Comparison of diffusion based Grasp Generation Models

Scientific thesis for the procurance of the degree B.Sc.
from the TUM School of Computation, Information and Technology
at the Technical University of Munich.

Supervised by	Univ.-Prof. Dr.-Ing. Sandra Hirche Ian Huang Chair of Information-Oriented Control
Submitted by	Aditya Verma aditya.verma@tum.de
Submitted on	Munich, 28.02.2025

Abstract

Diffusion models represent a state-of-the-art approach in generative modelling, capable of producing highly precise outputs. Since their introduction, they have gained significant popularity due to their unique combination of flexibility and tractability– two properties that were often at odds in earlier probabilistic generative models. These models draw inspiration from non-equilibrium statistical physics, particularly non-equilibrium thermodynamics. Just as heat spreads in a system until thermal equilibrium is reached, diffusion models progressively add noise to data in a forward process until it becomes nearly indistinguishable from pure noise. This controlled increase in entropy allows for an effective reverse process, where noise is gradually removed to reconstruct high-quality data samples.

This paper compares three state-of-the-art diffusion-based grasp generation models: SE(3)-DiffusionFields, Constrained Grasp Diffusion Fields, and GraspLDM. The models are evaluated based on their ability to generate stable and reliable grasps in a simulated environment. Specifically, we assess grasp success by verifying whether the object remains in the gripper after being lifted. We also measure the Earth Mover's Distance to quantify the similarity between generated grasp distributions and real-world datasets. Additionally, we analyse grasp robustness under perturbations such as shaking, rotations and random movement. To further assess grasp quality, we compute metrics including the smallest singular value, grasp wrench space volume, and Force Closure . Our findings highlight the strengths and limitations of each model, providing insights into their applicability for real-world robotic grasping tasks.

Code available at:
`https://github.com/Vaditya-61299/DiffusionGrasping_Evaluation`

Contents

Chapter 1

Introduction

1.1 Motivation and Background

Autonomous robotic systems must interact with objects in unstructured and dynamic environments, where constraints such as clutter, object irregularities, and task-specific grasping requirements make robotic grasping challenging. Unlike humans, who can instinctively grasp objects in multiple ways, robots must optimise grasp generation through computational models.

The problem of grasp generation has been extensively studied in robotics and artificial intelligence (AI). Traditional approaches rely on geometric heuristics and physics simulations [BK00], [LS88]. However, these approaches struggle with generalisation, especially when encountering new objects or dynamic environments. With the rise of deep learning, data-driven grasp generation has shown promise, with models trained on large datasets to predict feasible grasp poses [MEF19], [MME$^+$20].

Recently, diffusion models have gained attention as a promising technique for generative modelling. Inspired by non-equilibrium thermodynamics [SDWMG15], diffusion models learn by progressively corrupting data with noise and then learning to reverse this process to recover meaningful structures. These models have shown significant success in image generation (e.g., DALL-E, Stable Diffusion) and audio synthesis, demonstrating state-of-the-art performance in complex generative tasks.

Given their ability to model high-dimensional distributions and generate diverse outputs, diffusion models have recently been applied to robotic grasp generation. This emerging field aims to leverage diffusion-based generative processes to improve the quality, diversity, and feasibility of robot-generated grasps. Several works have explored this approach, including:

- SE(3)-DiffusionFields [UFPC23], which formulates grasp generation as an energy minimisation problem using diffusion.

- Constrained Grasp Diffusion Fields (CGDF) [SKK+24], which introduces grasp generation with spatial constraints, improving dual-arm manipulation.
- GraspLDM [BOR+24], which applies diffusion in the latent space of a Variational Auto-encoder (VAE) for efficient grasp synthesis.

While these works have made notable progress, the effectiveness and limitations of diffusion models for grasping remain an open research question. This thesis aims to provide a comprehensive analysis and comparison of these models, investigating their strengths, weaknesses, and potential improvements.

1.2 Problem Statement

The primary objective of this thesis is to explore the effectiveness of diffusion models in robotic grasp generation by analysing and comparing three state-of-the-art approaches: SE(3)-DiffusionFields, CGDF, and GraspLDM. Specifically, this thesis seeks to answer the following research questions:

- Understanding Diffusion for Grasping: How do diffusion models work, and why are they relevant for robotic grasp generation? What are the key advantages of diffusion-based approaches over existing grasp generation methods?
- Evaluation of State-of-the-Art Models: How do SE(3)-DiffusionFields, CGDF, and GraspLDM approach grasp generation differently? What are the strengths and weaknesses of each method in terms of grasp quality, generalisation, and efficiency?

This thesis will provide quantitative and qualitative comparisons of these models, offering insights into the current limitations and future directions for diffusion-based grasp generation.

1.3 Related Work

Early grasp generation models relied on analytical methods or supervised learning approaches using CNNs [RA15]. While these methods produced feasible grasps, they struggled with generalising to unseen objects. Generative models like VAEs and GANs have been explored, but VAEs suffer from blurry reconstructions, and GANs often fail to provide diverse and stable grasps. Diffusion models, which have recently revolutionised image generation, offer an alternative that combines high sample quality with strong diversity, making them a promising candidate for grasp synthesis [DN21].
The first diffusion model, introduced by Sohl-Dickstein et al. (2015) [SDWMG15], demonstrated a generative process that gradually perturbs data with Gaussian noise

and then learns to reverse the process. This avoids the tradeoff between model flexibility and tractability that plagued earlier probabilistic generative models. However, diffusion models were initially overlooked due to their slow sampling times and computational cost. Their resurgence in 2020 was driven by improvements in training and inference techniques, leading to state-of-the-art performance in generative modelling.

1.3.1 DDPM (Denoising Diffusion probabilistic Model)

Diffusion models are a class of generative models inspired by non-equilibrium thermodynamics, particularly Langevin dynamics and Markov chains. They have gained significant traction due to their ability to generate high-quality, diverse samples. The first diffusion-based generative model was introduced in 2015, establishing the foundational idea of gradually transforming a data distribution into noise and learning to reverse this process. However, it was the Denoising Diffusion Probabilistic Models (DDPMs) proposed by Ho et al. (2020) that fully demonstrated their potential in image generation and other generative tasks [HJA20]. DDPMs operate as a two-step process:

Forward Process (Noise Addition):

The input data x_0 undergoes a gradual Markovian transformation into pure Gaussian noise x_T over T steps using a pre-defined noise schedule β_T .

$$q(x_{1:T} \mid x_0) = \prod_{t=1}^{T} q(x_t \mid x_{t-1})$$

$$q(x_t \mid x_{t-1}) = \mathcal{N}(x_t; (1-\beta_t)x_{t-1}, \beta_t I)$$

This ensures a smooth transition from structured data to noise.

Reverse Process (Denoising):

A neural network $\epsilon_\theta(x_t, t)$ learns to approximate the true reverse process, removing noise step-by-step:

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$$

where $\mu_\theta(x_t, t)$ is the predicted mean, and $\Sigma_\theta(x_t, t)$ is the variance.

Training is done by minimizing a noise prediction loss:

$$L = \mathbb{E}_{x_0, \epsilon, t}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|^2\right]$$

This score-based learning approach allows DDPMs to generate high-quality samples by iteratively denoising noisy inputs.

