints between the two grippers, often producing grasp pairs that are physically inconsistent, unstable under external disturbances, or prone to collisions.

In this chapter, we introduce **DAGDiff: Dual-Arm Grasp Diffusion**, an end-to-end learning-based framework for directly generating dual-arm grasps in a unified and physically grounded manner (shown in Figure 5.1). Instead of decomposing the problem into independent sub-tasks, we formulate it as a joint generative process over the space of grasp pairs. Specifically, we build upon diffusion models and extend them to operate in the combined pose space $SE(3) \times SE(3)$, enabling the model to reason about both grippers simultaneously. To ensure physical validity, we incorporate *guidance signals* that encode key constraints such as force-closure stability and collision avoidance. These signals steer the generative process toward regions of the grasp space that are not only geometrically feasible but also physically stable.

A key insight of this work is that *stability and feasibility can be learned implicitly through guidance*, rather than enforced through explicit heuristics or post-processing. By integrating stability-aware and geometry-aware signals into the generative process, the model learns to discover suitable grasp regions and coordinated configurations directly from data. This removes the need for predefined region selection, semantic labeling, or exhaustive pairwise evaluation, leading to a more general and scalable solution.

Through this formulation, dual-arm grasp generation becomes a unified learning problem that captures both geometric and physical aspects of manipulation. The resulting framework enables the synthesis of diverse, stable, and collision-free grasp pairs on previously unseen objects, bridging the gap between grasp generation and execution in realistic dual-arm robotic systems. The project page and code can be accessed here: https://dag-diff.github.io/dagdiff/.

5.2 Related Works

5.2.1 Dual-Arm Grasping and Stability

Dual-arm grasping requires two grasps that are not only individually stable but also jointly satisfy stability criteria such as force-closure, where the combined contact forces can resist arbitrary external wrenches [1, 10]. Mesh-based approaches address this by densely sampling candidate single-arm grasps on object meshes and evaluating all possible grasp pairs using analytical force-closure metrics [20, 43]. While such methods provide strong theoretical guarantees, they rely on complete object geometry and exhaustive pairwise evaluation, making them computationally expensive and difficult to deploy in real-world settings.

To improve scalability, several works approximate dual-arm grasping by combining independently generated single-arm grasps. Our part-guided strategy, as introduced in CGDF [19], generates grasps in spatially distant regions of the object and forms dual-arm pairs heuristically. As discussed in *Chapter* 3, this provides a simple and effective mechanism to reduce the combinatorial search space. However, the lack of explicit modeling of interactions between the two grippers means that joint stability is not guaranteed, and the resulting grasp pairs may be physically incompatible. Similarly, VLM-based approaches such as UniDiffGrasp [61] attempt to identify graspable regions using semantic

reasoning. However, these methods depend heavily on region priors and are limited by the lack of precise 3D and physical understanding in vision-language models [93, 94], often leading to coarse or unstable grasp configurations.

An alternative direction is to explicitly model inter-gripper dependencies. DualAfford [44] formulates dual-arm grasping as a collaborative affordance prediction problem, where one gripper's pose is conditioned on the other. While this captures coordination between the arms, it typically requires object category-specific training and complex pipelines, limiting generalization to unseen objects. In contrast, our approach models dual-arm grasp generation as a unified process, directly learning the distribution of stable grasp pairs without relying on region heuristics or category-specific priors.

5.3 DAGDiff: Dual-Arm Grasp Diffusion

In this chapter, we present **DAGDiff**, a learning-based framework for directly generating dual-arm grasp pairs that are both stable and collision-free. Given an object point cloud $P \in \mathbb{R}^{n\times 3}$, our goal is to generate a set of M (user-defined) dual-arm grasp poses $H_i = (H_{i,1}, H_{i,2}) \in SE(3) \times SE(3), i \in [1, M]$, such that (i) each grasp establishes valid contact with the object and (ii) the pair jointly satisfies dual-arm force-closure constraints (as given in Equation 4.4). As highlighted in the DAGDiff framework (Figure 5.3), this requires reasoning over both geometric feasibility and physical stability.

To address this, we formulate dual-arm grasp generation as a diffusion process in the joint pose space of $SE(3) \times SE(3)$. Starting from random grasp pairs, the network iteratively refines them toward physically valid[1] configurations using a learned score function. In addition, we incorporate guidance signals that encode stability and collision constraints, enabling the model to generate coordinated grasp pairs without relying on explicit region selection.

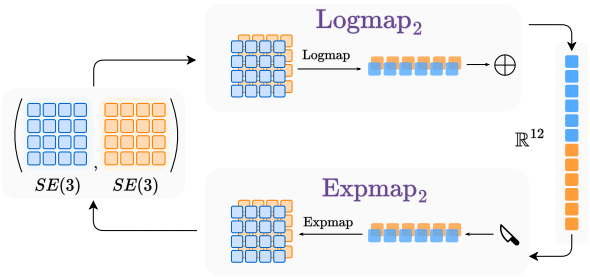

Figure 5.2: Mapping between dual-arm pose space and Euclidean space ($SE(3) \times SE(3) \leftrightarrow \mathbb{R}^{12}$).

To enable diffusion in the product Lie group $SE(3) \times SE(3)$, we first define mappings between the manifold and a Euclidean space (as shown in Figure 5.2). Let the dual-arm logarithmic map (Logmap_2) be defined as:

$$\mathbf{v} = \text{Logmap}_2(H) := \text{Logmap}(H_1) \oplus \text{Logmap}(H_2) \in \mathbb{R}^{12}$$

[1] By physically valid, we mean grasp poses that make proper contact with the object surface, rather than floating in free space.

where $\oplus$ denotes concatenation. Similarly, the dual-arm exponential map (Expmap_2) is defined as:

$$H = \text{Expmap}_2(\mathbf{v}) := \big(\text{Expmap}(\mathbf{v}_{[:6]}),\ \text{Expmap}(\mathbf{v}_{[6:]})\big)$$

where $\mathbf{v}_{[:6]}$) and $\mathbf{v}_{[6:]}$) refer to the first and the last six components of the vector $\mathbf{v} \in \mathbb{R}^{12}$.

This allows us to perform diffusion in $\mathbb{R}^{12}$ while maintaining consistency with the underlying pose manifold. In the following sections, we first describe the diffusion-based formulation for dual-arm grasp generation, followed by the guidance mechanisms and model architecture.

In the following sections, we first describe the diffusion-based formulation for joint $SE(3) \times SE(3)$ grasp generation (Section 5.3.1), followed by the incorporation of classifier-guidance modules that enforce stability and collision constraints (Section 5.3.2). Finally, we detail the overall model architecture and implementation of DAGDiff (Section 5.3.3).

5.3.1 Diffusion-based Dual-Arm Grasp Generation

We extend the $SE(3)$ diffusion framework proposed in $SE(3)$-DiF [26] to the dual-arm setting. A dual-arm grasp pose $H \in SE(3) \times SE(3)$ is first mapped to a Euclidean representation $\mathbf{v} \in \mathbb{R}^{12}$ using the logarithmic map (Logmap_2), enabling standard diffusion operations.

Energy-based formulation. Instead of directly predicting the added noise, we adopt a score-based formulation [74], where the model learns the gradient of the log-density of grasp poses. This formulation has been shown to provide more stable training and improved likelihood modeling, as the score function directly characterizes the structure of the underlying data distribution at different noise levels. Specifically, we define an energy-based model $E_\alpha \colon SE(3) \times SE(3) \times \mathbb{R}^{n\times 3} \times \mathbb{R} \to \mathbb{R}$ which assigns a scalar *energy* to a *grasp pair* conditioned on the object *point cloud* and diffusion *timestep*. The corresponding score function s_α is given by:

$$s_\alpha(H, P, t) = -\nabla_H E_\alpha(H, P, t) \in \mathbb{R}^{12} \tag{5.1}$$

This formulation enables the model to learn an energy landscape over grasp pairs, where lower energy corresponds to physically valid and stable configurations.

Forward and Reverse Diffusion. During training, ground truth grasp pairs are progressively perturbed by adding Gaussian noise in the Euclidean space:

$$\tilde{H}_t = \text{Expmap}_2\left(\text{Logmap}_2(H) + \epsilon_t\right) \tag{5.2}$$

where $\epsilon_t \sim \mathcal{N}(0, \sigma_t^2 \mathbf{I}_{12})$ and σ_t is the standard deviation at timestep t. This forward process generates noisy grasp samples at different timesteps, which are used to train the model to recover clean grasps.

