DAGDiff: Guiding Dual-Arm Grasp Diffusion to Stable and Collision-Free Grasps

***Abstract*— Reliable dual-arm grasping is essential for manipulating large and complex objects but remains a challenging problem due to stability, collision, and generalization requirements. Prior methods typically decompose the task into two independent grasp proposals, relying on region priors or heuristics that limit generalization and provide no principled guarantee of stability. We propose DAGDiff, an end-to-end framework that directly denoises to grasp pairs in the $SE(3) \times SE(3)$ space. Our key insight is that stability and collision can be enforced more effectively by guiding the diffusion process with classifier signals, rather than relying on explicit region detection or object priors. To this end, DAGDiff integrates geometry-, stability-, and collision-aware guidance terms that steer the generative process toward grasps that are physically valid and force-closure compliant. We comprehensively evaluate DAGDiff through analytical force-closure checks, collision analysis, and large-scale physics-based simulations, showing consistent improvements over previous work on these metrics. Finally, we demonstrate that our framework generates dual-arm grasps directly on real-world point clouds of previously unseen objects, which are executed on a heterogeneous dual-arm setup where two manipulators reliably grasp and lift them. Project Page: dag-diff.github.io/dagdiff/**

I. Introduction

Manipulating large, dual-arm relevant objects such as monitors, boxes, or chairs requires not only feasible grasps, but also reasoning about force balance and stable interaction between both arms. Imagine the task of picking up a monitor. Humans instinctively place their hands on the opposite sides of the monitor instead of grasping it at random points to balance forces and torques, ensuring stability. For robots, however, acquiring this kind of coordination is far more complex [1]. Developing this sense of dual-arm stability is essential for moving beyond single-arm grasping toward coordinated, physically robust manipulation of real-world objects [2]–[5].

While grasp pose generation has been explored extensively in the community, most efforts largely focus on single-arm settings. Most methods [6]–[11] follow a general recipe of curating a paired dataset consisting of ground truth grasps evaluated using physics simulators, followed by training encoder-decoder style models in a supervised setting. Recently, diffusion models have emerged as powerful generative frameworks for robotic grasping [12]–[16] due to their ability to model complex multimodal distributions. This enables them to sample diverse, high-quality grasp poses either uniformly across the object or constrained to specific parts.

While these methods have improved robustness and grasp quality on complex shapes, they are designed for single-arm grasps and lack mechanisms to ensure dual-arm stability. Moreover, extending these methods to dual-arm grasping is non-trivial, as exhaustive pair search is costly and naive single-arm extensions often yield unstable solutions [17]. Furthermore, these methods do not explicitly account for collisions and often produce grasps that intersect the object surface, a problem that becomes increasingly critical for larger and more geometrically complex shapes. A possible workaround would be to increase the resolution of the point cloud or latent representation to capture finer surface details, but this would greatly increase computation without guaranteeing collision-free grasps.

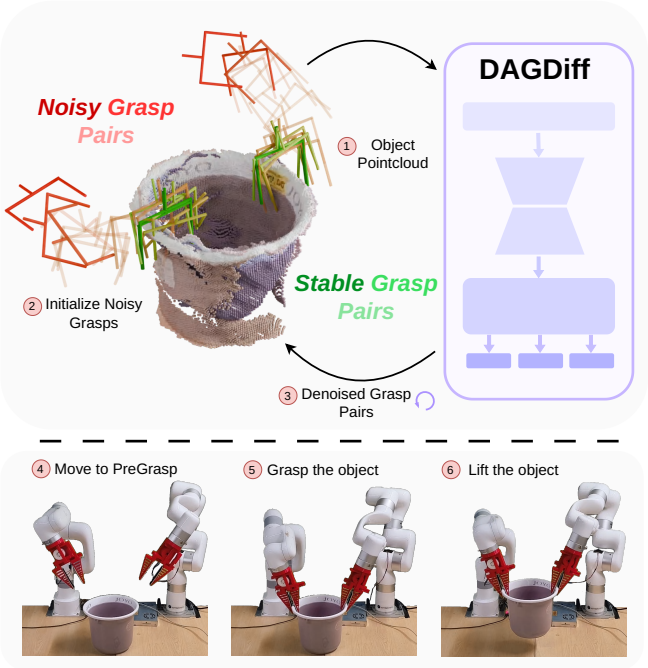

Fig. 1: We introduce DAGDiff: Dual-Arm Grasp Diffusion, a diffusion framework in $SE(3) \times SE(3)$ that takes an object point cloud and denoises noisy grasp pairs (in shades of red) into stable, collision-free dual-arm grasps (in shades of green), guided by multi-head outputs. These predicted grasps are further validated through real-world dual-arm executions, where objects are grasped and lifted successfully.