Limitations of DDPMs: Despite their success, DDPMs have several drawbacks: Firstly, the denoising process requires hundreds or even thousands of sequential steps, leading to slow sampling time. Secondly, the choice of β_t affects sample

quality but is not learned. Lastly, the model prioritizes sample quality over explicit likelihood maximization. As a result, the likelihood optimisation is limited.
Improvements with IDDPM: Nichol and Dhariwal (2021) [ND21] introduced *Improved Denoising Diffusion Probabilistic Models* to address these limitations:

- **Better Variance Selection:** Instead of using a fixed variance schedule, IDDPM introduces an improved strategy for computing $\Sigma_\theta(x_t, t)$, leading to sharper samples and better likelihood optimisation.
- **Hybrid Loss Function:** A combination of noise-matching loss and log-likelihood maximization improves training stability and sample fidelity.
- **Faster Sampling:** By modifying the noise schedule, IDDPM enables high-quality sample generation with fewer denoising steps (e.g., 250 instead of 1000).
- **Classifier-Guided Sampling:** A separate classifier provides additional signal during generation, enabling conditional sampling and improving sample realism.

These enhancements significantly improved the efficiency and practicality of DDPMs, paving the way for even more optimised variants like Denoising Diffusion Implicit Models (DDIM), which further reduce sampling time, and Latent Diffusion Models (LDMs), which operate in a compressed latent space for computational efficiency.

1.3.2 DDIM (Denosing Diffusion Implicit Model)

While DDPMs generate high-quality samples, their inefficiency in sampling remains a challenge. To address this, Song et al. (2020) [SME20] introduced Denoising Diffusion Implicit Models (DDIMs), which accelerate sampling without compromising image quality.

DDIMs replace the stochastic reverse diffusion process with a non-Markovian deterministic process that allows skipping multiple steps while maintaining quality:

$$x_{t-1} = \alpha_{t-1} \left(\alpha_t x_t - \alpha_t \epsilon_\theta(x_t, t) \right) + (1 - \alpha_{t-1}) \epsilon_\theta(x_t, t)$$

α_t is a function of the noise schedule, $\epsilon_\theta(x_t, t)$ is the estimated noise, The first term is a deterministic transformation, The second term controls stochasticity (which can be reduced to zero for deterministic generation).

Key Advantages of DDIMs:

- Fast Sampling: DDIMs can generate samples in as few as 25-50 steps, compared to the 1000+ steps required by DDPMs.

- Flexibility: DDIMs allow trade-offs between deterministic and stochastic generation, providing greater control over sample diversity.
- Reusability of DDPM Training: Since DDIMs retain the same training objective as DDPMs, they can be applied to pretrained DDPM models without additional training.

1.3.3 Diffusion models in Grasp Generation

Existing grasp synthesis methods can be categorised into:

- Analytical models based on force-closure principles [EKLB13], [BK00].
- Learning-based approaches, including CNN-based grasp detection [RA15] and generative grasp synthesis [VMT17], [MEF19].
- Diffusion-based approaches, which remain relatively unexplored but offer promising benefits in grasp diversity and quality.

While grasp detection models such as Dex-[MLN$^+$17] and Contact-GraspNet [SMTF21] have advanced the field, recent works have begun leveraging diffusion models for grasp generation. Notable examples include:

- SE(3)-DiffusionFields [UFPC23], which learns smooth cost functions for joint grasp and motion optimisation.
- Constrained 6-DoF Grasp Generation [SKK$^+$24], which improves grasp diversity while adhering to physical constraints.
- GraspLDM [BOR$^+$24], which applies latent diffusion models to 6-DoF grasp synthesis for improved efficiency.

While Language-Driven 6-DoF Grasping [NVH$^+$24] integrates diffusion models with language prompts for goal-directed grasping, this thesis does not explore language-conditioned grasping, focusing instead on general diffusion-based grasp synthesis methods.

Chapter 2

Technical Approach

Grasp generation has traditionally been evaluated based on task success-whether a robotic gripper can pick up and hold an object. However, real-world grasping demands robustness, stability, and adaptability to dynamic conditions. In this chapter, we present our methodology for comparing diffusion-based grasp generation models beyond simple task completion. We focus on three models:

- SE(3)-DiffFields - A diffusion-based energy model that enables smooth cost function optimisation.
- CGDF (Constrained 6-DoF Grasp Diffusion) - An extension of SE(3)-DiffFields, improving latent space encoding for better grasp representation.
- GraspLDM - A novel approach that leverages two generative models to improve diversity and robustness.

To ensure a fair and rigorous comparison, all models are trained on the ACRONYM dataset and evaluated within the IsaacGym simulator.
We assess performance using three key methodologies:

- Grasp Distribution Matching: We use an existing code in [UFPC23] to measure Earth Mover's Distance (EMD), which quantifies how well the generated grasp distributions align with real grasp data. The same script is also used to evaluate grasp success rates, where a grasp is considered successful if the object can be lifted and held without slipping.
- Robustness Under Perturbations (Custom Implementation): To assess real-world feasibility, we introduce controlled external forces (vibration, rotation and random movement).
- Grasp Stability and Quality Analysis (Custom Implementation): We evaluate grasp quality to assess long-term stability, focusing on Force Closure and the algebraic properties of the Grasp Matrix. This ensures that the generated grasps are not only successful initially but also reliable and repeatable over time.

While using an existing script for EMD and grasp success ensures consistency with prior work, it has limitations:

- The script does not explicitly capture grasp robustness beyond initial success cases.
- Additional robustness metrics (e.g., post-perturbation drift) were implemented separately, meaning results from different evaluation stages are not directly comparable without normalisation.

2.1 Grasp Diffusion Models

2.1.1 SE(3)-DiffFields

The SE(3) Diffusion Fields paper by J. Uraine et al. [UFPC23] proposes a novel grasp generation framework by learning a smooth cost function in SE(3) space. Unlike traditional grasp generators that directly predict discrete grasp poses, this method models grasp generation as an energy-based diffusion process, ensuring smooth and continuous optimisation of grasp quality.
Grasp configurations in SE(3) space are represented as 4x4 transformation matrices, which follow the principles of Lie algebra. The advantage of defining a grasp cost function in this space is twofold:

- Avoiding Unfeasible Grasp Solutions: Traditional data-driven grasp samplers often produce kinematically infeasible solutions, whereas a smooth energy function allows for gradient-based refinement.
- Preventing Non-Informative Gradients: Scalar field-based grasp generators often suffer from plateau regions, where the gradient provides little information for optimisation. By formulating the grasping problem as an SE(3) cost function, the model ensures smooth gradient propagation for optimisation tasks.

Additionally, this approach leverages multi-modal diffusion, allowing the model to learn multiple grasp solutions for the same object. Furthermore, since the grasp cost function is differentiable, it can be combined with other cost functions to optimise complex manipulation tasks.
Since the grasp cost function is formulated in SE(3) space, all necessary operations, such as sampling and loss computation, must be performed within the Lie algebra framework. Transformations between SE(3) and Gaussian R6R6 space are handled using exponential and logarithmic maps, allowing efficient optimisation.

Sampling in SE(3)-Space

To sample a grasp pose from the learned distribution, a perturbation is applied in Lie space:

$$H = H \cdot \text{ExpMap}(\epsilon), \quad \epsilon \sim \mathcal{N}(0, \sigma_k^2 I)$$

where H is the grasp transformation matrix, and the perturbation ϵ is drawn from a zero-mean Gaussian with variance σ_k^2.

Loss Function for SE(3) Diffusion Fields

The training objective is based on denoising score matching (DSM), where the loss function is formulated as:

$$L_{\text{DSM}} = \frac{1}{L} \sum_{k=0}^{L} \mathbb{E}_{H,H^*} \left[\| s_\theta(H^*, k) - D_{H^*} \log q(H^* \mid H, \sigma_k I) \| \right]$$

where H^* represents a perturbed grasp pose, $s_\theta(H^*, k)$ is the learned score function, and $q(H^* \mid H, \sigma_k I)$ represents the perturbed grasp distribution.