At inference, grasp pairs are initialized randomly and iteratively refined using a Langevin-style reverse diffusion update:

$$H_{t-1} = \text{Expmap}_2 \left(\frac{\eta_t^2}{2} s_\alpha(H_t, P, t) + \eta_t \epsilon \right) H_t \tag{5.3}$$

where $\epsilon \sim \mathcal{N}(0, \mathbf{I}_{12})$ and $\eta_t \geq 0$ controls the step size. This reverse process gradually removes noise, guiding samples toward low-energy regions corresponding to stable and physically feasible dual-arm grasps.

Training objective. The model is trained using a denoising objective that matches the predicted score with the normalized noise:

$$\mathcal{L}_{\text{diff}} = \left\| s_\alpha(H_t, P, t) - \frac{\epsilon_t}{\sigma_t} \right\|_1 \tag{5.4}$$

This encourages the model to accurately recover clean grasp poses from noisy inputs and learn the underlying distribution of valid dual-arm grasps.

5.3.2 Classifier-Guidance Modules

Naively training a diffusion model for dual-arm grasp generation does not guarantee physically valid solutions, as there is no explicit mechanism enforcing joint stability or collision avoidance. To address this, we adopt *classifier-guided diffusion* [98], which steers the generative process toward desired regions of the sample space by incorporating gradients of auxiliary classifiers during the reverse diffusion process.

We introduce two classifier-guidance modules that provide complementary physical constraints: (i) a *force-closure classifier* for stability and (ii) a *collision classifier* for geometric validity. As illustrated in the method overview (Figure 5.3), these modules operate jointly with the energy-based diffusion model to guide grasp generation.

Force-Closure Guidance. We define a force-closure classifier $C_\beta^{\text{fc}}(H, P) = p(y = 1 \mid H, P; \beta)$, with learnable parameters β, which predicts the probability that a dual-arm grasp pair satisfies the force-closure stability criterion. This classifier is trained to distinguish stable and unstable grasp configurations using supervision derived from analytical force-closure evaluation.

During inference, the gradient of the log-probability with respect to the grasp pose, given by, $\nabla_H \log C_\beta^{\text{fc}}(H, P)$, is used to guide the diffusion process toward regions of the pose space that satisfy dual-arm stability. Intuitively, this encourages the generated grasp pairs to achieve coordinated force balance rather than treating the two grasps independently.

Collision Guidance. To ensure geometric feasibility, we introduce a collision classifier, $C_\gamma^{\text{col}}(H, P) = p(y = 1 \mid H, P; \gamma)$, with learnable parameters γ, which predicts whether either gripper intersects

the object geometry. Unlike force-closure, collision is a local geometric property that depends on precise alignment with the object surface.

During inference, the gradient $\nabla_H \log\left(1 - C_\gamma^{\text{col}}(H, P)\right)$ is used to push grasp candidates away from collision regions. Importantly, collision guidance is activated only after an initial denoising phase, since early-stage samples are typically far from the object surface and do not require refinement. This staged guidance improves stability of the diffusion process while enabling fine-grained geometric corrections in later steps.

Training Objective. Both classifiers are trained using binary cross-entropy loss:

$$\mathcal{L}_{\text{fc}} = \text{BCE}\big(C_\beta^{\text{FC}}(H, P),\, y_{\text{fc}}\big) \tag{5.5}$$

$$\mathcal{L}_{\text{col}} = \text{BCE}\Big(C_\gamma^{\text{Col}}(H, P),\, y_{\text{col}}\Big) \tag{5.6}$$

where $y_{\text{fc}} \in \{0, 1\}$ indicates whether the grasp pair satisfies force closure, and $y_{\text{col}} \in \{0, 1\} \times \{0, 1\}$ encodes collision labels for the grasp pair.

Finally, the overall training objective combines the diffusion loss with classifier supervision:

$$\mathcal{L} = \mathcal{L}_{\text{diff}} + \mathcal{L}_{\text{fc}} + \mathcal{L}_{\text{col}} \tag{5.7}$$

Guided Score Function. At inference time, the classifier gradients are combined with the base diffusion score to guide sampling:

$$\tilde{s}(H, P, t) = s_\alpha(H, P, t) + \nabla_H \log C_\beta^{\text{fc}}(H, P) + \begin{cases} 0, & t < t_c \\ \nabla_H \log\left(1 - C_\gamma^{\text{col}}(H, P)\right), & t \geq t_c \end{cases} \tag{5.8}$$

where t_c is a predefined threshold controlling when collision refinement is supposed to be activated.

This formulation allows the diffusion process to jointly optimize for *feasibility*, *stability*, and *collision avoidance*, resulting in grasp pairs that are not only low-energy but also physically valid. Unlike prior approaches that rely on explicit region selection or post-hoc filtering, the proposed guidance mechanism directly shapes the generative dynamics, enabling the model to discover suitable grasp configurations in a unified manner.

5.3.3 DAGDiff Architecture

Our architecture (Figure 5.3) extends diffusion-based grasp generation to the dual-arm setting by combining a geometry-aware feature encoder (introduced in Chapter 3) with multi-head outputs that jointly enforce feasibility, stability, and collision avoidance.

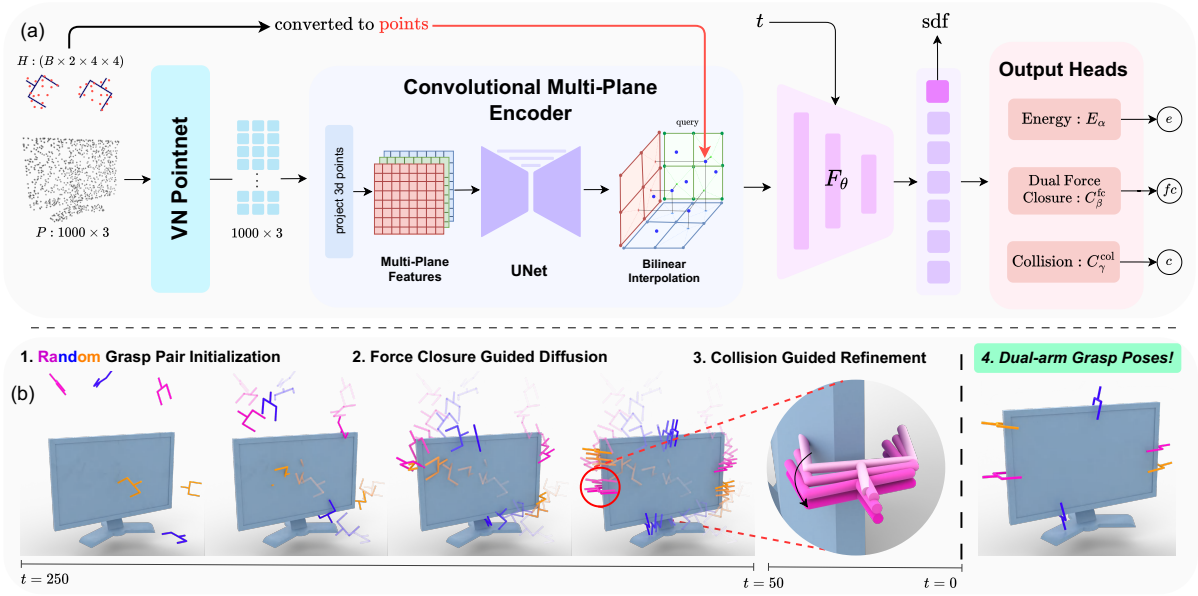

Figure 5.3: Overview of DAGDiff: (a) The input point cloud P is encoded into dense geometric features. Randomly initialized dual-arm grasps H define query points for feature sampling, which are processed by F_θ (conditioned on timestep t) to predict SDF and latent features. These are used by three heads to estimate energy (E_α), force-closure (C_β^{fc}), and collision (C_γ^{col}) signals that guide diffusion. **(b)** Starting from noisy grasp pairs ($t = 250$), the model iteratively denoises to stable, collision-free grasps ($t = 0$), driven by energy and refined by stability and collision guidance.

Feature Encoding. Inspired from our CGDF architecture (*Chapter* 3), given an input point cloud $P \in \mathbb{R}^{n\times 3}$, we first extract geometric features using a VN-PointNet encoder [77], which provides $SO(3)$-equivariant per-point representations. These features are then projected onto multiple canonical planes and processed using a convolutional UNet backbone [70], resulting in dense multi-plane feature maps that capture both local and global geometric context.

