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

A Hybrid Optimization Framework for Grasp Synthesis under Partial Observations

Wenzheng Zhang[1,*], Fahira Afzal Maken[2], Tin Lai[1], Fabio Ramos[1,3]

***Abstract*— We propose a hybrid grasp synthesis framework that combines a learning-based Energy-Based Model (EBM) with an analytical Iterative Closest Point (ICP) method to generate robust grasps from partially observed point clouds. The learned energy function acts as a prior within a Stein Variational Gradient Descent (SVGD) framework, guiding iterative refinement of grasp configurations. Evaluated on 67 objects with 5,360 grasp attempts, our method achieves an average success rate of 60.9%, outperforming AnyGrasp (31.1%) and Grasp Pose Detection (48.4%) and AS-ICP (56.6%). These results highlight the strong generalization ability of our approach and demonstrate how combining data-driven learning with geometric optimization addresses the limitations of either strategy in isolation.**

I. INTRODUCTION

Grasp synthesis remains a challenging problem in robotic manipulation, particularly for unseen objects under partial observation. Traditional methods relying on complete object models often fail to generalize to the diversity of everyday objects [1]. To address this, we propose a hybrid grasp synthesis framework that integrates a data-driven Energy-Based Model (EBM) [2] with Iterative Closest Point (ICP) [3] within the Stein Variational Gradient Descent (SVGD) framework [4].

Point cloud completion is one strategy for handling partial observations, but generating accurate shapes remains difficult [5]. Although some methods generate grasps directly from partial point clouds [6], they still require full object models to build training datasets [7]. Annealed Stein ICP (AS-ICP) [8] is an optimization-based approach capable of generating grasps from partial inputs. However, the method is sensitive to gripper aperture, requires multiple preshapes to ensure coverage.

To overcome these limitations, we propose a hybrid approach that combines learned global priors with geometric optimization. First, an EBM is trained on data generated by AS-ICP, assigning low energy to successful grasps and high energy to failures. Integrating the EBM gradient into SVGD enables iterative refinement of grasp poses, blending learned grasp-quality cues with local geometric alignment to improve robustness.

We evaluated our approach on 57 Google Scanned Objects and 10 KIT Dataset objects, comprising 5360 grasp attempts. Our method achieves an average success rate of 60.9%, outperforming baselines such as AnyGrasp (31.1%), GPD (48.4%), and AS-ICP (56.6%).

Contributions:

- A hybrid framework that integrates EBMs and ICP within SVGD for robust grasp synthesis from partial point clouds.
- A comprehensive ablation study analyzing model architecture, loss functions, and dataset design.
- Extensive experiments demonstrating improved robustness, repeatability, and generalization compared to analytic, data-driven, and weakly integrated hybrid baselines.

II. RELATED WORK

Grasp synthesis approaches can be broadly categorized into optimization-based and data-driven methods [9].

Optimization-based methods formulate grasp planning as an explicit optimization problem over full-object models. Examples include CPO-PPO [10], ISF [11], MD-ISF [12], and PPO-JPO [13], which jointly optimize gripper pose and finger joints. AS-ICP [8] simplifies the problem by optimizing only the grasp pose, reducing complexity while maintaining robustness.

Data-driven methods learn grasp quality directly from sensory data such as RGB-D or point clouds. DexNet [14] employs multiview CNNs, but still relies on full object models for training. GraspNet [15], AnyGrasp [6], and GPD [7] predict grasps from partial point clouds, though datasets are typically generated using full models. Recent work [16] follows this paradigm, often refining candidates with ICP [17].

Hybrid methods aim to combine analytical reasoning with learning, but most are weakly integrated. Typical strategies include the use of analytical quality metrics during training [18], as evaluation criteria [7], or as post-processing [19].

Finally, Energy-Based Models (EBMs) [2] have shown promise for grasp synthesis due to their flexibility in modeling distributions. However, challenges include intractable partition functions [20] and highly non-convex energy landscapes [21]. Stein Variational Gradient Descent (SVGD) [4] provides an efficient inference strategy and has been successfully applied in robotics, including trajectory [22] and motion [23] planning, and localization [24].

We integrate EBMs with ICP, combining learned priors with geometric optimization under the SVGD framework,

* Corresponding author: wzha2981@uni.sydney.edu.au
[1] School of Computer Science, The University of Sydney, Australia
[2] Data61, CSIRO, Australia
[3] NVIDIA, USA

targeting robust grasp synthesis from partial point clouds.

III. Preliminaries

A. Energy Based Model

Energy-Based Models (EBMs) offer a flexible framework to represent complex probability distributions. In an EBM, each input x is assigned a scalar energy $E_\theta(x)$, where lower energy values indicate more plausible or desirable configurations. The probability density is defined as:

$$p_\theta(x) = \frac{\exp(-E_\theta(x))}{Z_\theta}, \tag{1}$$

where $E_\theta(x)$ is a non-linear energy function such as a deep neural network parameterized by θ, and the partition function Z_θ is given by:

$$Z_\theta = \int \exp(-E_\theta(x))\, dx. \tag{2}$$

Inference with EBMs involves finding the configuration x that minimizes the energy function [2]:

$$\hat{x} = \arg\min_x E_\theta(x). \tag{3}$$

Training EBMs is challenging due to the intractability of the partition function. Various techniques—such as contrastive divergence [25] and score matching [26]—have been proposed to address this difficulty. In our work, we focus on integrating the gradient of the EBM into the Stein Variational Gradient Descent (SVGD) framework to optimize grasp synthesis from partial point cloud data. We used different weights to balance the gradients from EBM and SGD-ICP.