Grasp Refinement using the Langevin Process

Grasp refinement is achieved via an inverse Langevin sampling process, which iteratively updates the grasp pose:

$$H_{k-1} = \text{ExpMap}\left(2\alpha_k^2 s_\theta(H_k, k) + \alpha_k \epsilon\right)$$

where α_k controls the step size and $s_\theta(H_k, k)$ guides the refinement towards more feasible grasps.

The model is trained using the ACRONYM dataset, which provides object meshes, grasp poses, and Signed Distance Fields (SDFs). The training pipeline consists of the following steps:

- Feature Encoding: Each object's point cloud and associated grasp poses are encoded into a latent space.
- SDF-Based Learning: The model jointly learns to predict the object's Signed Distance Field (SDF) along with grasp energy, improving grasp accuracy.
- Lie Space Sampling: Successful grasp poses are perturbed in Lie space and mapped back into the object's reference frame.
- Grasp Energy Estimation: The model extracts grasp features, flattens them, and decodes the representation to predict an energy score, determining grasp feasibility.

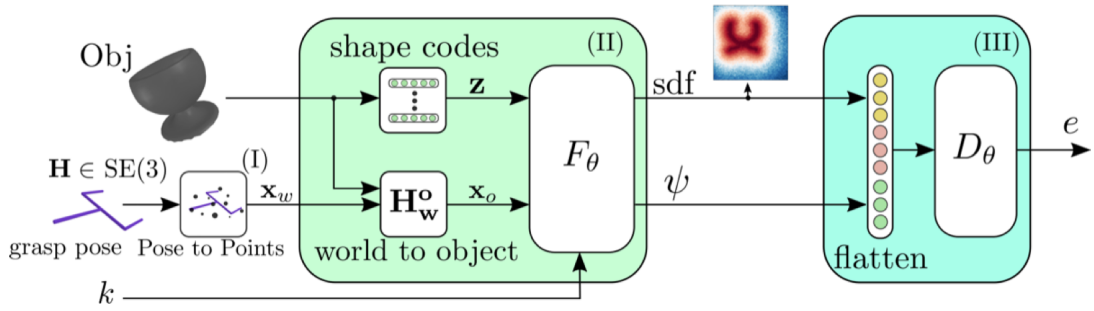

Figure 2.1: *Training Architecture of SE(3)-DiffFields [UFPC23]*

The training loss consists of two terms: L2 loss on SDF prediction to ensure object geometry is accurately modelled and Denoising score matching loss to refine grasp predictions in SE(3) space.

The advantages lie in the smooth Grasp Cost Function, enabling gradient-based refinement and integration with other optimisation objectives. Moreover the Multi-Modal Grasp Generation allows learning multiple valid grasps instead of a single deterministic output. Finally the SE(3) Representation ensures physically valid 6-DoF grasp generation by operating in the correct transformation space. However, the model is computationaly expensive and parameter sensitive.

2.1.2 CGDF: Constrained Grasp Diffusion Fields

The Constrained Grasp Diffusion Field (CGDF) [SKK+24] builds upon SE(3)-DiffFields by introducing constraints on the grasp generation process, allowing grasps to be restricted to specific regions of the object. This capability is particularly useful in settings such as dual-arm manipulation, where each arm must grasp a specific part of an object.

For the purposes of this thesis, we focus solely on single-arm unconstrained grasping, as our primary goal is to analyse how changes in object encoding impact grasp quality. Unlike SE(3)-DiffFields, which rely on PointNet-based feature encoding, CGDF employs a Convolutional Occupancy Network [PNM+20] for encoding object geometry. This design choice aims to enhance the representation of object surfaces and, in turn, improve the accuracy of generated grasp poses.

By comparing SE(3)-DiffFields with CGDF, we can assess whether alternative encoding mechanisms lead to tangible improvements in grasp success rates. Furthermore, this comparison highlights how diffusion-based grasp models remain an evolving field, with architectural refinements leading to qualitative differences in the grasp configurations they generate.

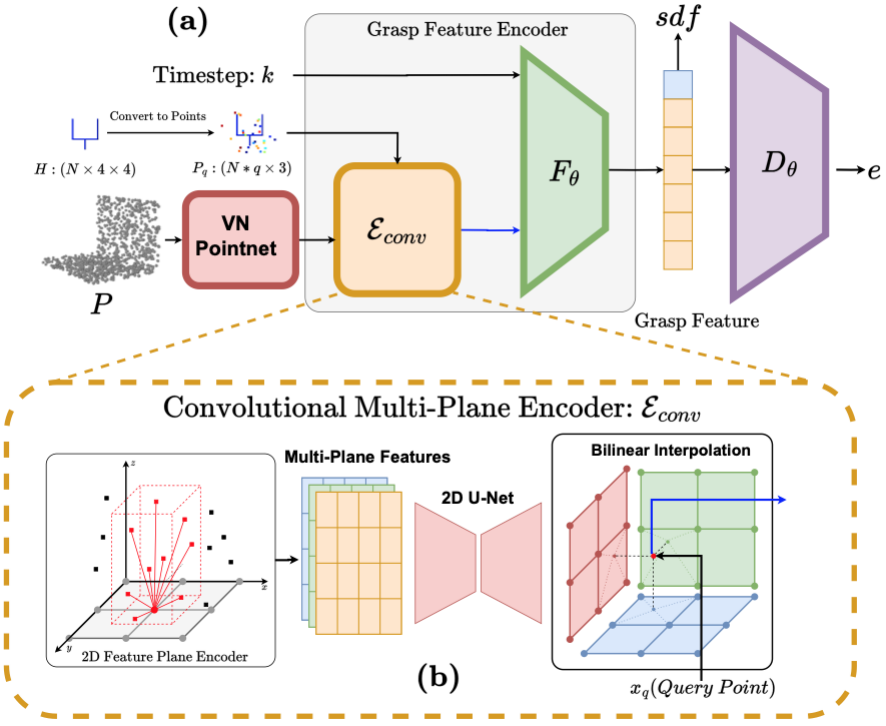

Figure 2.2: *Training Architecture of CGDF [SKK+24]*

CGDF retains the core diffusion framework of SE(3)-DiffFields but introduces a modified encoding pipeline for object representation. The encoding process consists of the following steps:

- **Point Cloud Encoding with VN-PointNet:**
  Similar to SE(3)-DiffFields, the point cloud is initially processed using Vector-Neuron (VN) PointNet, which ensures SE(3) equivariance. This step helps preserve spatial relationships in the grasp generation process.
- **Feature Extraction via Convolutional Occupancy Networks (ConvONet):**
  Instead of directly using PointNet features, CGDF maps object points onto three orthogonal feature planes(XY, YZ, and XZ planes). This representation captures spatial context more effectively, allowing the model to better encode object geometry.
- **Feature Sampling using Bilinear Interpolation:**
  For a given query point x_q the model retrieves a feature vector z_{x_q} from the feature planes. This is achieved through bilinear interpolation, which ensures that features are smoothly extracted from the 2D projections.
- **Final Feature Encoding and Grasp Prediction:** The encoded feature vector is processed in a similar manner to SE(3)-DiffFields. It is

passed through a series of fully connected layers to produce an energy-based grasp score, which is used during the diffusion process.