In this work, we introduce **DAGDiff: Dual-Arm Grasp Diffusion**, an end-to-end dual-arm grasp generation framework that leverages diffusion models guided by classifier signals. Our method extends $SE(3)$ diffusion to the dual-arm setting to generate grasp pairs that are simultaneously stable under dual-arm force closure and collision free with respect to the object's surface.

We frame dual-arm grasp generation as the task of generating two grasps on the object point cloud, each falling in a suitable region, such that they are jointly stable and

physically valid (by physically valid we mean grasps that are collision free, and make stable surface contact with the object). One of the key challenges is region selection: heuristic approaches such as choosing farthest regions [12] often fail when those regions itself are physically incompatible, while VLM-based reasoning [18] remains limited in 3D and physical understanding [19], [20]. It is further hindered by the fact that graspable regions rarely have semantic names, leaving no reliable basis for prediction. In contrast, our approach does not rely on region-specific training but instead learns suitable grasp regions implicitly from guidance signals, and we observe that it naturally discovers the right pairs of regions for stable dual-arm grasps (given in supplementary video). Specifically, a **force-closure module** distinguishes stable from unstable grasp pairs and provides gradients that bias the diffusion process toward stability, while a **collision module** identifies grasp–object intersections and pushes generated grasps away from collisions. Together, these signals guide the diffusion model to diverse, stable, and physically valid dual-arm grasps.

Our evaluation demonstrates the effectiveness of the proposed method in generating stable grasps within a dual-arm setup. Analytical evaluation based on dual-arm force-closure criteria [17] confirms that the generated grasp pairs satisfy fundamental stability requirements, while physics simulation-based evaluation [21] highlights the robustness of our approach across diverse objects and grasp configurations. Finally, real-world demonstrations show that our framework, trained entirely on synthetic data, transfers effectively to real point clouds, producing physically realizable dual-arm grasps on previously unseen objects like cooking utensils, buckets, drones etc as shown in Figure 1. To summarize the contributions:

1) We present **DAGDiff**, the first framework to the best of our knowledge, for dual-arm grasp generation that extends $SE(3)$ diffusion with guidance signals, enabling the synthesis of grasp pairs that are both force-closure stable and collision-free on large, geometrically complex objects.
2) Unlike prior methods that rely on region identification using VLMs or geometric heuristics, our architecture employs geometry-, stability-, and collision-aware multi-head outputs that directly guide the diffusion process toward valid regions of the dual-arm grasp space (Figure 2).
3) We show substantial improvements over prior methods and adapted baselines through analytical metrics and large-scale simulations (Table I), and further validate reliable zero-shot transfer on a heterogeneous real-world dual-arm setup with real point clouds and previously unseen objects (Figure 1).

II. Related Works

A. Dual-Arm Grasping and Stability

Dual-arm grasping requires two parallel-jaw grasps that are not only individually stable but also jointly satisfy stability criteria such as force-closure, i.e., the contact forces must counteract any external wrench on the object [22]. Mesh-based approaches like [17], [23] address this by densely sampling single-arm grasps on object meshes and evaluatin g all grasp pairs with a dual-arm force-closure test. While this ensures stability in simulation, the reliance on complete meshes and exhaustive pair evaluation makes these methods impractical for real-world perception and deployment. To reduce this combinatorial complexity, CGDF [12] employs a part-guided diffusion strategy that generates grasps in the two farthest regions of the point cloud, forming dual-arm pairs. Further, UniDiffGrasp [18] extends this idea by incorporating a VLM to identify object parts for dual-arm grasping. However, these methods still treat the problem as combining two single-arm proposals without explicitly enforcing joint stability. In contrast, DualAfford [24] directly predicts dual-arm grasp poses through collaborative affordance learning, generating one gripper's grasp conditioned on the other. While this captures inter-gripper dependencies, the method relies on object category-specific training and an intricate pipeline, which limits its generalization. To overcome these issues, we propose an end-to-end diffusion framework to implicitly learn the distribution of stable dual-arm grasps without relying on external region-proposals or object-centric pipelines.