B. Stein Variational Gradient Descent

Stein Variational Gradient Descent (SVGD) [4] is a particle-based variational inference method that approximates a complex target distribution using a set of interacting particles. Unlike traditional variational inference methods that assume a fixed parametric form for the approximate posterior, SVGD represents the target distribution non-parametrically as an empirical distribution over particles.

In the SVGD framework, the jth particle, θ_j, is updated as follows [4]:

$$\theta_j \leftarrow \theta_j + \eta\, \hat{\phi}^*(\theta_j), \tag{4}$$

where η is the step size and the optimal update direction $\hat{\phi}^*(\theta)$ is given by:

$$\hat{\phi}^*(\theta) = \frac{1}{K}\sum_{j=1}^{K}\Big[k(\theta_j,\theta)\,\nabla_{\theta_j}\log p(\theta_j) + \nabla_{\theta_j} k(\theta_j,\theta)\Big]. \tag{5}$$

Here, $k(\cdot,\cdot)$ is a positive definite kernel that couples the particles. The first term can be seen as an attractive force that pulls the particles toward regions of high probability density, and the second term acts as a repulsive force that prevents the particles from collapsing to a single mode, thereby encouraging diversity in the particle set.

IV. Stein Energy-Based Grasp Synthesis

A. Problem Formulation

In this work, we address the problem of grasp synthesis for unknown objects using a hybrid optimization framework that integrates data-driven and analytical approaches. Given the gripper's inner surface point cloud, $\mathcal{S}$, and the object's point cloud captured from a single viewpoint, $\mathcal{R}$, our goal is to determine an optimal grasp pose, parameterized by $\theta = (t, q)$, where $t = x, y, z$ represents the 3D translation and $q = q_w, q_x, q_y, q_z$ represents the unit quaternion (with q_w being the scalar and q_x, q_y, q_z the vector part). Initial poses are uniformly sampled and refined using Stein Variational Gradient Descent (SVGD) to minimize a hybrid cost function, combining geometric alignment and learned grasp-quality assessments.

Formally, we define our grasp optimization as follows:

$$\min_\theta \quad \mathcal{L}(y,x) + E(y,x) \quad s.t. \quad dist(y, SDF(x)) > 0, \tag{6}$$

where $x = T_\theta(\mathcal{S})$ denotes the transformed gripper surface point cloud under pose θ, and $y = \mathcal{R}$ denotes the object's observed point cloud. Here, $\mathcal{L}(y,x)$ represents the geometric loss computed via Iterative Closest Point (ICP) matching [27]. $E(y,x)$ is the learned energy term, computed from an Energy-Based Model (EBM) trained on grasp outcomes (success/failure) evaluated in simulation, reflecting learned probabilistic knowledge of grasp quality. The constraint ensures that the resulting grasp pose is physically plausible and collision-free, enforced by evaluating the distance from the object points to the gripper's Signed Distance Field (SDF). A summary of our method is provided in Figure 1.

B. Data Generation

To build our dataset, we use the AS-ICP algorithm [8] on single-view point clouds of objects. Specifically, we select 57 distinct objects from the Google Scanned Objects dataset [28] and 10 distinct objects from the KIT dataset [29], capture eight different point clouds per object with four different orientations (0, 90, 180 and 270 degree), each with two camera elevations (0.1 and 0.7 m). For each point cloud, AS-ICP is applied to optimize collision-free grasp poses. These optimized poses are subsequently validated in the Isaac Gym simulator [30]. A grasp is considered successful if the gripper maintains a firm hold on the object even after it is subjected to shaking. This process yields a labeled dataset of grasp poses.

C. Network Architecture

Our EBM is designed to capture the interaction between the object and the gripper by processing their point clouds and the associated grasp pose information. The model consists of three primary modules:

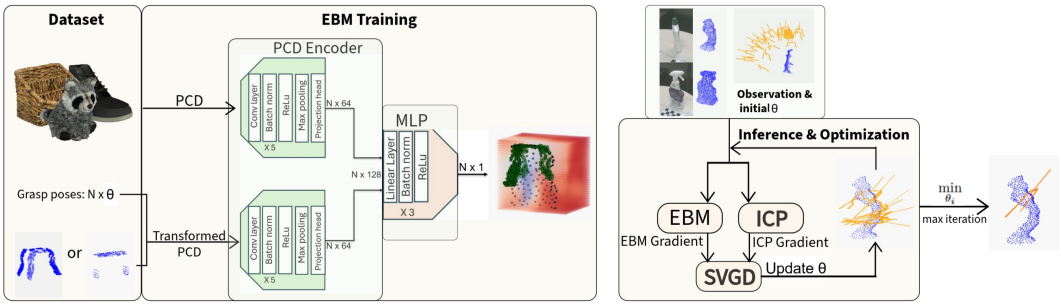

Fig. 1: A brief summary of our algorithm. Both the object and gripper point clouds are fed into separate point cloud encoders to produce 64-dimensional features. These features are then concatenated and scored by the EBM to output a single energy value. The system uses SVGD to iteratively update the transformation parameters θ by leveraging gradients from the EBM and the cost function of ICP. The best pose is selected with minimum energy and matching error.