Compared to SE(3)-DiffFields, CGDF offers several advantages that enhance its suitability for grasp generation tasks. First, CGDF provides a more expressive object encoding by utilising feature planes rather than relying solely on point-cloud-based encoding. This approach allows the network to capture object surfaces more effectively, leading to a richer representation of object geometry. Additionally, CGDF improves grasp quality by leveraging the spatial structure introduced by ConvONet, which contributes to generating more accurate and stable grasp poses. Furthermore, CGDF is better suited for constrained grasping scenarios, as it enables grasp generation to be restricted to specific object regions. This capability is particularly beneficial for applications involving dual-arm manipulation and structured tasks, where precise control over grasp locations is essential.

Despite its advantages, CGDF has certain limitations in the context of simulation-based grasp generation that warrant further investigation. One key concern is its computational complexity, as the additional encoding step introduces more parameters, potentially leading to slower training and inference times compared to other diffusion-based models. Additionally, CGDF's focus on constrained dual-arm grasping raises questions about its effectiveness in unconstrained single-arm grasping tasks, which are common in simulation environments. These considerations highlight the need for further research to assess the trade-offs and broader applicability of CGDF in diverse grasping scenarios.

2.1.3 GraspLDM

GraspLDM [BOR+24] introduces a novel grasp generation framework by leveraging diffusion models in the latent space of a Variational Autoencoder (VAE). Traditional VAE-based grasp generators suffer from poor sampling quality due to the limitations of their posterior distribution [MEF19]. On the other hand, diffusion models generate high-quality and diverse samples, but are often computationally expensive when applied directly in high-dimensional spaces like SE(3).
GraspLDM aims to combine the efficiency of VAEs with the expressiveness of diffusion models. Instead of running the diffusion process directly on the high-dimensional SE(3) grasp space, it operates within a lower-dimensional latent space. This allows for higher-quality grasp generation compared to traditional VAEs, more efficient sampling than direct diffusion on SE(3) transformations and improved diversity of generated grasps by leveraging latent-space diffusion.

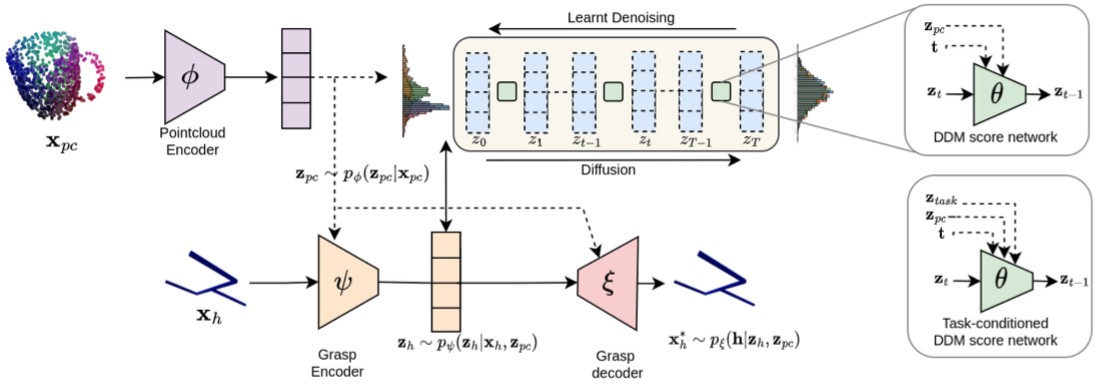

Figure 2.3: *Training Architecture of GraspLDM [BOR+24]*

GraspLDM has two stage training process. In the first stage, the point cloud encoder, grasp encoder, and grasp decoder are jointly trained to maximize the Evidence Lower Bound (ELBO):

$$L_{\text{ELBO}}(\phi, \psi, \xi) = \mathbb{E}\left[\log p_\xi(H^* \mid z_h, z_{\text{pc}})\right] - \lambda D_{\text{KL}}\left(q_\psi(z_h \mid H, z_{\text{pc}}) \parallel \mathcal{N}(0, I)\right)$$

where H^* represents the ground-truth grasp pose, z_{pc} is the encoded object representation from the point cloud encoder, z_h is the grasp embedding in the learned latent space and D_{KL} ensures that the learned latent distribution does not diverge excessively from a standard Gaussian, regularizing the latent space.

After the VAE training converges, the second phase starts in which the diffusion process is trained separately within the learned latent space. The diffusion model refines the generated grasp poses by denoising samples in the latent space. By keeping the VAE's latent space fixed, the diffusion model learns more stable gradients, leading to faster and more efficient training.

Rather than training the entire model end-to-end, GraspLDM first optimises the VAE independently. This prevents the diffusion model from constantly adapting to a shifting latent space, which would otherwise slow down convergence and reduce training efficiency.

GraspLDM offers several key benefits over previous grasp generation methods. Its ability to generate a greater diversity of grasps helps mitigate mode collapse, which is a common issue in traditional VAEs, enabling it to produce a wider range of valid grasp poses. By performing diffusion in a lower-dimensional latent space, GraspLDM improves sampling efficiency and reduces inference times compared to methods operating directly in the high-dimensional SE(3) space. Additionally, the combination of Evidence Lower Bound (ELBO) training and denoising diffusion enhances the quality of generated grasps, resulting in more stable and physically valid solutions. These

advantages make GraspLDM a highly efficient and effective approach for generating diverse and high-quality grasps in robotic manipulation tasks.

Despite its advantages, GraspLDM has certain limitations that must be considered. A major drawback is the limited interpretability of its latent space. Unlike SE(3)-based methods, the learned latent representations in GraspLDM lack a clear physical meaning, making it more challenging to analyse and understand the generated grasps. Furthermore, GraspLDM requires an additional Variational Autoencoder (VAE) pretraining stage, which increases the computational overhead compared to methods that directly train the diffusion model. This extra step adds complexity and time to the training process. These factors highlight the need for further refinement to improve interpretability and reduce training time.

2.2 Implementation

All experiments in this study were conducted using the Isaac Gym simulator to ensure efficient and controlled evaluations. To optimise computational efficiency, we utilised the evaluation script from the SE(3)-DiffFields repository, which measures key performance metrics such as Earth Mover's Distance (EMD) and grasp success rate. A grasp is considered successful if the object is lifted and remains within a predefined distance threshold (0.3m) from the gripper fingers after lifting, ensuring stability.
The software implementation is based on PyTorch Framework, utilising Python 2.0.1 with CUDA 11.8 as a prerequisite. The experiment was conducted on Ubunutu 20.04. The dataset used for training and evaluation is ACRONYM [EMF20]. The hardware implementation was tested on an NVIDIA RTX4060 graphic card (GB VRAM) with 16GB RAM.

2.2.1 Metrics 1: Robustness Check

To assess grasp stability beyond static lifting, we extended the evaluation framework to incorporate robustness tests that simulate real-world disturbances. A parameter k is randomly sampled within the range [0.4,0.8] to introduce variability in movement parameters. Values close to 0 result in minimal movement, whereas values near 1 induce extreme perturbations that do not reflect realistic, controlled motion. The robustness evaluation consists of the following tests:

– **Shake Test** - The object is lifted and subjected to oscillatory motion

along the x, y, and z axes one by one, with the following formula:

$$A \sin\left(\frac{t}{T}\right)$$

where A=0.02*k is the shaking amplitude, and T=10*k is the time period of oscillation. There are a total of 5 oscillation in each direction.