B. Diffusion Models for Grasp Generation

Diffusion models, which are particularly suited for capturing multimodal distributions, have emerged as a powerful alternative to previous classical as well as deep-learning based methods for grasp synthesis [7]–[9], [11], [22], [25]. SE3Diff [13] introduced diffusion in the $SE(3)$ space for sampling diverse single-arm grasps, and [26] implemented this idea for partial point clouds along with refinement using collision spheres. CGDF [12] extended diffusion in $SE(3)$ with improved feature representations for constrained grasping on complex shapes. More recently, [14] combined diffusion with transformers to scale grasp generation to large datasets with strong sim-to-real performance. Beyond single-arm settings, diffusion has also been explored for dexterous and multi-fingered hands, with recent works [15], [16], [27] demonstrating its effectiveness for generating stable and generalizable grasps in high-DOF gripper settings. Together, these approaches demonstrate the flexibility of diffusion for both simple and high-DOF grasp generation. However, they do not naturally extend to the dual-arm setting, where the challenge is not only generating individually valid grasps but also ensuring their joint stability. In this work, we address this challenge by introducing guidance signals that steer the diffusion process toward grasp pairs that are both feasible and dual-arm stable.

III. Methods

Given an object point cloud $P \in \mathbb{R}^{n\times 3}$, our goal is to generate M pairs of force closure stable and collision-free parallel-jaw grasp poses $H_i = (H_{i,1}, H_{i,2}) \in SE(3) \times SE(3), i \in [1, M]$ on P. The feasibility of H is determined