1) Point Cloud Encoder: We adopt a PointNet-based encoder [31], as point clouds are unordered sets of 3D points without a regular grid structure. PointNet handles this by using permutation-invariant operations, making it well-suited for extracting robust geometric features for grasping. The Point Cloud Encoder module applies five one-dimensional convolutional layers (with kernel size 1), each followed by batch normalization and ReLU activations. After these convolutions, a global max pooling operation aggregates features across all points to produce a fixed-dimensional embedding. This embedding is further projected into a lower-dimensional feature space via a fully connected layer, layer normalization, and a ReLU activation.

2) Energy Function Network: The features are then fed into the Energy Function network, which is implemented as a multi-layer perceptron. This network comprises three fully-connected layers (interleaved with batch normalization and ReLU activations) and culminates in a single linear output that represents the energy.

D. Network Training

1) Loss Function: We employ a contrastive loss function to train our EBM for grasp synthesis. The objective is to assign low energy values to positive grasp samples and high energy values to generated negative samples. We use a smooth, differentiable variant of the hinge loss as follows [2]:

$$L = \log\left(1 + \exp\left(E^{+} - E^{-}\right)\right). \tag{7}$$

Here, E^{+} represents the energy of a positive sample, while E^{-} denotes the energy of a corresponding negative sample. In the context of grasping, even a small deviation in the positive grasp pose may lead to failure, resulting in many subtle negative samples. Thus, preserving this specific pairing is crucial for effective model performance.

2) Dataset: The dataset generated in Section IV-B is highly imbalanced, containing significantly more failure cases than successful grasp poses. This imbalance can negatively impact the performance of our EBM training. In Section V-A, we discuss various data-processing strategies to mitigate such imbalances.

For training, we selected 50 objects. Given that most objects are non-symmetrical, each point cloud captured under different orientations and elevations is treated as a distinct group. From each group, we randomly select 80 successful grasp poses. In cases where a group contains fewer than 80 successful examples, we duplicate the available examples until the total reaches 80. Thus, the total number of positive samples used for training is 32,000.

3) Training: For training our Energy-Based Model (EBM), we used a batch size $\mathcal{N} = 64$ with the Adam optimizer, setting the learning rate to 1×10^{-3} and applying a weight decay of 1×10^{-5}. The model was trained for 25 epochs, and the resulting model was then used to compute the gradient for SVGD using PyTorch's autograd. Our primary goal is not to achieve the best performance from the EBM but rather to demonstrate how the integration of a data-driven approach with optimization can effectively overcome the limitations inherent to each method.

E. SVGD Optimization

In our work, we adapt SVGD within the Annealed Stein ICP (AS-ICP) framework [8] for robust grasp synthesis. In AS-ICP [32], [8], the gradient $\nabla_\theta \log p(\theta)$ in Equation (5) is replaced by the gradient of the SGD-ICP cost function with respect to the transformation parameters [27], [33]:

$$\nabla \log p(\theta) \approx -\gamma(t)\left(N\,\bar{g}(\theta, \mathcal{M}) + \nabla_\theta \log p(\theta)\right), \tag{8}$$

where $\bar{g}(\theta, \mathcal{M})$ denotes the averaged gradient of the ICP cost function over a mini-batch $\mathcal{M}$, N is a normalization factor equal to the number of particles, and $\gamma(t)$ is the annealing schedule. The prior gradient term, $\nabla_\theta \log p(\theta)$, is computed using Gaussian priors for translation and von Mises priors for rotation. However, in practical grasping scenarios, assuming a uniform prior over grasp poses fails to improve optimization performance, as it does not reflect the inherent structure of feasible grasps. To overcome this limitation, we propose augmenting the Stein ICP framework with a learned model that provides a more informative prior. Specifically, we replace the prior gradient $\nabla_\theta \log p(\theta)$ in Equation (8) with the gradient

derived from an Energy-Based Model (EBM), denoted by $\nabla_\theta E(y,x)$:

$$\nabla \log p(\theta) \approx -\gamma(t)\,(w \times N\,\bar{g}(\theta,\mathcal{M}) + \nabla_\theta E(y,x)). \quad (9)$$

We observed that the magnitudes of the learned EBM gradient $\nabla_\theta E(y,x)$ and the matching gradient $N\bar{g}(\theta,\mathcal{M})$ differ significantly. This discrepancy can cause imbalanced updates during optimization, potentially degrading performance. To address this issue, we introduce a dynamic weighting factor w:

$$w = \frac{\|\nabla_\theta E(y,x)\|}{\|N\bar{g}(\theta_k,\mathcal{M})\|}. \quad (10)$$

We use the RBF kernel for translations and the dot product kernel for rotations. For the dot product kernel, the bandwidth parameter was set using the median heuristic [34]. However, in our experiments, we found that setting a fixed kernel bandwidth of $\sigma = 3$ for translation consistently yielded better results than the median heuristic.

We utilize the Adam optimizer with a learning rate of 1×10^{-2} for SVGD parameter updates. Although the optimization requires a large number of iterations to converge, further increases in the learning rate result in parameter divergence and numerical instability. Additionally, we note that the ICP matching gradient provides strong guidance for translation parameters when the gripper is distant from the object, but its magnitude diminishes significantly as the gripper approaches the object's surface.