To incorporate grasp-specific information, a fixed query point cloud $P_q \in \mathbb{R}^{30\times 3}$ is transformed using the dual-arm grasp pose H, producing a set of query points P_H that represent the local region around the grippers. The projected locations of these query points are then used to sample features from the multi-plane representations via bilinear interpolation. The sampled features are aggregated and passed through a feature encoder F_θ, which is conditioned on the diffusion timestep t. In addition to producing a latent feature vector, F_θ also predicts the signed distance function (SDF) values of the query points, providing auxiliary geometric supervision during training. As shown in Figure 5.3 this pipeline enables the network to reason about fine-grained geometry in the vicinity of the grasp.

Multi-Head Outputs. The learned feature representation is passed through three specialized output heads, each responsible for a distinct aspect of grasp generation:

1. **Energy Head:** Outputs a scalar energy $E_\alpha(H, P, t) \in \mathbb{R}$, which defines the score function used in diffusion. Lower energy corresponds to grasp pairs that are physically grounded on the ob-

ject and closer to valid configurations. This head forms the backbone of the generative process, guiding denoising toward feasible regions of the grasp space.

2. **Force-Closure Head:** Predicts the probability $C_{\beta}^{\text{fc}}(H, P)$ that a grasp pair satisfies dual-arm force-closure constraints. Trained jointly with the energy head, it ensures that the learned representation encodes physically meaningful stability cues. During inference, its gradients act as guidance signals that bias sampling toward stable grasp configurations.

3. **Collision Head:** Predicts the probability $C_{\gamma}^{\text{col}}(H, P)$ of collision between the grippers and the object. Importantly, this head is trained after the main network has converged, thereby allowing it to specialize in detecting fine-grained geometric collisions without impacting the base generative dynamics. During inference, its gradients are activated in later diffusion stages to refine grasp poses and eliminate collisions.

The three heads operate in a complementary manner. The *Energy Head* drives the generative dynamics and captures the overall structure of valid grasps, the *Force-Closure Head* enforces joint stability between the two arms, and the *Collision Head* ensures geometric feasibility. This coordinated design allows the model to directly generate dual-arm grasp pairs that are both stable and collision-free, without relying on explicit region proposals or post-hoc filtering.

5.4 Experiments

In this section, we evaluate the proposed DAGDiff framework for dual-arm grasp generation. We describe the dataset used for training and evaluation, the baselines considered for comparison, and the metrics used to assess performance. For each run, we initialize a batch of B random dual-arm grasp poses and progressively refine them through a denoising process over $T = 250$ timesteps. The last 50 timesteps are dedicated to collision refinement, since by $T = 200$ most grasps have settled near valid regions of the object, although some may still be in collision.

5.4.1 Dataset

We conduct our experiments on the DG16M dataset [20], which was introduced in Chapter 4. DG16M is a large-scale dataset specifically designed for dual-arm grasping and contains 4,143 objects with approximately 2,000 grasp pairs per object, annotated with both positive and negative labels based on force-closure evaluation. For our experiments, we adopt a random train-test split, where 400 objects are reserved for testing and the remaining objects are used for training. All quantitative results are reported on this unseen test set to evaluate generalization.

During inference, object meshes are converted into point clouds by uniformly sampling 1000 points per object, which serve as input to the model. Additionally, to train the collision classifier, we construct a synthetic dataset of colliding and non-colliding grasp pairs by sampling grasp poses and checking for intersections with the object geometry.

5.4.2 Baselines

To evaluate the effectiveness of DAGDiff, we compare against representative methods across three categories of dual-arm grasping approaches:

1. Farthest-region based methods. These methods reduce the combinatorial complexity of dual-arm grasping by selecting spatially distant regions on the object. **CGDF** [19] generates grasps in the two farthest regions of the point cloud and combines them into grasp pairs. **VCGS** [66] generates constrained single-arm grasps using a variational model and is adapted to the dual-arm setting following [19].

2. VLM-guided methods. These approaches leverage vision-language models (VLMs) to identify graspable regions. **UniDiffGrasp** [61] uses GPT-4V [99] and VL-Part [100] to segment object parts and applies diffusion-based grasp generation within those regions. Additionally, we evaluate variants based on RoboBrain 2.0 [101], which predict grasp regions using bounding boxes (**RoboBrainGrasp-BB**) and keypoints (**RoboBrainGrasp-KP**), respectively.

3. Affordance-based methods. **DualAfford** [44] formulates dual-arm grasping as a collaborative affordance prediction problem, where the grasp of one arm is conditioned on the other. Due to its category-specific training pipeline, we evaluate it in a zero-shot manner on its test objects, following the same protocol used in the above baselines.

5.4.3 Metrics

We evaluate all methods using three complementary metrics that capture analytical stability, physical robustness, and geometric feasibility:

1. Force Closure Evaluation (FCE). FCE measures whether a grasp pair satisfies the dual-arm force-closure condition, i.e., whether the contact forces can resist arbitrary external wrenches under friction constraints. Following the formulation in DG16M [20], this metric provides a theoretical guarantee of grasp stability.

2. Grasp Success Rate (GSR). GSR evaluates the physical robustness of generated grasps in simulation using Isaac Gym [37]. Each grasp is executed by initializing the grippers at the predicted poses, closing the fingers, and enabling gravity. A grasp is considered successful if the object is lifted and remains stably grasped. This metric reflects real-world execution feasibility.

3. Grasp Collision Rate (GCR). GCR measures the percentage of grasp pairs that result in collisions with the object geometry. Lower values indicate better geometric validity and effective collision avoidance during generation.

Together, these metrics provide a comprehensive evaluation of dual-arm grasp quality by jointly assessing stability, executability, and collision-free interaction.

5.4.4 Real-World Dual-Arm Setup

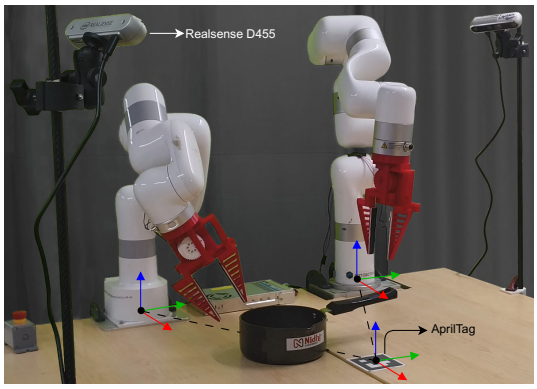

Figure 5.4: Real-world setup. A dual-arm system comprising an *XArm7* and *XArm6 Lite*, calibrated via an AprilTag and uses two RealSense D455 cameras for perception.

We also validate DAGDiff on a heterogeneous dual-arm setup (Figure 5.4) consisting of an xArm7 and an xArm6 Lite. The scene is observed using two Intel RealSense D455 RGB-D cameras, and the captured point clouds are fused (via ICP) to obtain a unified representation [102]. An AprilTag-based calibration [103] is used to align the camera and robot coordinate frames.

Given the reconstructed point cloud P (with 1000 points extracted), DAGDiff generates candidate dual-arm grasp pairs, which are filtered to retain only kinematically feasible configurations. Each grasp is executed by moving both arms to the predicted poses, closing the grippers, and lifting the object. A trial is considered successful if the object is lifted and remains stably grasped. We evaluate on diverse, previously unseen objects such as trays, buckets, pans, and drones. The results (in Table 5.3) demonstrate reliable execution and strong sim-to-real generalization, highlighting the robustness of DAGDiff under real-world sensing and actuation constraints.

5.4.5 Results

Qualitative Resullts We qualitatively evaluate the performance of DAGDiff by comparing the generated dual-arm grasps with prior methods across a diverse set of objects, as shown in Figure 5.5. These comparisons highlight the ability of our method to generate physically stable and collision-free grasp pairs without relying on heuristic region selection or semantic priors.

Methods based on farthest-region heuristics, such as CGDF [19] and VCGS [66], often fail to select physically compatible grasp regions. While these approaches ensure spatial separation between the two grippers, they do not account for force balance or feasibility. As a result, they frequently generate grasp pairs where one gripper is placed in unstable or unreachable regions. For example, in objects such as buckets, one of the grasps is often forced toward geometrically unsuitable regions (e.g., corners or edges), leading to poor contact and unstable configurations. Even in cases where both grasps lie on valid surfaces, the resulting configurations often require unbalanced or excessive contact forces, making them physically unreliable.