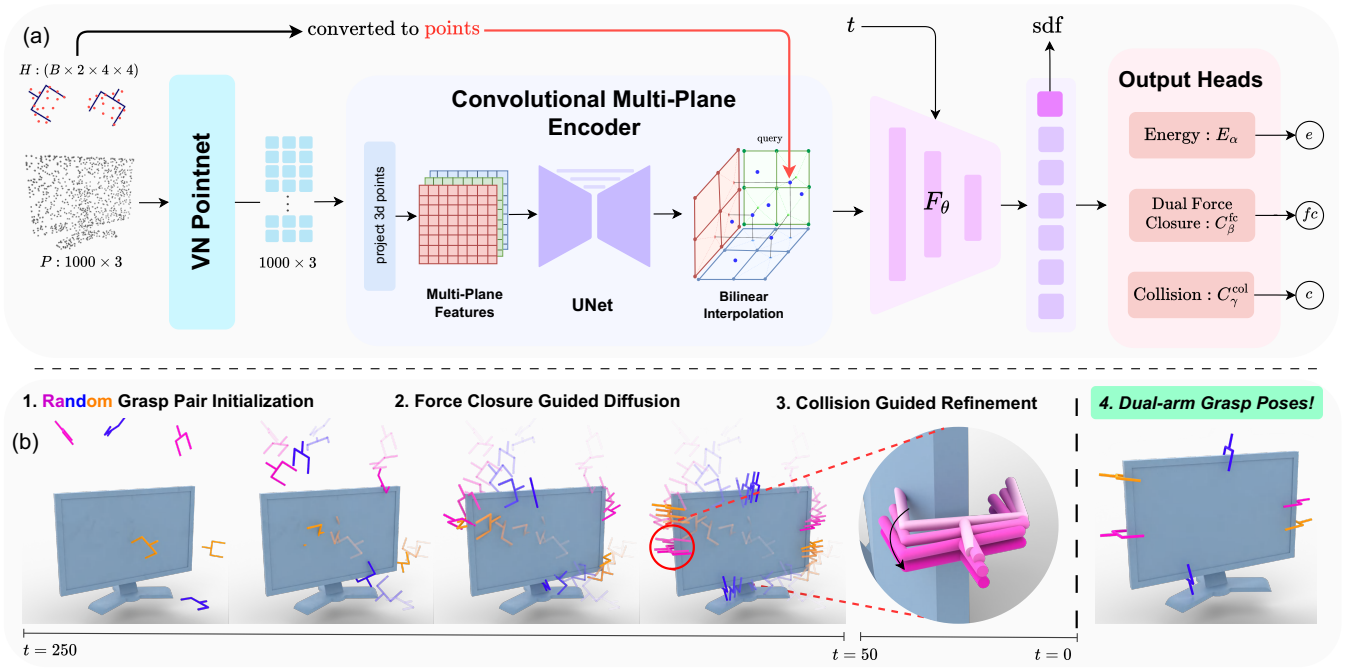

Fig. 2: **Overview of the proposed method: (a)** Given an object point cloud P, our network encodes geometric features into dense feature maps. Next, randomly initialized dual-arm grasps H are used to transform a fixed query cloud into query points, followed by feature sampling through bilinear interpolation. Conditioned on the noise step t, these features are passed through F_θ, which predicts both the SDF of the query points and a feature vector. This vector is used by three output heads that predict energy (E_α), force-closure probability (C_β^{fc}), and collision probability (C_γ^{col}), jointly guiding the diffusion process. **(b)** At inference, denoising proceeds from random initializations ($t = 250$) to refined grasps ($t = 0$). The energy head drives the generative dynamics, while the force-closure and collision heads bias the generation until stable, collision-free dual-arm grasps emerge.

jointly by (i) whether each grasp maintains a stable, non-colliding contact with the object surface, and (ii) whether the pair jointly satisfies the dual-arm force-closure constraints [17]. To address this, we formulate dual-arm grasp generation as a diffusion process in $SE(3) \times SE(3)$. Starting from random pairs of initial poses, we iteratively refine grasp candidates toward physically valid configurations using the score function learned by an energy-based model. In addition to it, we jointly train two classifier-guidance modules — a force-closure classifier and a collision classifier, that provide gradient signals during inference. Together, these components enable the model to generate low-energy, dual-arm stable, and collision-free grasp pairs as shown in Figure 2.

To formalize diffusion in the $SE(3) \times SE(3)$ space, we introduce the following notation. The dual-arm logarithmic map $\text{Logmap}_2 \colon SE(3) \times SE(3) \to \mathbb{R}^{12}$ is defined as:

$$\mathbf{v} = \text{Logmap}_2\,(H) := \text{Logmap}\,(H_1) \oplus \text{Logmap}\,(H_2)$$

where $\oplus$ is the vertical concatenation operator. Similarly, we define the dual-arm exponential map $\text{Expmap}_2 \colon \mathbb{R}^{12} \to SE(3) \times SE(3)$ as:

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

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

In this section, we first discuss the formulation of a diffusion-based dual-arm grasp generation model **(A)**, then formulate the classifier guidance modules for force-closure and collision **(B)** and finally outline the architecture of the complete generation model **(C)**.

A. Diffusion-based Dual-arm Grasp Generation

We adapt the diffusion formulation of SE3Diff [13] to the dual-arm setting. A dual-arm grasp pose H lies in $SE(3) \times SE(3)$, where each element of the pair lies in a Lie group that does not allow direct Euclidean operations. To address this, we map H to a vector $\mathbf{v} \in \mathbb{R}^{12}$ using Logmap_2, which allows the diffusion process to operate in $\mathbb{R}^{12}$, while Expmap_2 is used to convert perturbed samples back to $SE(3) \times SE(3)$. The denoiser is then defined as a vector field s_α that predicts a vector $d \in \mathbb{R}^{12}$ given a dual-arm grasp pose H, the object point cloud P, and the noise step $t \in [0, T)$, where T is the total number of noise steps.

Energy Based Formulation: In standard diffusion models, one option is to directly predict the added Gaussian noise at each noise step. However, learning the score function [28] has shown to yield more stable training and better likelihood modeling, since the score directly captures the structure of the noisy distribution. Formally, the score is defined as: $s(H, P, t) = \nabla_H \log p_t(H|P)$, where p_t is the distribution of the noisy grasps at noise step t.