To take advantage of this behavior, we explicitly incorporate the matching gradient as a regularization term to accelerate early-stage translation optimization, gradually reducing its influence in later iterations. The final adaptive update rule for the transformation parameters is as follows:

$$\theta_{t+1} = \theta_t + \hat{\phi}_t^*(\theta_k) + \left(1 - \frac{k}{K}\right)\bar{g}(\theta_k,\mathcal{M}), \quad (11)$$

where $\hat{\phi}_t^*(\theta_k)$ denotes the SVGD update, and the adaptive term $\left(1-\frac{k}{K}\right)$ progressively decreases the regularization effect of the matching gradient as the iteration index k progresses toward the maximum iteration K.

At the end of the optimization process, we evaluate the matching error for each particle by selecting the minimum non-collision error across a predefined set of gripper preshapes, following the procedure described in [8]. The complete optimization procedure is summarized in Algorithm 1.

Algorithm 1: Stein Energy-Based Grasp

Input: Point cloud of: Gripper $\mathcal{S} = \{s_i\}_{i=1}^N$, Target Object $\mathcal{R} = \{r_i\}_{i=1}^M$, initial parameters $\Theta_0 = \{\theta_0^j\}_{j=1}^J$, number of iteration K, SDF of gripper preshapes
Output: θ that minimizes the sum of energy and matching error

```
while k ≤ K do
    for each θ_k^j ∈ Θ_k in parallel do
        S_k^j ← Transform the source cloud with θ_k^j
        ∇_θ E(y,x)^j and ḡ(θ_k, M)^j ← Compute Stein
          variational gradient (9) then (5)
        Collision avoidance using SDF
        Update θ (11)
    end
    k = k + 1
end
for each θ^j ∈ Θ_K do
    matching error^j ← min(matching error^j_preshapes)
end
return θ = argmin_θj (norm(energy) + norm(matching error))
```

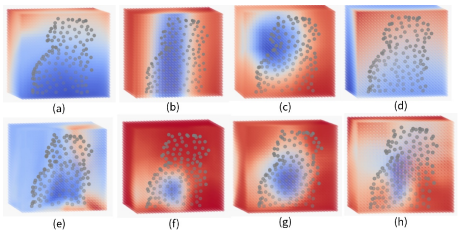

Fig. 2: Energy landscapes for a top-down grasp under different training schemes. Blue indicates low energy (favorable grasps) and red indicates high energy (unfavorable grasps). Panels (a,e) use the full dataset; (b,f) use only positives with an equal number of negatives; (c,g) use negatives generated by uniform sampling; and (d,h) balance positives across groups (object orientation/elevation), yielding the best results. The top row excludes the gripper point cloud, while the bottom row includes it, leading to markedly improved performance.

V. Experiments

A. Ablation

Data Processing

Raw Data. Training on the imbalanced AS-ICP dataset (28,555 positive vs. 72,461 total) limited the effectiveness of the contrastive loss, yielding poor separation between positive and negative examples (Figures 2(a), (e)). Balancing positives with an equal number of random negatives improved balance but flattened the energy field, making the model object-invariant (Figures 2(b), (f)).

Localized Negative Sampling. Generating negatives around each positive improved separation and introduced collision cases absent from the raw dataset. However, vertical grasp positions were poorly captured due to bias in the AS-ICP data (Figures 2(c), (g)).

Group-based Sampling. Partitioning data by orientation and elevation, balancing positives per group, and sampling localized negatives produced the best energy distributions (Figures 2(d), (h)).

Including the gripper point cloud as input further improved results (bottom row of Figure 2). However, even in the best configuration, energy minima were biased downward, reflecting dataset limitations—particularly the underrepresentation of top-down grasps and grasps near the table surface.

Overall, these results highlight that dataset distribution matters more than dataset size: well-structured training data yields a more accurate and reliable energy landscape, even with fewer examples.

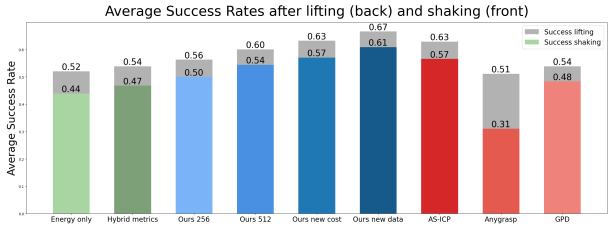

Fig. 3: Plots of average success rate after lifting and shaking. The results demonstrate a clear trend of increasing performance as the model improves. Our method outperforms individual analytical and data-driven methods, other hybrid approaches, and baseline methods.

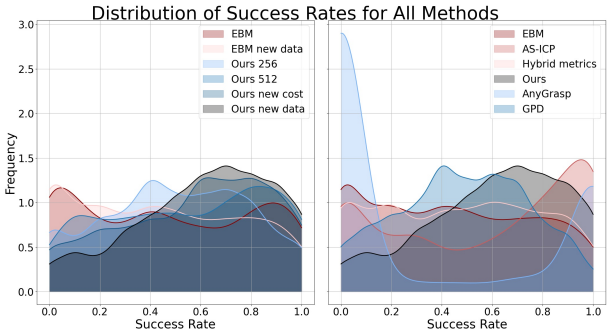

Fig. 4: Kernel density estimates (KDEs) of success rates for all evaluated methods. The KDEs transform discrete success rate measurements into smooth, continuous probability distributions. **Left:** Hybrid model variants with different architectures and training data. **Right:** Baselines including analytical, learning-based, and hybrid approaches. Our model achieves sharper, higher-density peaks near high success rates, indicating improved reliability over baselines.