- **Rotation Test** -A randomly generated quaternion defines the target orientation, which the robotic system attempts to achieve. Quaternions are preferred over Euler angles due to their advantages in avoiding gimbal lock and ensuring computational efficiency. However, Euler angles could also be used, since the primary focus of this study is to evaluate grasp stability, specifically examining whether the object remains securely held during and after the rotational motion is applied.
- **Rotation and Translation Test** - This test introduces a combination of random rotations (as in the Rotation Test) and translational movements. Unlike the Shake Test, which applies oscillatory motion, this test simulates controlled translational movement more representative of real-world scenarios. A random direction (axis) is chosen, and the object undergoes a simple translation with a displacement of:

  $$\text{Translation Length} = 0.05 \cdot k$$

  This movement ensures a more controlled perturbation, reflecting realistic interactions where objects are moved while maintaining a stable grasp.

2.2.2 Metric 2: Algebraic Properties

To quantitatively assess grasp quality [LS88], we use the grasp matrix.

Grasp Matrix

A grasp matrix is used to analyse the forces and torques that a robotic hand or gripper can apply to an object. It mathematically describes the relationship between the contact forces exerted by the fingers (or gripper) and the resulting wrench (force and torque) on the object. Mathematically, it combines wrench at every point for every point.

$$G = [G_1 \; G_2 \; G_3 \; \ldots \; G_N]$$

$$W_i = \begin{bmatrix} F_i \\ \tau_i \end{bmatrix}$$

Where G_i represents the wrench at contact point i, and N is the total number of contact points.

We can analyse this grasp matrix to evaluate the generated grasps.

- **Smallest Singular Value** - The smallest singular value $\sigma_{\min}(G)$ serves as a crucial stability indicator, where higher values suggest more stable grasps, while $\sigma_{\min}(G) = 0$ indicates a singular configuration that cannot resist external forces.
- **Volume in wrench space** - This metric, defined as the product of all singular values of G, provides a global measure of grasp stability. A larger grasp wrench space volume suggests greater stability, especially when two configurations share the same smallest singular value. This approach is based on prior research by Li and Sastry (1987) [LS88] on task-oriented optimal grasping.

Singular Value Decomposition

Singular Value Decomposition (SVD) is a mathematical technique used to analyse a matrix by breaking it down into three components:

$$G = U\Sigma V^T$$

where:

- U is an **orthogonal/unitary matrix** of shape $m \times m$, containing the **left singular vectors**.
- Σ is a **diagonal matrix** of shape $m \times n$, containing the **singular values** in descending order.
- V is an **orthogonal/unitary matrix** of shape $n \times n$, containing the **right singular vectors**.

The singular values in Σ represent the significance of independent components in the transformation defined by G. A small singular value indicates a weak or nearly non-existent contribution in that direction. If the smallest singular value is close to zero, it suggests that there exists a direction in which external forces cannot be effectively resisted.

Eigenvector Interpretation The left singular vectors (columns of U) correspond to the eigenvectors of GG^T, while the right singular vectors (columns of V) correspond to the eigenvectors of G^TG. The squared singular values in Σ correspond to the eigenvalues of these matrices:

$$GG^TU = U\Lambda, \quad G^TGV = V\Lambda$$

where Λ is a diagonal matrix containing the squared singular values of G.

Application to Grasp Analysis: In grasp analysis, SVD helps us understand the quality of a grasp by analysing the singular values of the grasp matrix G. Specifically:

- The **smallest singular value** $\sigma_{\min}(G)$ provides a measure of grasp stability. If $\sigma_{\min}(G)$ is near zero, the grasp cannot resist forces in certain directions.
- The **left singular vectors** (columns of U) reveal the dominant motion directions of the grasped object when forces are applied.
- The **right singular vectors** (columns of V) indicate the directions in which the contact forces are most effective in generating wrenches.

By analysing $\sigma_{\min}(G)$, we can assess how well a grasp resists external forces and whether the grasp configuration is stable.

2.2.3 Metric 3: Force Closure

For a grasp with frictionless contact points, the grasp map has the same form as the grasp matrix given above [MSZ94]. A grasp achieves **force-closure** if and only if the set of positive linear combinations of the columns of the grasp matrix G spans $\mathbb{R}^6$, that is:

$$G_{FC} = \mathbb{R}^6.$$

To numerically verify force-closure, we check the following two conditions:

1. **Convexity Condition:** The convex hull of the columns of G must contain a neighbourhood around the origin. This ensures that any external wrench can be counteracted by feasible contact forces. For the strictness of this metric, we chose to include the origin as well as the deviation of 0.001 of the convex hull.
2. **Singular Value Condition:** The smallest singular value of G must be at least **0.1**. This threshold quantifies the grasp's ability to resist external forces and indicates grasp stability.

In addition to these conditions, we perform an empirical test by generating **10 random wrenches** and using `linprog()` to determine whether each wrench can be balanced using feasible contact forces. If all 10 wrenches can be resisted, this further confirms that the grasp satisfies force-closure.

2.2.4 Metric 4: Model Success Metrics

To evaluate and validate the performance of the models, we consider the following two success metrics:

- **Success Rate:** A grasp is deemed successful if the gripper can lift the object and maintain a stable hold. Specifically, if the distance between the object and the gripper fingers remains less than 0.3m after lifting, the grasp is classified as successful.
- **Earth Mover's Distance (EMD):** This metric quantifies how closely the generated grasp samples resemble those in the dataset. A lower EMD value indicates better alignment with the distribution of real-world or ground-truth grasps, reflecting improved grasp generation quality.

The implementation is designed to be easily reproducible, with an extended evaluation script enabling verification of grasp quality across different models. A detailed setup guide is provided in the repository to facilitate the reproduction of results.

This implementation is specifically tailored to the ACRONYM dataset, with a hardcoded dataloader that requires modifications to support other datasets. Due to project time constraints, generalising the evaluation script to multiple datasets was not feasible; however, this remains a potential avenue for future work. Expanding the script to incorporate a more flexible dataloader would enable broader applicability across different grasping datasets.

Chapter 3

Evaluation

During the evaluation, we consider two subsets:
Subset 1) A set with 15 random classes, selecting the first five objects with at least one good grasp in the subset.
Subset 2) A class of mugs with 101 different objects.

For each object, 50 grasps were generated and evaluated. Additionally, force closure and robustness are considered in two cases: absolute and relative. Absolute values are determined by the total number of grasps, while relative values are calculated based on successful grasps.

3.1 Experimental Results

We first analyse the results of the model along with robustness checks. For SE3-DiffusionFields, the average success rate is approximately 58.08%, with 11.84% of the successful grasps classified as robust and 7.02% satisfying the force closure criteria. Comparatively, the Constrained Grasp Diffusion Fields model shows a slightly lower performance, achieving a success rate of 52.56%, with 7.6% of successful grasps being robust and 5.93% meeting the force closure criterion. Meanwhile, GraspLDM had a success rate of 49.33%, but exhibited the highest robustness relative to successful grasps at 16.91%, while 6.32% met the force closure requirement.

The Earth Mover's Distance (EMD) results show relatively low values for Subset 2, consistent with findings from the original SE3-DiffusionFields paper, with values slightly increased for Subset 1. CGDF displayed varying EMD values for Subset 1 while showing better values for Subset 2. Meanwhile, GraspLDM also displays values similar to the original findings.

The smallest singular values range between 0.4 and 0.003, while the grasp wrench space volume typically falls between 0.1 and 100,000. Notably, as the smallest singular value increases, the grasp wrench space volume also increases. While some deviations from the mean are present, their frequency remains relatively low.