Vision-language model (VLM) based approaches, such as UniDiffGrasp [61] and RoboBrainGrasp [101], attempt to address region selection by predicting semantically meaningful object parts. However, these methods suffer from limitations in spatial precision and physical reasoning. As observed in Figure 5.5 (c) and (d), predicted keypoints or bounding boxes are often coarse, misaligned, or even located in free space. This leads to grasp proposals that either fail to establish proper contact or result

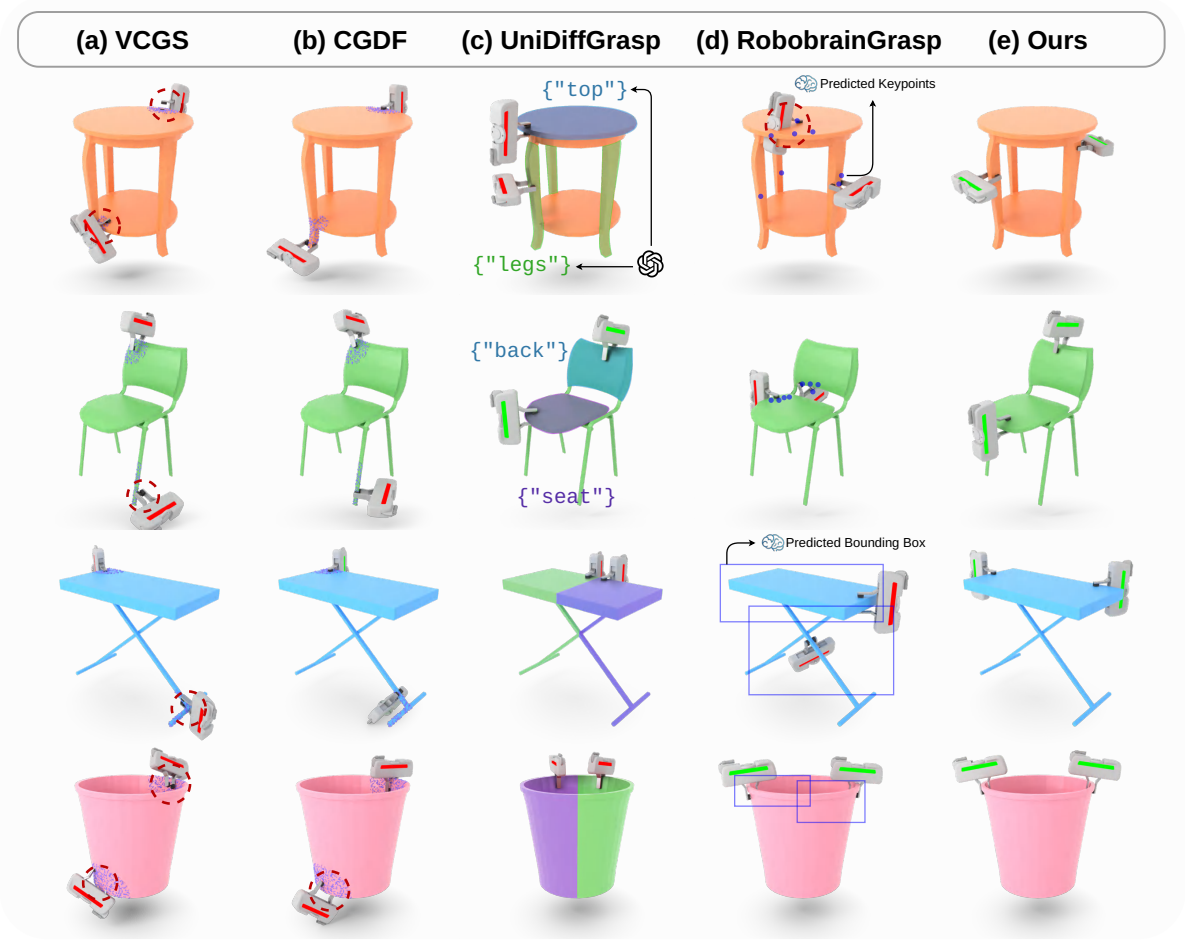

Figure 5.5: Qualitative comparison of dual-arm grasps. VCGS and CGDF rely on farthest-region heuristics, often leading to physically infeasible grasps (e.g., unreachable regions in the bucket). UniDiffGrasp depends on semantic segmentation, but ambiguous labels result in unstable region selection. RoboBrainGrasp predicts keypoints or bounding boxes, which are often misaligned or too coarse for accurate grasping. In contrast, our method directly generates stable and collision-free grasp pairs by reasoning over physical constraints, without relying on heuristics or semantic cues. (*Red circles indicate invalid grasps. RoboBrainGrasp-KP and -BB are collectively referred to as RoboBrainGrasp.*)

in unstable configurations. Additionally, when semantic cues are ambiguous or unavailable, these methods tend to default to arbitrary region splits, which further degrades grasp quality [2].

In contrast, DAGDiff directly generates dual-arm grasp pairs by reasoning over geometric and physical constraints during the diffusion process. By incorporating force-closure and collision-aware guidance, the model learns to identify compatible grasp regions implicitly, without relying on external heuristics or semantic annotations. As a result, the generated grasps exhibit consistent alignment with object geometry, balanced placement across both arms, and minimal collision with the object surface.

[2]This is by construction in [61] and not an output of the VLMs used in these methods.

Method	FCE(%)↑	GSR(%)↑	GCR(%)↓
DAGDiff (ours)	**60.14**	**72.50**	**15.10**
CGDF [19]	35.14	56.25	30.55
VCGS [66]	16.85	23.36	74.73
UniDiffGrasp [61]	10.10	31.68	59.90
RoboBrainGrasp-KP [101]	9.80	27.85	66.30
RoboBrainGrasp-BB [101]	7.12	27.81	70.26
Ours-DA†	**56.45**	**68.80**	**18.59**
DualAfford†† [44]	–	54.33	–

†Evaluated on Dual-Afford objects in a zero-shot setting.

††Values reported directly from the DualAfford paper.

Table 5.1: Comparison of dual-arm grasp generation methods. Higher values are better for Force Closure Error (FCE), which reflects grasp stability, and Grasp Success Rate (GSR), which measures execution success. Lower values are better for Grasp Collision Rate (GCR), which indicates the frequency of collisions.

Across all evaluated scenarios, DAGDiff produces grasp pairs that are both physically stable and geometrically valid. The grasps are well-distributed across structurally meaningful regions of the object, enabling coordinated dual-arm manipulation. These qualitative results demonstrate that explicitly incorporating physical constraints into the generative process is critical for reliable dual-arm grasp synthesis, particularly for large and complex objects.

Quantitative Results We quantitatively evaluate DAGDiff against existing dual-arm grasp generation methods using three metrics: Force Closure Evaluation (FCE), Grasp Success Rate (GSR), and Grasp Collision Rate (GCR). These metrics jointly capture analytical stability, physical robustness in simulation, and geometric validity of the generated grasps.

Table 5.1 summarizes the performance across all methods. DAGDiff achieves the best performance on all three metrics, with an FCE of 60.14%, GSR of 72.50%, and GCR of 15.10%. These results indicate that the proposed method generates grasps that are not only analytically stable but also execute reliably in simulation while maintaining low collision rates.

Farthest-region based methods such as CGDF [19] and VCGS [66] show significantly lower performance in both FCE and GSR. This is primarily because these approaches decouple grasp generation from stability considerations, relying on spatial heuristics rather than physical constraints. While CGDF improves over VCGS due to better grasp generation, both methods suffer from selecting geometrically incompatible regions, which leads to unstable grasp pairs and higher collision rates.

VLM-based approaches, including UniDiffGrasp [61] and RoboBrainGrasp [101], perform poorly in all metrics. Their reliance on semantic region prediction results in coarse or inaccurate localization, which does not align with physically feasible grasp regions. Consequently, these methods produce grasps that frequently violate force-closure conditions and exhibit high collision rates.

Variant	FCE(%)↑	GSR(%)↑	GCR(%)↓
DAGDiff (ours)	**60.14**	**72.50**	**15.10**
w/o FC Head	24.94	30.06	16.85
w/o Collision Head	50.01	55.67	23.50
Post-hoc FC Head	32.37	46.42	18.36

Table 5.2: Results of the ablation study across three variants: without the force-closure head, without the collision head, and with a post-hoc force-closure head.

Compared to DualAfford [44], which uses a category-specific training, DAGDiff is able to demonstrate stronger generalization by achieving higher grasp success rates in a zero-shot setting on unseen objects. This highlights the advantage of modeling grasp generation as a unified, physically grounded generative process rather than relying on object-specific affordances.