Hybrid Model

The main focus of this work is to introduce and evaluate a hybrid optimization framework for grasp synthesis. While ICP and EBM are used as representative analytic and learning-based components, the framework is modular and can be extended to other methods, though a full exploration of alternatives is beyond the present scope. The aim is not to train the single best EBM, but to demonstrate that as the learned model improves, grasp performance correspondingly increases.

To evaluate the framework, we conduct ablation studies and systematically compare it against analytical (ICP), learning-based (EBM, GPD, AnyGrasp), and hybrid baselines (Hybrid metrics). A common strategy is to apply ICP as post-processing to learning-based grasps. However, as shown in [8], if the initial grasps are poor, ICP refinement provides little improvement.

Our ablations examine both model architecture and dataset composition.

- **Dataset 1:** 32,000 positive examples from AS-ICP. Models include:
  - **EBM:** Baseline (256-dim encoder, energy only).
  - **Ours 256:** Adds matching error.
  - **Ours new cost:** Adds two extra loss terms.
  - **Ours 512:** Larger encoder (512-dim) with combined point cloud features.
- **Dataset 2:** 3,280 positives generated by "Ours new cost". Models include:
  - **EBM new data:** Energy only (256-dim).
  - **Ours new data:** Energy (256-dim) + matching error + additional losses.

Method	Mean (↑)	Std (↓)	≥ 0.1 (↑)	≥ 0.5 (↑)	≥ 0.9 (↑)
Energy only	0.439	0.318	0.756	0.381	0.056
Hybrid metrics	0.468	0.312	0.795	0.431	0.054
Ours 256	0.500	0.285	0.864	0.451	0.057
Ours 512	0.544	0.310	0.851	0.513	0.093
Ours new cost	0.571	0.291	0.890	0.586	0.082
Ours new data	**0.609**	0.270	**0.920**	**0.623**	0.095
AS-ICP	0.566	0.367	0.784	0.567	0.192
AnyGrasp	0.311	0.440	0.347	0.306	**0.254**
GPD	0.484	**0.259**	0.873	0.424	0.022

TABLE I: Comparison of methods by group-level mean, standard deviation, and fraction of groups exceeding selected success thresholds. The arrows indicate whether higher (↑) or lower (↓) values are preferred. Thresholds (0.1, 0.5, 0.9) denote the fraction of groups whose success rate is above 10%, 50%, and 90%, respectively, providing a sense of performance across low, moderate, and high success regimes.

The additional loss terms used for Ours new cost are defined as follows:

$$\text{PartialError}(R_i) = \text{sigmoid}\left(10\left(\left|(R_i \mathbf{e}_z)_y\right| - 0.5\right)\right), \tag{12}$$

where $\text{sigmoid}(x) = \frac{1}{1+e^{-x}}$, R_i is the i-th rotation matrix, and $\mathbf{e}_z = [0,0,1]^T$, which penalizes near-horizontal grasps prone to unseen collisions.

$$\text{PointInGrasp} = \exp\left(-\frac{1}{10}\sum_{i=1}^{N}(p_i \in \mathcal{B})\right), \tag{13}$$

where $\mathcal{B}$ is the volume within the grasp (i.e., the region enclosed by the gripper), which encourages object points within the gripper volume.

Figure 3 and Figure 4 (left) show progressive improvement: the raw EBM produces a wide spread of success rates, while adding costs and model capacity yields more peaked, consistent outcomes. Refining the dataset further boosts performance, with "Ours new data" achieving the highest density near high success rates.

By contrast, Figures 4 (right) benchmarks performance against baselines. AS-ICP and AnyGrasp show extreme outcomes; GPD produces a spread closer to our hybrids. Overall, our framework yields higher mean success and tighter distributions, though—as Table I shows—it produces fewer near perfect (90%) cases, reflecting the stochasticity of learned components.

In Figure 5, grasp diversity is plotted in terms of translation (x-axis) and rotation (y-axis), computed from pairwise distances between successful grasps. Translation is measured by standard deviation, while rotation differences are weighted more heavily for nearby poses. Because results are drawn from the best of ten trials, these plots reflect repeatability.