Hence, we create an energy-based model E_α, with learnable parameters α, to learn a scalar energy landscape over *dual-arm grasp* poses and an input object *point cloud* along with the current *noise step*. Formally, $E_\alpha \colon SE(3) \times SE(3) \times \mathbb{R}^{n\times 3} \times \mathbb{R} \to \mathbb{R}$. The denoising vector field is then obtained as the negative gradient of the

energy, given by,

$$s_\alpha(H, P, t) = -\nabla_H E_\alpha(H, P, t) \in \mathbb{R}^{12} \tag{1}$$

This formulation is equivalent to score-based diffusion, with the advantage that the energy function also provides a natural way to rank grasp candidates. Lower energy indicates grasp pairs that are physically grounded on the object rather than arbitrarily floating in free space.

Forward and Reverse diffusion: During training, the forward diffusion process perturbs the ground truth dual-arm grasp poses by adding noise in $\mathbb{R}^{12}$. The perturbed vector is converted back to $SE(3) \times SE(3)$ using the Expmap_2 operation as:

$$\tilde{H}_t = \text{Expmap}_2(\text{Logmap}_2(H) + \epsilon_t), \tag{2}$$

where $\epsilon_t \sim \mathcal{N}\big(0, \sigma_t^2 I_{12}\big)$ and σ_t is the standard deviation at noise step t. This forward process progressively produces noisier versions of the grasp poses, which are used to train the model to denoise. At inference, the denoising process iteratively removes noise using a Langevin-style update. Given a noisy sample, H_t, the reverse step is defined as:

$$H_{t-1} = \text{Expmap}_2\left(\frac{\eta_t^2}{2}\, s_\alpha(H_t, P, t) + \eta_t \epsilon\right) H_t, \tag{3}$$

where $\epsilon \sim \mathcal{N}(0, I_{12})$ and $\eta_t \geq 0$ is a step-dependent coefficient controlling the update magnitude.

Loss function: The diffusion network is trained using the regular denoising loss function as the L1 norm between the normalized sampled noise $\frac{\epsilon_t}{\sigma_t}$ and predicted vector field $s_\alpha(H_t, P, t)$. Formally,

$$\mathcal{L}_{\text{diff}} = \left\| s_\alpha(H_t, P, t) - \frac{\epsilon_t}{\sigma_t} \right\|_1 \tag{4}$$

B. Classifier Guidance Modules

Naively training a diffusion model in the dual-arm setting leads to poor generalization, since there is no explicit constraint enforcing the generation of stable dual-arm grasps. To address this, we adopt classifier-guided diffusion [29], which steers the generative process toward desired regions of the sample space by incorporating classifier log-likelihood gradients at each reverse step, thereby biasing the generation toward samples with specific properties.

We employ two classifier-guidance modules. The first is a **force-closure classifier** $C_\beta^{\text{FC}} = p(y = 1|H, P; \beta)$ with learnable parameters β, which predicts the probability that a dual-arm grasp pair H on object point cloud P satisfies the force-closure stability criterion. During inference, its gradient with respect to the grasp pose guides the diffusion process toward force-closure stable configurations. The second is a **collision classifier** $C_\gamma^{\text{Col}} = p(y = 1|H, P; \gamma)$ with learnable parameters γ, which predicts the probability that a grasp pose H is in collision with the object point cloud P. During inference, its gradient is used to refine generated candidates by pushing them away from collisions with the object.