Overall, the results show that incorporating stability and collision guidance directly into the diffusion process leads to substantial improvements in grasp quality. DAGDiff approximately doubles the stability and success rates compared to prior methods while significantly reducing collisions, demonstrating its effectiveness for reliable dual-arm grasp generation.

Ablation Study We conduct ablation studies to evaluate the contribution of key components in DAGDiff, particularly the role of classifier-guided diffusion in enforcing stability and collision-free grasp generation. The results are summarized in Table 5.2.

(i) Effect of Force-Closure Guidance: To assess the importance of explicit stability supervision, we remove the Force-Closure (FC) head and rely solely on the Energy head for grasp generation. This leads to a significant drop in both FCE (60.14% $\rightarrow$ 24.94%) and GSR (72.50% $\rightarrow$ 30.06%), indicating that the diffusion backbone alone is insufficient to capture dual-arm stability. The FC head provides explicit supervision for force-closure, and its gradients guide the model toward physically stable grasp configurations during generation.

(ii) Effect of Collision Guidance: Next, we remove the Collision head and evaluate the generated grasps without collision refinement. This results in a notable increase in collision rate (15.10% $\rightarrow$ 23.50%), along with a corresponding drop in FCE and GSR. This highlights that geometric validity cannot be implicitly enforced by the energy-based formulation alone. The Collision head provides targeted gradients that push grasp candidates away from object intersections, improving both feasibility and execution robustness.

(iii) Post-hoc Training of the FC Head: To examine whether the FC head can be introduced after training, we train the Energy head and encoder first, and subsequently train the FC head while keeping the backbone fixed. This results in a substantial degradation in performance (FCE: 32.37%, GSR: 46.42%), as the learned features are not adapted to capture stability-relevant information. This

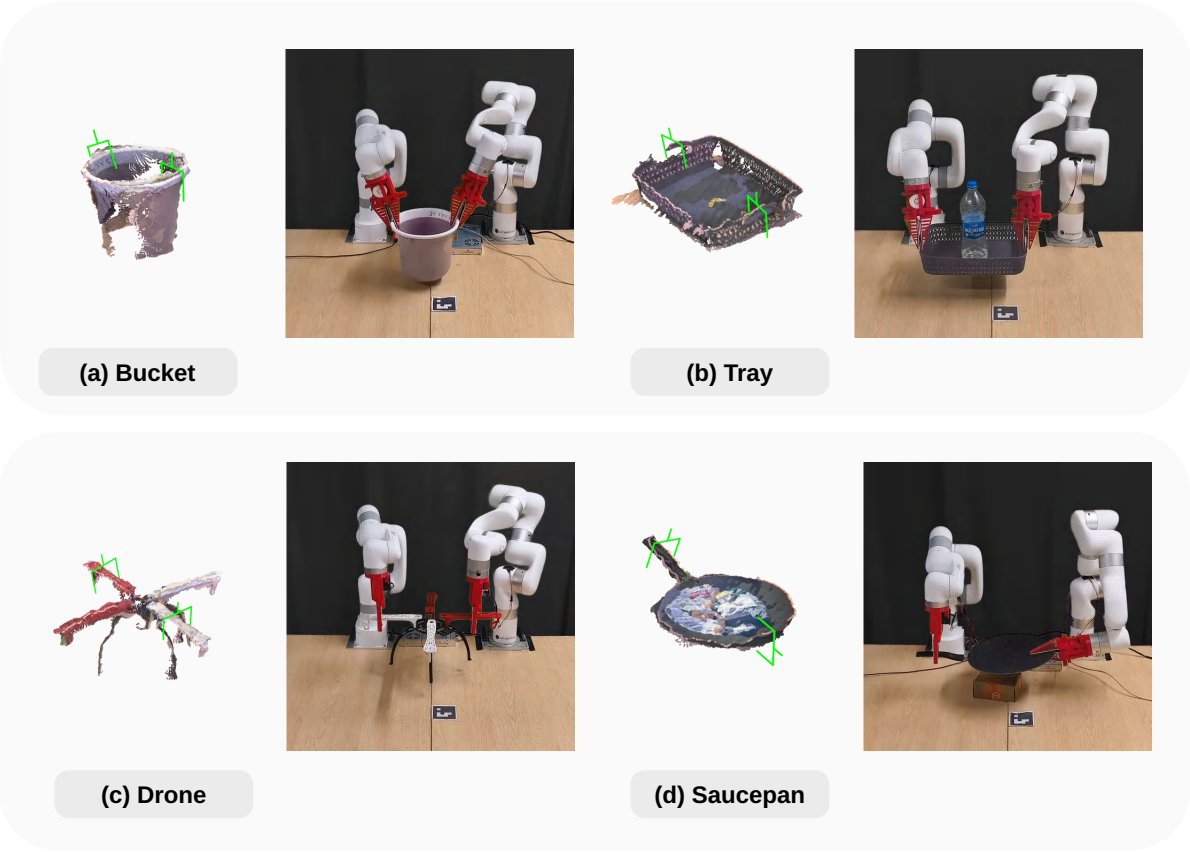

Figure 5.6: Real-world dual-arm grasp execution. Qualitative results on previously unseen objects: bucket, tray, drone, and saucepan. For each example, the left image shows the reconstructed point cloud with predicted grasp poses, while the right image shows successful execution on a real dual-arm robotic system. The results demonstrate reliable grasp generation and execution across diverse object geometries.

demonstrates that the FC head must be jointly trained with the backbone to learn meaningful and stability-aware representations that effectively guide the diffusion process.

Overall, these ablations highlight that each component of DAGDiff plays a critical role. The FC head is essential for enforcing dual-arm stability, the Collision head ensures geometric feasibility, and their joint integration within the diffusion framework is necessary for generating physically valid grasp pairs.

Real-life Results We evaluate the real-world performance of DAGDiff on a heterogeneous dual-arm robotic setup, as shown in Figure 5.6. The system consists of an XArm7 and an XArm6 Lite, calibrated using an AprilTag, and observed by two Intel RealSense D455 cameras for multi-view point cloud acquisition. We conduct experiments on a set of previously unseen real-world ob-

Object	Tray	Bucket	Saucepan	Frypan	Drone
Success	6/10	8/10	7/10	6/10	5/10

Table 5.3: Real-world dual-arm grasp execution results, where each entry denotes the number of successful grasps out of the total attempts.

jects, including trays, buckets, saucepans, frypans, and drones. For each object, 10 grasp trials are executed. The results are summarized in Table 5.3, where most objects achieve high success rates (e.g., 8/10 for buckets, 7/10 for saucepans), demonstrating reliable execution across diverse object geometries.

Failures are primarily attributed to inaccuracies in point cloud reconstruction, which occasionally result in missing or noisy geometric details. These imperfections can lead to incorrect grasp predictions, particularly in regions with insufficient surface information. Despite this, the majority of grasps successfully achieve stable object lifting, validating both the physical correctness of the generated grasps and the robustness of the method under real sensing conditions. Importantly, the model is trained entirely on synthetic data and directly applied to real-world point clouds without any fine-tuning. The successful execution of dual-arm grasps on unseen objects highlights the strong generalization capability of DAGDiff and its ability to transfer from simulation to real-world settings.

Overall, these experiments demonstrate that incorporating stability and collision-aware guidance within the diffusion process enables the generation of physically realizable dual-arm grasps that can be reliably executed on real robotic systems.

5.5 Chapter Summary

This chapter introduces **DAGDiff**, a unified framework for directly generating dual-arm grasps by jointly modeling geometry and physical stability within a diffusion-based formulation over $SE(3) \times SE(3)$. Building on CGDF and the optimization-based insights from DG16M, we move beyond disjoint pipelines and instead learn a distribution over grasp pairs that are both stable and collision-free. By incorporating classifier-guided signals for force closure and collision avoidance, the model is able to implicitly reason about inter-gripper coordination and generate physically consistent grasps without relying on heuristic region selection or post-hoc filtering. Our experimental results demonstrate significant improvements over prior approaches across analytical, simulation, and real-world evaluations, highlighting the effectiveness of integrating generative modeling with physical constraints. Despite these advances, the framework depends on the quality of learned guidance signals and the availability of accurate, complete object point clouds, which can limit robustness under partial or noisy observations. A promising direction for future work is to extend this formulation to operate under partial observations and integrate perception-aware reasoning, enabling reliable dual-arm grasp generation directly from real-world sensory inputs.

Chapter 6

Conclusion

It takes something more than intelligence to act intelligently.