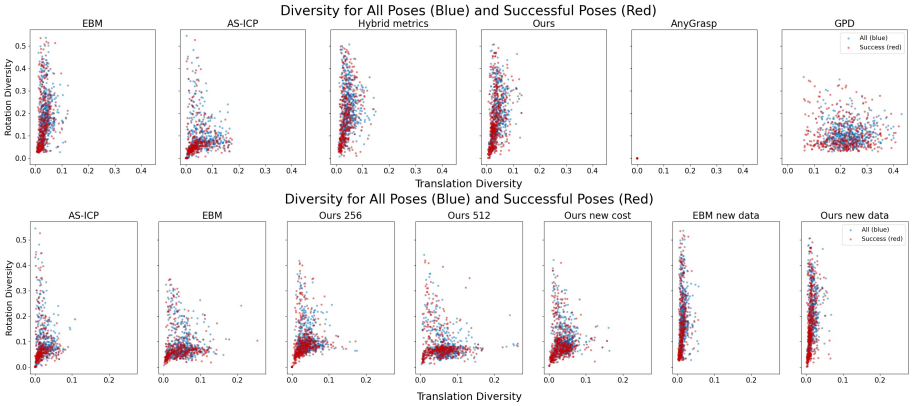

Fig. 5: Comparison of grasp diversity across baseline and hybrid methods. Translation diversity (x-axis, up to 98th percentile) and rotation diversity (y-axis) are shown for all attempted (blue) and successful (red) grasps. Among baselines, AnyGrasp produces identical poses, GPD yields high translation but low rotation diversity, AS-ICP concentrates at low diversity, and EBM shows high rotational diversity. Hybrid models generally mirror their underlying EBM: the initial dataset leads to high translation diversity, while the refined dataset reduces translation but substantially increases rotational diversity.

Among baselines (top), GPD exhibits high translation but low rotation diversity, AnyGrasp produces nearly identical poses, and AS-ICP clusters at low diversity. EBMs spread more broadly, especially in rotation. Hybrid models (bottom) inherit the behavior of their learned components: the initial EBM-trained model resembles ICP, while the refined "Ours new data" variant shows lower translation but higher rotation diversity, yielding consistent contact points with flexible approach angles.

Overall, dataset refinement and model design improve repeatability, whereas purely analytic or sampling-based methods lack consistency, and generative models risk failure by producing identical grasps. The hybrid framework achieves a favorable balance of stability and adaptability.

B. Simulation

We selected 67 objects from the Google Scanned Objects dataset and captured their point clouds in the Isaac Gym simulator [30]. Some objects were rescaled to ensure that the Franka robotic arm could grasp them securely—large enough for a stable hold, yet not so small as to fit entirely within the gripper. All training and optimization experiments were conducted on a laptop equipped with an RTX 2070 GPU. The grasp pose's origin is aligned with the known object origin. In real experiments, the origin is approximated by the centroid of the segmented object point cloud, computed from a stable reference view to avoid inconsistencies caused by occlusions and changing viewpoints.

For evaluation, we compared our approach with three baselines—AnyGrasp [6], Grasp Pose Detection (GPD) [7], and AS-ICP [8]—all of which generate grasps from partial point clouds. For each method, a set of candidate grasps was generated, the top-scoring pose was executed, and this was repeated ten times to compute the average success rate.

Simulation results for the 67 objects are shown in Figure 6. Our method achieved an average success rate of 60.9% over 5,360 grasps, outperforming AnyGrasp (31.1%), GPD (48.4%), and AS-ICP (56.6%). We also achieved higher object-based mean minimum and maximum success rates, as indicated by the error bars. This differs from our previous analysis, which was group-based. However, conventional object-level analysis is less meaningful in our setting, since partial point clouds from different viewpoints can differ significantly (Figure 7). For example, Figure 8 shows that there is no obvious visual distinction between object categories with less than 50% success rate and those above 70%, suggesting that occlusion and viewpoint-specific visibility are key factors.

Our approach also demonstrates strong generalization ability: although the model was trained on only the first 41 objects, it achieved a higher success rate on unseen objects (64.6%) than on the training set (58.6%).

Ours (Franka)	**Time (s)**
SVGD with ICP only	0.653
Initial EBM loading	1.17
Energy and gradient computation	0.458
Overall	2.28
AS-ICP (Franka)	**Time (s)**
SVGD followed by SGD	2.42

TABLE II: Computation time breakdown for the hybrid method compared to original AS-ICP.

As shown in Table II, the use of a single preshape during optimization significantly reduces ICP computation time (0.653 s) compared to AS-ICP (2.42 s), which used 7 preshapes to ensure a high success rate. The complete hybrid optimization process (including model loading time and optimization time) requires approximately 2.28s, demonstrating that it is faster than the AS-ICP method even when including the overhead of loading the EBM.

C. Real Experiment

To validate our algorithm, we conducted experiments using a KG3 gripper mounted on a Kinova arm. We

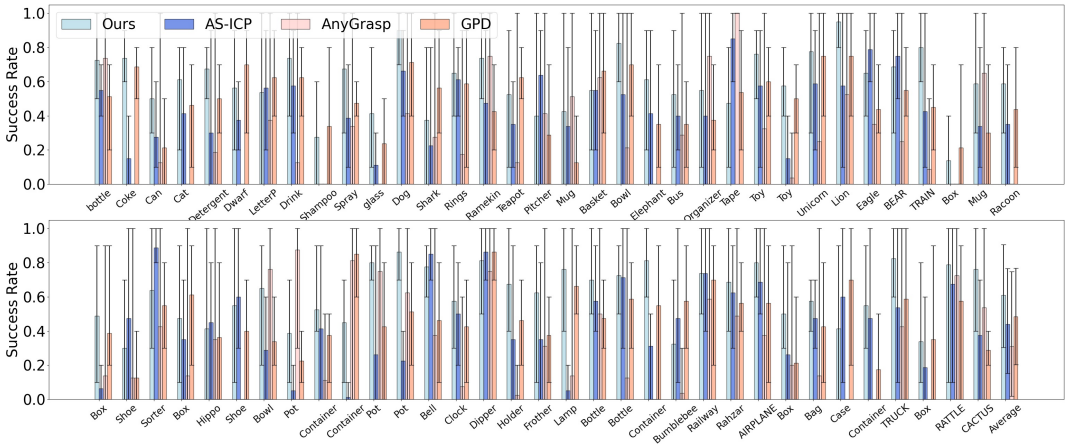

Fig. 6: Simulation results comparing our approach with baseline methods. Our method achieved an average success rate of 60.9%, outperforming AnyGrasp (31.1%), GPD (48.4%), and AS-ICP (56.6%). We also achieved higher object-based (not group-based) minimum and maximum success rates, as indicated by the error bars.