3.1.1 Subset 1 Results

Table 3.1: Comparison of grasping performance metrics for Subset 1

Metric (in %)	**SE3-DiffusionFields**	**Constrained Grasp Diffusion Fields**	**GraspLDM**
Success cases	58.08	52.56	49.33
Robust cases absolute	6.8	4	8.3
Robust cases relative	11.84	7.6	16.91
Force Closure absolute	4.08	3.12	3.12
Force Closure relative	7.02	5.93	6.32

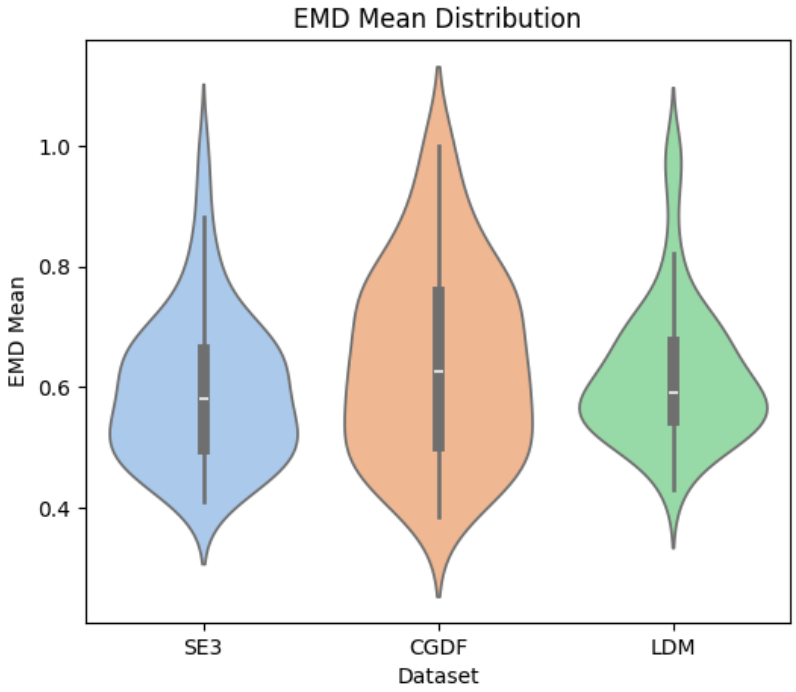

Figure 3.1: Earth Mover's Distance (EMD) results for Subset 1.

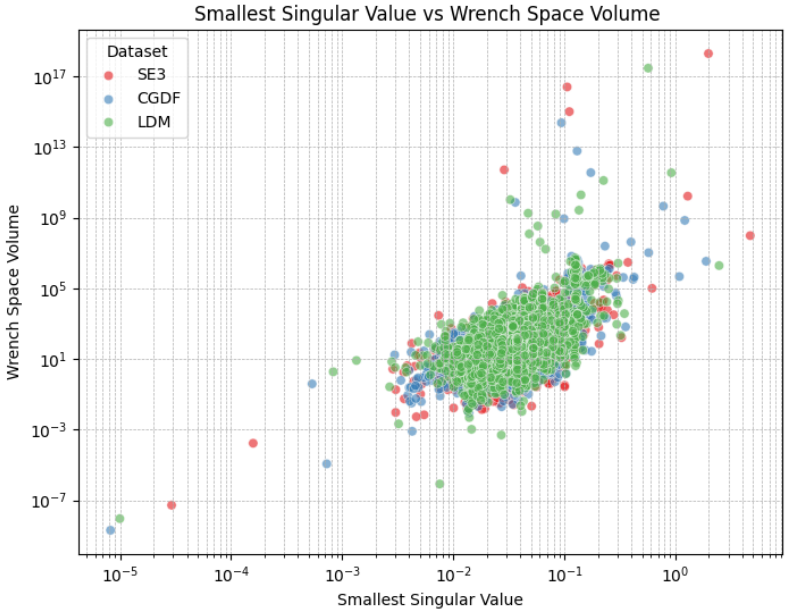

Figure 3.2: Smallest singular values and grasp wrench space volume for Subset 1.

3.1.2 Subset 2 Results

Table 3.2: Performance on the Mug class (Subset 2)

Metric (in %)	**SE3-DiffusionFields**	**Constrained Grasp Diffusion Fields**	**GraspLDM**
Success cases	80.06	77.62	69.82
Robust cases absolute	12.02	7.8	14.35
Robust cases relative	14.89	10.1	20.56
Force Closure absolute	3.6	3.08	3.2
Force Closure relative	4.5	3.97	4.6

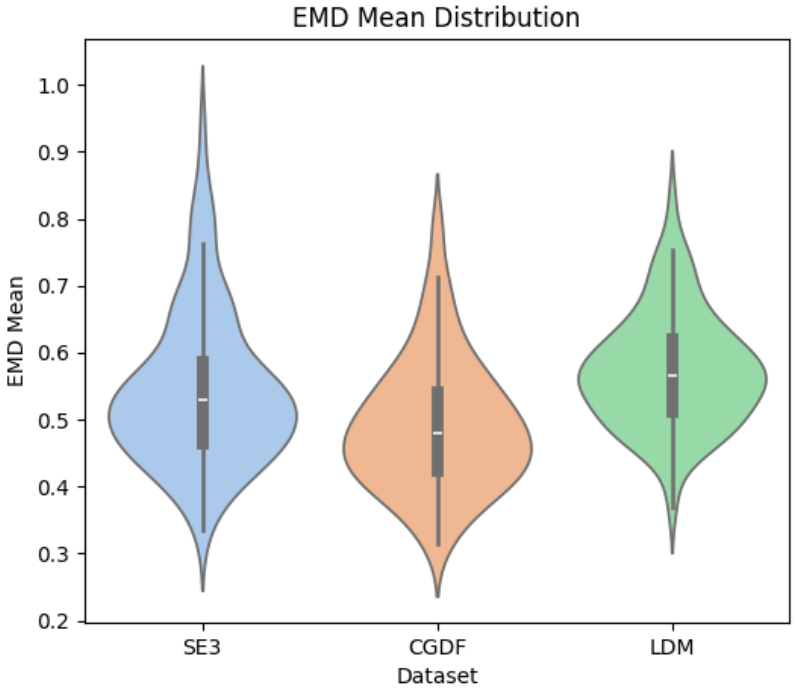

Figure 3.3: Earth Mover's Distance (EMD) results for Mug class (Subset 2).

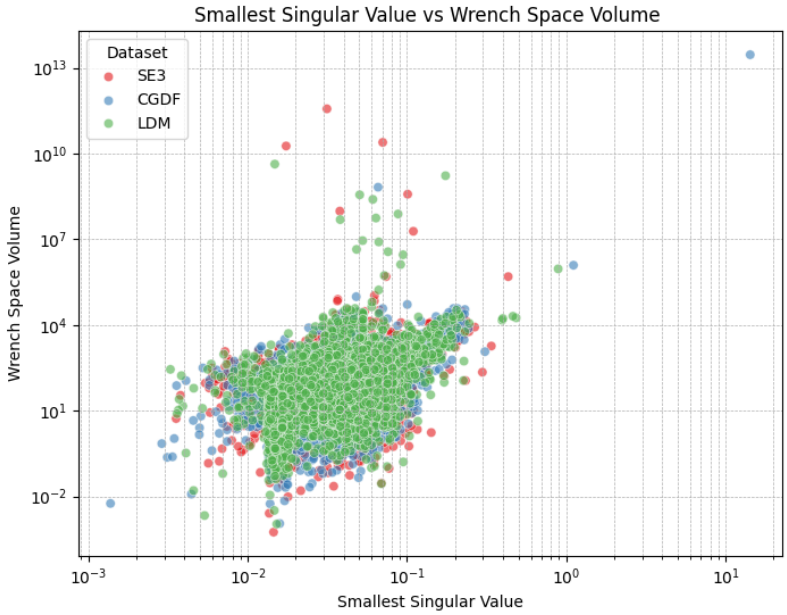

Figure 3.4: Smallest singular values and grasp wrench space volume for Mug class (Subset 2).