FYODOR DOSTOEVSKY in *Crime and Punishment*, 1866

6.1 Discussion

Robotic manipulation of large and complex objects often requires two robotic arms to coordinate, where the two grippers must act together to achieve force balance and ensure stable interaction. This makes dual-arm grasping inherently more challenging, as it requires not only generating feasible grasps but also ensuring that the pair is jointly stable. Existing approaches largely treat grasp generation and stability evaluation as separate problems, leading to a gap between geometric feasibility and physical robustness.

In this thesis, we address this gap through a structured progression that bridges generative modeling and analytical physical reasoning. We begin with CGDF, which enables dense and region-aware grasp generation on large and complex objects, providing a scalable way to explore the grasp space. Recognizing that generation alone is insufficient in dual-arm settings, we then introduce DG16M, a physically grounded framework for evaluating grasp stability using an optimization-based force-closure formulation, offering a principled understanding of stable grasp pairs.

Building on these components, we develop DAGDiff, which unifies grasp generation and stability reasoning into a single framework. By modeling grasp pairs jointly in $SE(3) \times SE(3)$ and incorporating stability and collision-aware guidance during generation, we move beyond heuristic pairing strategies and directly generate physically consistent dual-arm grasps. This progression highlights a key insight of our work: dual-arm grasping cannot be reduced to combining independent single-arm grasps, but instead requires joint reasoning over geometry and physical interaction.

Through this progression, we demonstrate a shift from disjoint pipelines toward unified, stability-aware generative models for dual-arm grasping, enabling more reliable and generalizable manipulation on complex objects.

6.2 Impact

From a methodological perspective, this thesis demonstrates that grasp generation and physical stability can be modeled jointly within a unified framework, rather than treated as separate stages. By extending diffusion-based generative models to the joint pose space $SE(3) \times SE(3)$ and incorporating stability and collision-aware guidance, we show that physical constraints can be integrated directly into the generative process. Additionally, the development of DG16M provides a physically grounded dataset for dual-arm grasp evaluation, contributing a useful resource for future research in stability-aware manipulation.

From an application perspective, the proposed framework enables more reliable grasping of large and complex objects, which is essential for real-world robotic systems. The successful transfer of our approach from simulation to real-world dual-arm setups demonstrates its practical feasibility under real-world observations. These results indicate that such methods can be extended to larger and more capable robotic platforms, including systems such as humanoids, where coordinated dual-arm manipulation is critical.

Overall, this work highlights the potential of combining generative modeling with physical reasoning to develop more robust and deployable robotic manipulation systems.

6.3 Future Research Directions

While this thesis presents a unified framework for dual-arm grasp generation, several limitations point to promising directions for future work. First, our methods assume access to full object point clouds, and extending them to operate under **partial observations** (e.g., single RGB-D views) with real-world noise remains an important challenge. Second, incorporating **kinematics and dynamics awareness** by accounting for reachability, joint limits, and execution constraints can enable more practical and task-aware grasp generation. Third, integrating **active and multimodal perception**, where the robot can gather better observations and leverage inputs such as language and videos, can improve scene understanding and guide grasping through learned affordances. Finally, the larger goal is to move **beyond grasping to full manipulation**, extending these ideas to long-horizon tasks where grasping is one component within a broader pipeline of planning, control, and interaction. Addressing these challenges will be crucial for enabling robust, generalizable, and real-world deployable robotic manipulation systems.

List of Related Publications

[P1] Gaurav Singh, Sanket Kalwar, Md Faizal Karim, Bipasha Sen, Nagamanikandan Govindan, Srinath Sridhar, and K Madhava Krishna; **Constrained 6-DoF Grasp Generation on Complex Shapes for Improved Dual-Arm Manipulation**, in proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2024.

[P2] Md Faizal Karim, Mohammed Saad Hashmi, Shreya Bollimuntha, Mahesh Reddy Tapeti, Gaurav Singh, Nagamanikandan Govindan, and K Madhava Krishna; **DG16M: A Large Scale Dataset for Dual-Arm Grasping with Force-Optimized Grasps**, in proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2025.

[P3] Md Faizal Karim, Vignesh Vembar, Keshab Patra, Gaurav Singh, and K Madhava Krishna; **DAGDiff: Guiding Dual-Arm Grasp Diffusion to Stable and Collision-Free Grasps**, in proceedings of IEEE International Conference on Robotics and Automation (ICRA), 2026.

Related publications (not included in the thesis):

[P4] Md Faizal Karim, Shreya Bollimuntha, Mohammed Saad Hashmi, Autrio Das, Gaurav Singh, Srinath Sridhar, Arun Kumar Singh, N Govindan, and K Madhava Krishna; **DA-VIL: Adaptive Dual-Arm Manipulation with Reinforcement Learning and Variable Impedance Control**, in proceedings of IEEE International Conference on Robotics and Automation (ICRA), 2025.

Bibliography

[1] A. Bicchi, "On the closure properties of robotic grasping," *Int. J. Rob. Res.*, vol. 14, no. 4, p. 319–334, Aug. 1995. [Online]. Available: https://doi.org/10.1177/027836499501400402

[2] H. Zhang, J. Tang, S. Sun, and X. Lan, "Robotic grasping from classical to modern: A survey," 2022. [Online]. Available: https://arxiv.org/abs/2202.03631

[3] C. Ferrari and J. Canny, "Planning optimal grasps," in *Proceedings 1992 IEEE International Conference on Robotics and Automation*, 1992, pp. 2290–2295 vol.3.

[4] C. Smith, Y. Karayiannidis, L. Nalpantidis, X. Gratal, P. Qi, D. V. Dimarogonas, and D. Kragic, "Dual arm manipulation—a survey," *Robotics and Autonomous Systems*, vol. 60, no. 10, pp. 1340–1353, 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S092188901200108X

[5] O. Kroemer, S. Niekum, and G. Konidaris, "A review of robot learning for manipulation: Challenges, representations, and algorithms," 2020. [Online]. Available: https://arxiv.org/abs/1907.03146

[6] A. Billard and D. Kragic, "Trends and challenges in robot manipulation," *Science*, vol. 364, no. 6446, p. eaat8414, 2019. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.aat8414

[7] A. Edsinger and C. C. Kemp, *Two Arms Are Better Than One: A Behavior Based Control System for Assistive Bimanual Manipulation*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 345–355. [Online]. Available: https://doi.org/10.1007/978-3-540-76729-9_27

[8] H. Ravichandar, A. S. Polydoros, S. Chernova, and A. Billard, "Recent advances in robot learning from demonstration," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 3, no. Volume 3, 2020, pp. 297–330, 2020. [Online]. Available: https://www.annualreviews.org/content/journals/10.1146/annurev-control-100819-063206

[9] M. F. Karim, S. Bollimuntha, M. S. Hashmi, A. Das, G. Singh, S. Sridhar, A. K. Singh, N. Govindan, and K. M. Krishna, "Da-vil: Adaptive dual-arm manipulation with reinforcement learning and

variable impedance control," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*, 2025, pp. 11 896–11 903.

[10] R. M. Murray, S. Sastry, and Z. Li, "A mathematical introduction to robotic manipulation," in *Proceedings of the Conference on Robotics Manipulation*, 1994. [Online]. Available: https://api.semanticscholar.org/CorpusID:108605633

[11] D. Prattichizzo and J. C. Trinkle, *Grasping*. Cham: Springer International Publishing, 2016, pp. 955–988. [Online]. Available: https://doi.org/10.1007/978-3-319-32552-1_38

[12] S. Bai, W. Song, J. Chen, Y. Ji, Z. Zhong, J. Yang, H. Zhao, W. Zhou, W. Zhao, Z. Li *et al.*, "Towards a unified understanding of robot manipulation: A comprehensive survey," *arXiv preprint arXiv:2510.10903*, 2025.

[13] Y. Wang and H. Kasaei, "Learning dual-arm coordination for grasping large flat objects," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*, 2025, pp. 7997–8003.

[14] M. Sundermeyer, A. Mousavian, R. Triebel, and D. Fox, "Contact-graspnet: Efficient 6-dof grasp generation in cluttered scenes," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 13 438–13 444.

[15] H. Cheng, Y. Wang, and M. Q.-H. Meng, "Grasp pose detection from a single rgb image," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021, pp. 4686–4691.

[16] J. Redmon and A. Angelova, "Real-time grasp detection using convolutional neural networks," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, 2015, pp. 1316–1322.

[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 Transactions on Robotics*, vol. 39, no. 5, pp. 3929–3945, 2023.