Object	Detergent	Washing Liquid	Spray	Holder	Bag	Mouse	Glass Container	Box	Water Bottle	Hand Help Tool	**Avg.**
AS-ICP	0.80	1.00	0.80	0.60	0.40	0.20	0.60	0.40	0.80	0.80	**0.64**
Ours	0.80	1.00	1.00	0.60	0.60	0.40	0.60	1.00	1.00	0.80	**0.78**

TABLE III: Comparison of average success rates per object between AS-ICP and our method.

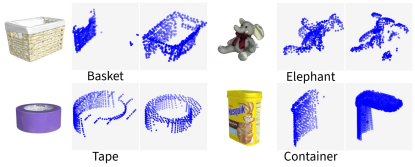

Fig. 7: Examples of objects and their corresponding point clouds.

Fig. 8: Examples of objects used in simulation for which our method achieves less than 50% success rate (left) and more than 70% (right). There is no obvious visual distinction between object categories with less than 50% success rate and those above 70%, suggesting that occlusion and viewpoint-specific visibility are key factors.

selected ten everyday objects and placed each in five different orientations, capturing the point clouds with a wrist-mounted camera. Overall, we achieved a 78% success rate across fifty grasps, as summarized in Table III. Figure 9 displays a single view of each object along with the corresponding point clouds and generated grasps.

Our algorithm is robust to noisy and partial point clouds. However, it struggles when the observed point cloud deviates significantly from the true object geometry—for instance, with the Holder (d) and Mouse (f), where table noise reduces accuracy, and the transparent Glass Container (g), where poor sensing leads to failures. In contrast, the Box (h) shows clear improvement over AS-ICP, highlighting the method's strength in reducing misalignment.

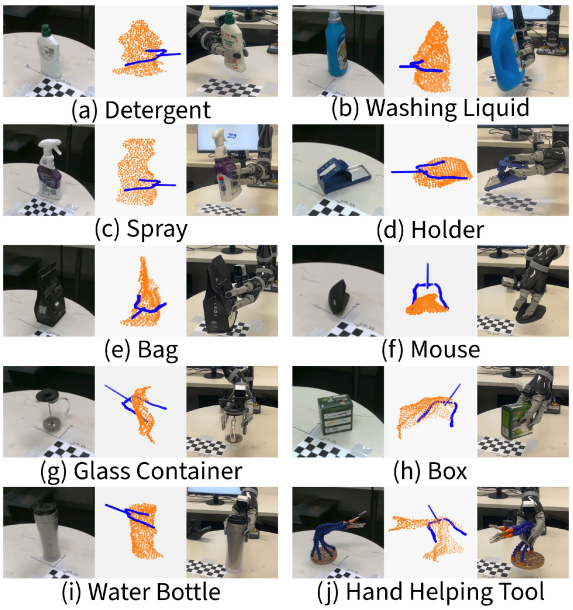

Fig. 9: Illustration of the real experiment. For each object: left—camera view, middle—scanned point cloud with predicted grasp, right—KG3 gripper executing the grasp.

VI. Summary and Discussion

Through ablation studies, we demonstrated the critical role of structured datasets in training EBMs and showed that the hybrid formulation consistently outperforms each component in isolation. Experimental results confirm that our method achieves high success rates on both seen and unseen objects.

Several promising directions emerge from this work. First, because AS-ICP is gripper-invariant, our hybrid

framework inherits this property. Training the EBM on data from multiple grippers would enable grasp synthesis across different end-effectors without architectural changes. Second, our method is not a competitor to existing approaches, but a modular framework into which other learning-based components can be integrated—such as alternative grasp synthesis networks or other analytical metrics. Third, our pipeline—generating data through an analytical method (AS-ICP), trains an EBM, then hybrid optimization—mirrors, in simplified form, the broader paradigm of simulation-based training. This perspective exposes a deeper opportunity: if simulators themselves are collections of analytical methods, then our hybrid framework—where a learned model augments the very method that generated its training data—suggests a path toward closed-loop refinement between simulation, learning, and optimization.

References

[1] 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.

[2] A. Dawid and Y. LeCun, "Introduction to latent variable energy-based models: A path towards autonomous machine intelligence," *arXiv preprint arXiv:2306.02572*, 2023. Les Houches Summer School Lecture Notes 2022 Preprint.

[3] P. J. Besl and N. D. McKay, "A method for registration of 3-d shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1992.

[4] Q. Liu and D. Wang, "Stein variational gradient descent: A general purpose bayesian inference algorithm," in *Proc. Neural Inf. Process. Syst.*, 2016.

[5] M. Kiatos, S. Malassiotis, and I. Sarantopoulos, "A geometric approach for grasping unknown objects with multifingered hands," *IEEE Transactions on Robotics*, vol. 37, no. 3, pp. 735–746, 2021.