Both classifiers are trained with the standard binary cross-entropy loss:

$$\mathcal{L}_{\text{fc}} = \text{BCE}\big(C_\beta^{\text{FC}}(H, P),\, y_{\text{fc}}\big) \tag{5}$$

$$\mathcal{L}_{\text{col}} = \text{BCE}\big(C_\gamma^{\text{Col}}(H, P),\, y_{\text{col}}\big) \tag{6}$$

where $y_{\text{fc}} \in \{0, 1\}$ indicates whether the grasp pair satisfies force closure, and $y_{\text{col}} \in \{0, 1\} \times \{0, 1\}$ denotes which grasp(s) in the pair collide with the object. The overall training objective then combines the diffusion loss (Equation 4) with the classifier losses (Equation 5, 6) as:

$$\mathcal{L} = \mathcal{L}_{\text{diff}} + \mathcal{L}_{\text{fc}} + \mathcal{L}_{\text{col}}. \tag{7}$$

At inference time, the gradients from these classifiers are combined with the base diffusion score, steering the sampling process toward low-energy regions that also satisfy stability and collision constraints ensuring that the final grasp pairs are dual-arm stable as well as physically valid. The final score is then defined as,

$$\tilde{s}(H, P, t) = s_\alpha(H, P, t) + \nabla_H \log C_\beta^{\text{FC}}(H, P) + \begin{cases} 0, & \text{if } t < t_c, \\ \nabla_H \log\big(1 - C_\gamma^{\text{Col}}(H, P)\big), & \text{if } t \geq t_c, \end{cases} \tag{8}$$

where t_c denotes a predefined threshold, after which collision guidance is activated to progressively refine the generated grasps, as refinement is unnecessary while the grasps remain in free space. The corresponding reverse update then follows the same formulation defined in Equation 3.

C. Model Architecture

Our framework (Figure 2) extends diffusion-based grasp generation to the dual-arm setting by introducing a geometry-aware Vision Encoder and a set of specialized Output Heads that jointly enforce feasibility, stability, and collision-free interaction.

Feature Encoding. Inspired from [12], the input point cloud P is encoded using a VN-PointNet [30] to extract $SO(3)$-equivariant per-point features, which are further processed through multi-plane projections and a UNet backbone [31]. Next, dual-arm grasp poses are transformed by a fixed query point cloud $P_q \in \mathbb{R}^{30\times3}$ to get P_H, which represents the local grasp region on the object point cloud. We then retrieve plane features at the projected locations of P_H using bilinear interpolation. These features are then aggregated and passed through the feature encoder F_θ that conditions on the diffusion step t and jointly predicts (i) a feature vector representation and (ii) the SDF of the query points for geometric supervision.

Multi-Head Output. The resulting feature vector is then passed through three heads that play distinct yet tightly coupled roles in the generation step:

1. *Energy Head*: It outputs a scalar energy $E_\alpha(H, P, t) \in \mathbb{R}$, where lower energy corresponds to more physically valid grasps. It serves as the backbone of the diffusion process: during training, its gradients teach the model to denoise noisy samples toward low-energy configurations, and during inference, it provides the base generative dynamics.