[18] I. Lenz, H. Lee, and A. Saxena, "Deep learning for detecting robotic grasps," 2014. [Online]. Available: https://arxiv.org/abs/1301.3592

[19] G. Singh, S. Kalwar, M. F. Karim, B. Sen, N. Govindan, S. Sridhar, and K. M. Krishna, "Constrained 6-dof grasp generation on complex shapes for improved dual-arm manipulation," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 7344–7350.

[20] M. F. Karim, M. S. Hashmi, S. Bollimuntha, M. R. Tapeti, G. Singh, N. Govindan, and K. M. Krishna, "Dg16m: A large-scale dataset for dual-arm grasping with force-optimized grasps," *arXiv preprint arXiv:2503.08358*, 2025.

[21] M. F. Karim, V. Vembar, K. Patra, G. Singh, and K. M. Krishna, "Dagdiff: Guiding dual-arm grasp diffusion to stable and collision-free grasps," in *2026 International Conference on Robotics and Automation (ICRA)*. IEEE, 2026.

[22] 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," 2017.

[23] A. Mousavian, C. Eppner, and D. Fox, "6-dof graspnet: Variational grasp generation for object manipulation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 2901–2910.

[24] Z. Jiang, Y. Zhu, M. Svetlik, K. Fang, and Y. Zhu, "Synergies between affordance and geometry: 6-dof grasp detection via implicit representations," *arXiv preprint arXiv:2104.01542*, 2021.

[25] H.-S. Fang, C. Wang, M. Gou, and C. Lu, "Graspnet-1billion: A large-scale benchmark for general object grasping," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 444–11 453.

[26] J. Urain, N. Funk, J. Peters, and G. Chalvatzaki, "Se(3)-diffusionfields: Learning smooth cost functions for joint grasp and motion optimization through diffusion," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 5923–5930.

[27] A. Murali, B. Sundaralingam, Y.-W. Chao, W. Yuan, J. Yamada, M. Carlson, F. Ramos, S. Birchfield, D. Fox, and C. Eppner, "Graspgen: A diffusion-based framework for 6-dof grasping with on-generator training," 2025. [Online]. Available: https://arxiv.org/abs/2507.13097

[28] Z. Weng, H. Lu, D. Kragic, and J. Lundell, "Dexdiffuser: Generating dexterous grasps with diffusion models," *IEEE Robotics and Automation Letters*, 2024.

[29] Z. Zhang, L. Zhou, C. Liu, Z. Liu, C. Yuan, S. Guo, R. Zhao, M. H. A. Jr., and F. E. Tay, "Dexgrasp-diffusion: Diffusion-based unified functional grasp synthesis method for multi-dexterous robotic hands," 2024. [Online]. Available: https://arxiv.org/abs/2407.09899

[30] Y. Zhong, Q. Jiang, J. Yu, and Y. Ma, "Dexgrasp anything: Towards universal robotic dexterous grasping with physics awareness," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 22 584–22 594.

[31] A. Miller and P. Allen, "Graspit! a versatile simulator for robotic grasping," *IEEE Robotics and Automation Magazine*, vol. 11, no. 4, pp. 110–122, 2004.

[32] J. Redmon and A. Angelova, "Real-time grasp detection using convolutional neural networks," 2015. [Online]. Available: https://arxiv.org/abs/1412.3128

[33] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," 2022. [Online]. Available: https://arxiv.org/abs/1312.6114

[34] K. Sohn, X. Yan, and H. Lee, "Learning structured output representation using deep conditional generative models," in *Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'15. Cambridge, MA, USA: MIT Press, 2015, p. 3483–3491.

[35] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International conference on machine learning*. pmlr, 2015, pp. 2256–2265.

[36] M. A. Roa and R. Suárez, "Grasp quality measures: review and performance," *Auton. Robots*, vol. 38, no. 1, p. 65–88, Jan. 2015. [Online]. Available: https://doi.org/10.1007/s10514-014-9402-3

[37] 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.

[38] E. Todorov, T. Erez, and Y. Tassa, "Mujoco: A physics engine for model-based control," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012, pp. 5026–5033.

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

[40] J. Gu, F. Xiang, X. Li, Z. Ling, X. Liu, T. Mu, Y. Tang, S. Tao, X. Wei, Y. Yao, X. Yuan, P. Xie, Z. Huang, R. Chen, and H. Su, "Maniskill2: A unified benchmark for generalizable manipulation skills," in *International Conference on Learning Representations*, 2023.

[41] E. Rohmer, S. P. N. Singh, and M. Freese, "Coppeliasim (formerly v-rep): a versatile and scalable robot simulation framework," in *Proc. of The International Conference on Intelligent Robots and Systems (IROS)*, 2013.

[42] C. Eppner, A. Mousavian, and D. Fox, "Acronym: A large-scale grasp dataset based on simulation," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 6222–6227.

[43] G. Zhai, Y. Zheng, Z. Xu, X. Kong, Y. Liu, B. Busam, Y. Ren, N. Navab, and Z. Zhang, "Da^2 dataset: Toward dexterity-aware dual-arm grasping," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, p. 8941–8948, 2022.

[44] Y. Zhao, R. Wu, Z. Chen, Y. Zhang, Q. Fan, K. Mo, and H. Dong, "Dualafford: Learning collaborative visual affordance for dual-gripper manipulation," in *International Conference on Learning Representations*, 2023.

[45] K. Mo, S. Zhu, A. X. Chang, L. Yi, S. Tripathi, L. J. Guibas, and H. Su, "PartNet: A large-scale benchmark for fine-grained and hierarchical part-level 3D object understanding," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[46] P. Ardón, E. Pairet, R. P. Petrick, S. Ramamoorthy, and K. S. Lohan, "Learning grasp affordance reasoning through semantic relations," *IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp. 4571–4578, 2019.

[47] Y. Ju, K. Hu, G. Zhang, G. Zhang, M. Jiang, and H. Xu, "Robo-abc: Affordance generalization beyond categories via semantic correspondence for robot manipulation," in *European Conference on Computer Vision*. Springer, 2024, pp. 222–239.

[48] T. Ma, Z. Wang, J. Zhou, M. Wang, and J. Liang, "Glover: Generalizable open-vocabulary affordance reasoning for task-oriented grasping," *arXiv preprint arXiv:2411.12286*, 2024.

[49] K. Mo, L. J. Guibas, M. Mukadam, A. Gupta, and S. Tulsiani, "Where2act: From pixels to actions for articulated 3d objects," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6813–6823.

[50] M. Liu, Y. Zhu, H. Cai, S. Han, Z. Ling, F. Porikli, and H. Su, "Partslip: Low-shot part segmentation for 3d point clouds via pretrained image-language models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 21 736–21 746.

[51] D. Turpin, L. Wang, E. Heiden, Y.-C. Chen, M. Macklin, S. Tsogkas, S. Dickinson, and A. Garg, in *European Conference on Computer Vision*. Springer, 2022, pp. 201–221.

[52] R. Zurbrügg, A. Cramariuc, and M. Hutter, "Graspqp: Differentiable optimization of force closure for diverse and robust dexterous grasping," *arXiv preprint arXiv:2508.15002*, 2025.

[53] B. Lei, W. Jiang, and K. Daniilidis, "Activegrasp: Information-guided active grasping with calibrated energy-based model," *arXiv preprint arXiv:2511.12795*, 2025.

[54] Y. Chen, R. Xu, Y. Lin, H. Chen, and P. A. Vela, "Kgnv2: Separating scale and pose prediction for keypoint-based 6-dof grasp synthesis on rgb-d input," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 2971–2978.

[55] Y. Chen, Y. Lin, R. Xu, and P. Vela, "Keypoint-graspnet: Keypoint-based 6-dof grasp generation from the monocular rgb-d input," *arXiv preprint arXiv:2209.08752*, 2022.

[56] Z. Qin, K. Fang, Y. Zhu, L. Fei-Fei, and S. Savarese, "Keto: Learning keypoint representations for tool manipulation," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 7278–7285.

[57] R. Xu, F.-J. Chu, and P. A. Vela, "Gknet: Grasp keypoint network for grasp candidates detection," *The International journal of robotics research*, vol. 41, no. 4, pp. 361–389, 2022.

[58] G. Tziafas and H. Kasaei, "Towards open-world grasping with large vision-language models," *8th Conference on Robot Learning (CoRL 2024)*, 2024.