[6] 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*, 2023.

[7] A. ten Pas, M. Gualtieri, K. Saenko, and R. Platt, "Grasp pose detection in point clouds," *The International Journal of Robotics Research*, vol. 36, no. 13-14, pp. 1455–1473, 2017.

[8] W. Zhang, F. A. Maken, T. Lai, and F. Ramos, "Grasping by parallel shape matching," in *ACRA 2024*, 2024.

[9] K. Kleeberger, R. Bormann, W. Kraus, and M. F. Huber, "A survey on learning-based robotic grasping," *Current Robotics Reports*, vol. 1, no. 4, pp. 239–249, 2020.

[10] Y. Fan, T. Tang, H.-C. Lin, and M. Tomizuka, "Real-time grasp planning for multi-fingered hands by finger splitting," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1571–1576, IEEE, 2018.

[11] Y. Fan, H.-C. Lin, T. Tang, and M. Tomizuka, "Grasp planning for customized grippers by iterative surface fitting," in *2018 IEEE 14th International Conference on Automation Science and Engineering (CASE)*, pp. 1445–1450, IEEE, 2018.

[12] Y. Fan and M. Tomizuka, "Efficient grasp planning and execution with multi-fingered hands by surface fitting," *IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp. 3995–4002, 2019.

[13] Y. Fan, X. Zhu, and M. Tomizuka, "Optimization model for planning precision grasps with multi-fingered hands," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1548–1554, 2019.

[14] J. Mahler, F. T. Pokorny, B. Hou, M. Roderick, M. Laskey, M. Aubry, K. Kohlhoff, T. Kröger, J. Kuffner, and K. Goldberg, "Dex-net 1.0: A cloud-based network of 3d objects for robust grasp planning using a multi-armed bandit model with correlated rewards," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1957–1964, 2016.

[15] H.-S. Fang, C. Wang, M. Gou, and C. Lu, "Graspnet-1billion: A large-scale benchmark for general object grasping," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11441–11450, 2020.

[16] X. Liu, C. Huang, J. Li, W. Wan, and C. Yang, "Two-stage grasp detection method for robotics using point clouds and deep hierarchical feature learning network," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 16, no. 2, pp. 720–731, 2024.

[17] E. Chisari, N. Heppert, T. Welschehold, W. Burgard, and A. Valada, "Centergrasp: Object-aware implicit representation learning for simultaneous shape reconstruction and 6-dof grasp estimation," *IEEE Robotics and Automation Letters*, 2024.

[18] J. Mahler, J. Liang, S. Niyaz, M. Laskey, R. Doan, X. Liu, J. A. Ojea, and K. Goldberg, "Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics," *arXiv preprint arXiv:1703.09312*, 2017.

[19] 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)*, pp. 13438–13444, IEEE, 2021.

[20] Y. Du and I. Mordatch, "Implicit generation and modeling with energy-based models," in *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[21] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," in *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[22] Z. Yin, T. Lai, S. Khan, J. Jacob, Y. Li, and F. Ramos, "Stein movement primitives for adaptive multi-modal trajectory generation," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 11901–11908, 2024.

[23] Z. Yin, T. Lai, L. Barcelos, J. Jacob, Y. Li, and F. Ramos, "Diverse motion planning with stein diffusion trajectory inference," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 15610–15616, IEEE, 2025.

[24] F. A. Maken, F. Ramos, and L. Ott, "Stein particle filter for nonlinear, non-gaussian state estimation," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 5421–5428, 2022.

[25] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Computation*, vol. 14, no. 8, pp. 1771–1800, 2002.

[26] A. Hyvärinen and P. Dayan, "Estimation of non-normalized statistical models by score matching.," *Journal of Machine Learning Research*, vol. 6, no. 4, 2005.

[27] F. A. Maken, F. Ramos, and L. Ott, "Speeding up iterative closest point using stochastic gradient descent," in *2019 International Conference on Robotics and Automation (ICRA)*, pp. 6395–6401, 2019.

[28] L. Downs, A. Francis, N. Koenig, *et al.*, "Google scanned objects: A high-quality dataset of 3d scanned household items," in *International Conference on Robotics and Automation*, pp. 2553–2560, 2022.

[29] A. Kasper, Z. Xue, and R. Dillmann, "The kit object models database: An object model database for object recognition, localization, and manipulation in service robotics," *International Journal of Robotics Research*, 2012.

[30] V. Makoviychuk *et al.*, "Isaac gym: High performance gpu-based physics simulation for robot learning." https://neurips.cc/datasets-benchmarks/2021, 2021.

[31] R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *IEEE Conference of Computer Vision and Pattern Recognition (CVPR)*, 2017.

[32] F. A. Maken, F. Ramos, and L. Ott, "Stein icp for uncertainty estimation in point cloud matching," *Robotics and Automation Letters*, 2022.

[33] F. Maken, F. Ramos, and L. Ott, "Bayesian iterative closest point for mobile robot localization," *The International Journal of Robotics Research*, vol. 41, no. 9-10, pp. 851–874, 2022.

[34] D. Garreau, W. Jitkrittum, and M. Kanagawa, "Large sample analysis of the median heuristic," *arXiv preprint arXiv:1707.07269*, 2017.