Chapter 4

Discussion

The increase in the success rate for the Mug class suggests that the models demonstrate improved performance on objects with specific geometric properties. Mugs, which generally exhibit cylindrical or slightly tapered shapes, may facilitate easier grasp generation compared to more complex or irregular objects. While SE3-DiffFields achieves a higher success rate compared to CGDF, both models exhibit relatively low force closure rates. This indicates that while the models generate kinematically feasible grasps, they do not always ensure dynamic stability.On the other hand, GraspLDM achieves a lower success rate than both SE3-DiffFields and CGDF but produces grasps that are more stable. This is due to the fact that the generated grasps by GraspLDM are more similar to those found in the ACRONYM dataset, which consists of stable grasps under perturbation. Since the dataset itself defines robustness, GraspLDM inherently produces stable grasps by virtue of its dataset alignment rather than any explicit stability training.

This difference in metrics arises due to two key reasons: **(1)** SE3-DiffFields and CGDF do not have a separate train-test split, as their datasets were randomly divided for training and testing. Consequently, these models were effectively exposed to all objects during training. In contrast, GraspLDM follows a more rigorous training paradigm by maintaining a distinct train-test split. To establish a fair baseline for evaluation, all objects in Subset 2 (Mugs) were included in the comparison, regardless of their involvement in training, since SE3-DiffFields and CGDF do not distinguish between train and test objects. For the same reason, Subset 1 also included objects that may have been included in the training of those models. **(2)** The EMD metric places greater emphasis on rotational similarity over translational alignment, leading to a bias toward grasps that are rotationally consistent with the dataset rather than optimally positioned for physical execution. Although GraspLDM has EMD values similar to SE3-DiffFields, its grasps tend to be closer to the object, making them more representative of the original ACRONYM dataset

[EMF20] and thus more stable in practice.

An additional noteworthy observation is the correlation between the smallest singular value and the grasp wrench space volume. The results indicate that an increase in the smallest singular value corresponds to an increase in the grasp wrench space volume. This aligns with theoretical expectations, as a well-conditioned grasp should exhibit a larger wrench space, thereby enhancing force distribution and potentially improving robustness.

4.1 Limitations

SE3-DiffFields, CGDF and GraspLDM demonstrate promising results, as evidenced by the Earth Mover's Distance (EMD), which indicates that the generated grasps are close to high-quality grasps in the original dataset. However, several limitations remain:

- **Low Force Closure Rates**: Despite achieving high success rates, all three models exhibit low force closure rates, indicating that the generated grasps are often kinematically feasible but not necessarily dynamically stable under external disturbances. This suggests that the models may prioritise generating grasps that geometrically align with high-quality examples in the training data rather than ensuring physical robustness.
- **EMD Weighting Bias**: The Earth Mover's Distance (EMD) implementation appears to prioritise rotational alignment over translational proximity. This means that grasps with similar orientations to the training dataset are favoured, even if their positions differ slightly. While this benefits models trained on datasets with stable grasps, it does not fully capture real-world execution feasibility. GraspLDM, despite having comparable EMD values to SE3-DiffFields, often produces grasps that are closer to the object, making them more stable in practical applications.
- **Variability in Robustness**: SE3-DiffFields demonstrates a slightly higher robustness rate compared to CGDF. However, GraspLDM, while exhibiting a lower success rate, produces more stable and physically robust grasps. This suggests a trade-off between maximising grasp success and ensuring real-world grasp stability.
- **Singular Value Variability**: The models exhibit a general trend linking singular values to grasp quality; however, deviations from the mean indicate inconsistencies in grasp stability across different objects. GraspLDM and SE3-DiffFields tends to generate grasps with a slightly higher grasp wrench space volume, whereas CGDF produces more stable but potentially less diverse grasps.

4.2 Comparison Between SE3-DiffFields and CGDF

CGDF introduces the Convolutional Occupancy Network into the SE3-DiffFields architecture to enhance feature extraction. Theoretically, this addition should improve grasp quality by enabling a more comprehensive understanding of object geometry. However, as evidenced by the results, SE3-DiffFields outperforms CGDF. This discrepancy can be attributed to the preprocessing applied to the dataset, which normalises objects to a tabletop size. During this process, many fine-grained features are lost. For small objects, detailed feature extraction may be unnecessary, as their general shape often provides sufficient information for effective grasp generation. Excessive feature extraction can lead to overfitting, as seen in the EMD results, where CGDF performs well for Subset 2 but exhibits high variance in Subset 1. This suggests that the model overfits to specific geometric characteristics.

Furthermore, the robustness and force closure rates of grasps generated by CGDF indicate a diminished ability to produce high-quality grasps. Although the percentage differences appear minor, they reflect a higher failure rate for CGDF-generated grasps. This trend is also evident in the scatter plot of singular values, where CGDF exhibits greater variability in Subset 1 and overfitting in Subset 2. These results suggest that while CGDF's architectural modifications enhance feature extraction, they may also contribute to reduced generalisation capability, particularly when object features are lost during preprocessing.

Overall, these findings highlight the importance of balancing feature extraction and generalisation to achieve both high success rates and stable grasp generation.

4.3 Comparison Between SE3-DiffFields and GraspLDM

Although both SE3-DiffFields and GraspLDM generate successful grasps, their fundamental approaches differ. SE3-DiffFields prioritises maximising the success rate, whereas GraspLDM indirectly produces more stable grasps due to its adherence to the ACRONYM dataset. SE3-DiffFields achieves a higher success rate, as it was effectively trained on all objects without a strict train-test split. However, GraspLDM's training methodology and model concept results in grasps that more closely resemble the stable grasps in ACRONYM. Despite similar EMD values, GraspLDM's grasps are closer to the object, leading to better engagement and stability. This suggests that SE3-DiffFields may favour

rotational similarity over practical grasp feasibility. GraspLDM's adherence to the train-test split enforces better generalisation, whereas SE3-DiffFields, trained on the full dataset, may exploit prior exposure to objects, leading to artificially inflated success rates.

4.4 Comparison Between CGDF and GraspLDM

CGDF extends SE3-DiffFields by incorporating the Convolutional Occupancy Network, which enhances object feature extraction. However, this modification does not necessarily lead to superior performance when compared to GraspLDM. CGDF's additional feature extraction layers aim to improve grasp quality, yet GraspLDM's advantage comes from closely following the ACRONYM dataset. While CGDF performs well in specific cases, its over-reliance on feature extraction may lead to overfitting, as seen in its high EMD variance across subsets. CGDF shows higher variance across different object categories, suggesting that its approach overfits to specific geometric features rather than generalising well. The additional features that are extracted are not too useful for small table top objects, like the ones in the ACRONYM dataset. In contrast, GraspLDM maintains more consistent performance across different shapes, leading to a more robust grasp generation process.