[59] C. Tang, D. Huang, W. Ge, W. Liu, and H. Zhang, "Graspgpt: Leveraging semantic knowledge from a large language model for task-oriented grasping," *IEEE Robotics and Automation Letters*, vol. 8, no. 11, pp. 7551–7558, 2023.

[60] C. Tang, D. Huang, W. Dong, R. Xu, and H. Zhang, "Foundationgrasp: Generalizable task-oriented grasping with foundation models," *IEEE Transactions on Automation Science and Engineering*, 2025.

[61] X. Guo, H. Hu, C. Song, J. Chen, Z. Zhao, Y. Fu, B. Guan, and Z. Liu, "Unidiffgrasp: A unified framework integrating vlm reasoning and vlm-guided part diffusion for open-vocabulary constrained grasping with dual arms," 2025. [Online]. Available: https://arxiv.org/abs/2505.06832

[62] A. Deshpande, Y. Deng, J. Salvador, A. Ray, W. Han, J. Duan, R. Hendrix, Y. Zhu, and R. Krishna, "Graspmolmo: Generalizable task-oriented grasping via large-scale synthetic data generation," in *Proceedings of The 9th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, J. Lim, S. Song, and H.-W. Park, Eds., vol. 305. PMLR, 27–30 Sep 2025, pp. 2983–3007. [Online]. Available: https://proceedings.mlr.press/v305/deshpande25a.html

[63] Y. Zhong, X. Huang, R. Li, C. Zhang, Z. Chen, T. Guan, F. Zeng, K. N. Lui, Y. Ye, Y. Liang, Y. Yang, and Y. Chen, "Dexgraspvla: A vision-language-action framework towards general dexterous grasping," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 40, no. 22, pp. 18 836–18 844, Mar. 2026. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/38953

[64] Y. Qian, X. Zhu, O. Biza, S. Jiang, L. Zhao, H. Huang, Y. Qi, and R. Platt, "Thinkgrasp: A vision-language system for strategic part grasping in clutter," *8th Conference on Robot Learning (CoRL 2024)*, 2024.

[65] A. Murali, W. Liu, K. Marino, S. Chernova, and A. Gupta, "Same object, different grasps: Data and semantic knowledge for task-oriented grasping," in *Conference on robot learning*. PMLR, 2021, pp. 1540–1557.

[66] J. Lundell, F. Verdoja, T. N. Le, A. Mousavian, D. Fox, and V. Kyrki, "Constrained generative sampling of 6-dof grasps," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, Oct. 2023, p. 2940–2946.

[67] S. Peng, M. Niemeyer, L. Mescheder, M. Pollefeys, and A. Geiger, "Convolutional occupancy networks," in *European Conference on Computer Vision*. Springer, 2020, pp. 523–540.

[68] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 652–660.

[69] H. Zhao, L. Jiang, J. Jia, P. H. Torr, and V. Koltun, "Point transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 16 259–16 268.

[70] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[71] A. Simeonov, Y. Du, A. Tagliasacchi, J. B. Tenenbaum, A. Rodriguez, P. Agrawal, and V. Sitzmann, "Neural descriptor fields: Se (3)-equivariant object representations for manipulation," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 6394–6400.

[72] N. Thomas, T. Smidt, S. Kearnes, L. Yang, L. Li, K. Kohlhoff, and P. Riley, "Tensor field networks: Rotation-and translation-equivariant neural networks for 3d point clouds," *arXiv preprint arXiv:1802.08219*, 2018.

[73] F. Fuchs, D. Worrall, V. Fischer, and M. Welling, "Se(3)-transformers: 3d roto-translation equivariant attention networks," in *Advances in neural information processing systems*, vol. 33, 2020, pp. 1970–1981.

[74] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," *arXiv preprint arXiv:2011.13456*, 2020.

[75] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[76] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," *Advances in neural information processing systems*, vol. 32, 2019.

[77] C. Deng, O. Litany, Y. Duan, A. Poulenard, A. Tagliasacchi, and L. J. Guibas, "Vector neurons: A general framework for so (3)-equivariant networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 12 200–12 209.

[78] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu, "Shapenet: An information-rich 3d model repository," 2015. [Online]. Available: https://arxiv.org/abs/1512.03012

[79] R. Diankov, "Automated construction of robotic manipulation programs," 2018.

[80] D. Kappler, J. Bohg, and S. Schaal, "Leveraging big data for grasp planning," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, 2015, pp. 4304–4311.

[81] A. Miller, S. Knoop, H. Christensen, and P. Allen, "Automatic grasp planning using shape primitives," in *2003 IEEE International Conference on Robotics and Automation*, vol. 2, 2003, pp. 1824–1829 vol.2.

[82] M. Veres, M. Moussa, and G. Taylor, "An integrated simulator and dataset that combines grasping and vision for deep learning," 02 2017.

[83] A. ten Pas and R. Platt, *Using Geometry to Detect Grasp Poses in 3D Point Clouds*. Cham: Springer International Publishing, 2018, pp. 307–324.

[84] H. Zhang, D. Yang, H. Wang, B. Zhao, X. Lan, J. Ding, and N. Zheng, "Regrad: A large-scale relational grasp dataset for safe and object-specific robotic grasping in clutter," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 2929–2936, 2022.

[85] 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 Robotics and Automation Letters*, vol. 7, no. 1, pp. 470–477, 2021.

[86] D. Guo, Y. Xiang, S. Zhao, X. Zhu, M. Tomizuka, M. Ding, and W. Zhan, "Phygrasp: generalizing robotic grasping with physics-informed large multimodal models," in *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2025, pp. 14 915–14 922.

[87] A. Depierre, E. Dellandréa, and L. Chen, "Jacquard: A large scale dataset for robotic grasp detection," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 3511–3516.

[88] L. Pinto and A. Gupta, "Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours," in *2016 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2016, pp. 3406–3413.

[89] D. Morrison, P. Corke, and J. Leitner, "Egad! an evolved grasping analysis dataset for diversity and reproducibility in robotic manipulation," *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 4368–4375, 2020.

[90] S. Diamond and S. Boyd, "Cvxpy: A python-embedded modeling language for convex optimization," *Journal of Machine Learning Research*, vol. 17, no. 83, pp. 1–5, 2016.

[91] H. Liang, X. Ma, S. Li, M. Görner, S. Tang, B. Fang, F. Sun, and J. Zhang, "Pointnetgpd: Detecting grasp configurations from point sets," in *2019 international conference on robotics and automation (ICRA)*. IEEE, 2019, pp. 3629–3635.

[92] P. Song, J. A. C. Ramón, and Y. Mezouar, "Dynamic evaluation of deformable object grasping," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 4392–4399, 2022.

[93] W. Chow, J. Mao, B. Li, D. Seita, V. Guizilini, and Y. Wang, "Physbench: Benchmarking and enhancing vision-language models for physical world understanding," 2025. [Online]. Available: https://arxiv.org/abs/2501.16411

[94] H. Sun, Q. Gao, H. Lyu, D. Luo, Y. Li, and H. Deng, "Probing mechanical reasoning in large vision language models," *arXiv preprint arXiv:2410.00318*, 2024.

[95] L. Puyin, T. Xiang, E. Mao, S. Wei, X. Chen, A. Masood, L. Fei-Fei, and E. Adeli, "Quantiphy: A quantitative benchmark evaluating physical reasoning abilities of vision-language models," *arXiv preprint arXiv:2512.19526*, 2025.

[96] S. Motamed, L. Culp, K. Swersky, P. Jaini, and R. Geirhos, "Do generative video models understand physical principles?" in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2026, pp. 948–958.

[97] X. Xu, P. Bu, Y. Wang, B. F. Karlsson, Z. Wang, T. Song, Q. Zhu, J. Song, Z. Ding, and B. Zheng, "Deepphy: Benchmarking agentic vlms on physical reasoning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 40, no. 40, 2026, pp. 34 160–34 168.

[98] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," 2021. [Online]. Available: https://arxiv.org/abs/2105.05233

[99] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[100] P. Sun, S. Chen, C. Zhu, F. Xiao, P. Luo, S. Xie, and Z. Yan, "Going denser with open-vocabulary part segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 15 453–15 465.

[101] B. R. Team, "Robobrain 2.0 technical report," 2025. [Online]. Available: https://arxiv.org/abs/2507.02029

[102] Q.-Y. Zhou, J. Park, and V. Koltun, "Open3d: A modern library for 3d data processing," 2018. [Online]. Available: https://arxiv.org/abs/1801.09847

[103] E. Olson, "AprilTag: A robust and flexible visual fiducial system," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, May 2011, pp. 3400–3407.