Chapter 5

Conclusion

In this thesis, we explored the performance of SE3-DiffFields, CGDF and GraspLDM in generating feasible and stable grasps. Our findings indicate that while all three models achieve reasonable success rates, their force closure rates remain low, suggesting a trade-off between grasp feasibility and stability. SE3-DiffFields consistently outperforms CGDF in success rate and robustness, indicating that a more general grasping approach might be preferable over an overly complex feature extraction process, especially for small objects. However, CGDF can be used for all kinds of object of all sizes with the right amount of training over a larger dataset. However grasp stability will still remain an issue that would require attention.

One of the key insights from our evaluation is that GraspLDM generates more stable grasps than SE3-DiffFields and CGDF, despite having a lower success rate. This can be attributed to its adherence to the ACRONYM dataset, which inherently prioritises robustness by ensuring grasps are stable under perturbations. Unlike SE3-DiffFields and CGDF, which were trained without a strict train-test split, GraspLDM's structured training process leads to better generalisation and inherently stable grasps. However, all three models exhibit relatively low force closure rates, indicating that while they generate kinematically feasible grasps, they do not always ensure dynamic stability.

Additionally, we observed a strong correlation between the smallest singular value and the grasp wrench space volume, aligning with theoretical expectations. An increase in the smallest singular value corresponds to a larger grasp wrench space, enhancing force distribution and improving robustness. SE3-DiffFields and GraspLDM tend to generate grasps with a slightly higher grasp wrench space volume, whereas CGDF, despite its additional feature extraction layers, demonstrates higher variability and potential overfitting.

Our findings also highlight the limitations of EMD as an evaluation metric. Since EMD prioritises rotational similarity over translational alignment, it may

not fully capture real-world grasp execution feasibility. Although GraspLDM has EMD values similar to SE3-DiffFields, it produces grasps that are closer to the object, making them more physically stable. This suggests that future grasp evaluation metrics should consider both rotational and translational accuracy for a more balanced assessment.

Future models should explicitly incorporate force closure constraints into training objectives to ensure grasps are dynamically stable under external disturbances. A new metric that balances rotational and translational alignment could provide a more holistic evaluation of grasp quality. This can be achieved by manually adding weight to translation and rotation to give them a more balanced approach while calculating EMD using the implementation of [UFPC23]. Moreover, our experiments focused on simulated datasets, but real-world robotic evaluations, including sensor noise and dynamic interactions, would provide a more comprehensive assessment of these models.

In summary, this thesis demonstrates the strengths and weaknesses of diffusion-based grasp generation models and provides insights into their trade-offs. While SE3-DiffFields offers a more generalisable approach for normalised tabletop objects, CGDF's feature extraction method may be better suited for large-object grasping scenarios. GraspLDM, despite a lower success rate, produces inherently stable grasps due to its adherence to the ACRONYM dataset. Future research should focus on improving force closure rates, refining evaluation metrics, and validating these models in real-world robotic environments to bridge the gap between simulated and practical grasp execution.

List of Figures

List of Tables

Bibliography

[BK00] A. Bicchi and V. Kumar. Robotic grasping and contact: a review. In *Proceedings 2000 ICRA. Millennium Conference. IEEE International Conference on Robotics and Automation. Symposia Proceedings (Cat. No.00CH37065)*, volume 1, pages 348–353 vol.1, 2000. doi:10.1109/ROBOT.2000.844081.

[BOR+24] Kuldeep R. Barad, Andrej Orsula, Antoine Richard, Jan Dentler, Miguel A. Olivares-Mendez, and Carol Martinez. GrasplDM: Generative 6-dof grasp synthesis using latent diffusion models. *IEEE Access*, 12:164621-164633, 2024. URL: http://dx.doi.org/10.1109/ACCESS.2024.3492118, doi:10.1109/access.2024.3492118.

[DN21] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. *CoRR*, abs/2105.05233, 2021. URL: https://arxiv.org/abs/2105.05233, arXiv:2105.05233.

[EKLB13] Sahar El-Khoury, Miao Li, and Aude Billard. On the generation of a variety of grasps. *Robotics and Autonomous Systems*, 61(12):1335–1349, 2013. URL: https://www.sciencedirect.com/science/article/pii/S0921889013001437, doi:10.1016/j.robot.2013.08.002.

[EMF20] Clemens Eppner, Arsalan Mousavian, and Dieter Fox. ACRONYM: A large-scale grasp dataset based on simulation. *CoRR*, abs/2011.09584, 2020. URL: https://arxiv.org/abs/2011.09584, arXiv:2011.09584.

[HJA20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *CoRR*, abs/2006.11239, 2020. URL: https://arxiv.org/abs/2006.11239, arXiv:2006.11239.

[LS88] Z. Li and S.S. Sastry. Task-oriented optimal grasping by multifingered robot hands. *IEEE Journal on Robotics and Automation*, 4(1):32–44, 1988. doi:10.1109/56.769.

[MEF19] Arsalan Mousavian, Clemens Eppner, and Dieter Fox. 6-dof graspnet: Variational grasp generation for object manipulation.

CoRR, abs/1905.10520, 2019. URL: http://arxiv.org/abs/1905.10520, arXiv:1905.10520.

[MLN+17] Jeffrey Mahler, Jacky Liang, Sherdil Niyaz, Michael Laskey, Richard Doan, Xinyu Liu, Juan Aparicio Ojea, and Ken Goldberg. Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics. *CoRR*, abs/1703.09312, 2017. URL: http://arxiv.org/abs/1703.09312, arXiv:1703.09312.

[MME+20] Adithyavairavan Murali, Arsalan Mousavian, Clemens Eppner, Chris Paxton, and Dieter Fox. 6-dof grasping for target-driven object manipulation in clutter. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6232–6238, 2020. doi:10.1109/ICRA40945.2020.9197318.

[MSZ94] Richard M. Murray, S. Shankar Sastry, and Li Zexiang. *A Mathematical Introduction to Robotic Manipulation*. CRC Press, Inc., USA, 1st edition, 1994.

[ND21] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021.

[NVH+24] Toan Nguyen, Minh Nhat Vu, Baoru Huang, An Vuong, Quan Vuong, Ngan Le, Thieu Vo, and Anh Nguyen. Language-driven 6-dof grasp detection using negative prompt guidance. In *European Conference on Computer Vision*, pages 363–381. Springer, 2024.

[PNM+20] Songyou Peng, Michael Niemeyer, Lars M. Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. *CoRR*, abs/2003.04618, 2020. URL: https://arxiv.org/abs/2003.04618, arXiv:2003.04618.

[RA15] Joseph Redmon and Anelia Angelova. Real-time grasp detection using convolutional neural networks. In *2015 IEEE international conference on robotics and automation (ICRA)*, pages 1316–1322. IEEE, 2015.

[SDWMG15] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. pmlr, 2015.

[SKK+24] 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, 2024. URL: https://arxiv.org/abs/2404.04643, arXiv:2404.04643.

[SME20] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *CoRR*, abs/2010.02502, 2020. URL: https://arxiv.org/abs/2010.02502, arXiv:2010.02502.

[SMTF21] Martin Sundermeyer, Arsalan Mousavian, Rudolph Triebel, and Dieter Fox. Contact-graspnet: Efficient 6-dof grasp generation in cluttered scenes. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13438–13444, 2021. doi:10.1109/ICRA48506.2021.9561877.

[UFPC23] Julen Urain, Niklas Funk, Jan Peters, and Georgia 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)*, pages 5923–5930. IEEE, 2023.

[VMT17] Matthew Veres, Medhat Moussa, and Graham W. Taylor. Modeling grasp motor imagery through deep conditional generative models. *CoRR*, abs/1701.03041, 2017. URL: http://arxiv.org/abs/1701.03041, arXiv:1701.03041.

License