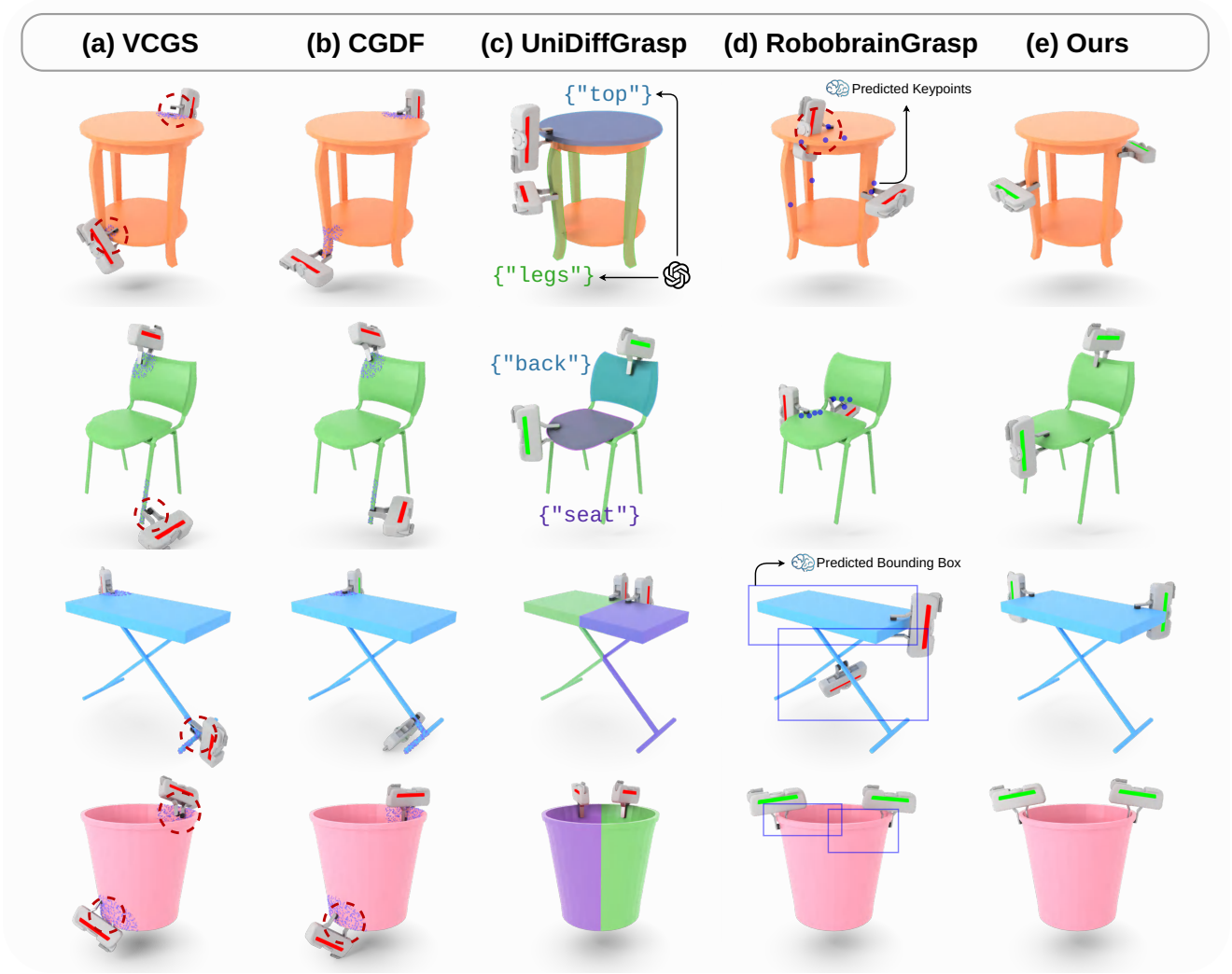

Fig. 3: **Qualitative comparison of dual-arm grasps.** VCGS often fails due to poor grasp generation and its farthest-region heuristic. CGDF shares the latter limitation, for example, on the bucket, one gripper cannot reach the bottom corner. UniDiffGrasp relies on GPT-4V and VLPart for semantic segmentation, but when region labels are ambiguous (e.g., last two rows), it defaults to naive splits, yielding unstable grasps. RoboBrainGrasp predicts keypoints or bounding boxes, yet the keypoints are frequently misaligned (e.g., inside the stool) and the boxes too coarse for precise grasping. In contrast, our method directly generates stable, collision-free dual-arm grasps by reasoning over physical constraints, without heuristics or vague semantic cues. (*Red circles indicate grasps that either collide or fail to contact the object. RoboBrainGrasp-KP and -BB are referred to jointly as RoboBrainGrasp.*)

2. *Force-Closure (FC) Head*: It predicts the probability that a candidate grasp pair achieves force-closure. Importantly, it is trained jointly with the Energy Head, ensuring that the backbone generative process is directly coupled with physically meaningful stability supervision. During inference, its gradients act as guidance signals that bias generation toward dual-arm stable regions of the pose space.

3. *Collision Head*: It predicts the probability of either gripper intersecting the object point cloud. This head is trained after the Vision Encoder, Energy, and FC heads have converged, allowing it to specialize in detecting fine-grained collisions without disturbing the generative objectives. During inference, its gradients are activated after an initial denoising stage, progressively refining candidate grasps by pushing them away from collisions. While our network also predicts SDFs, these are not used for collision refinement since differentiable collision detection is non-trivial and SDF gradients do not guarantee effective grasp refinement.

Together, the three heads form a cooperative architecture: the Energy Head drives diverse and physically valid generation, the FC Head enforces joint grasp stability, and the Collision Head ensures geometric validity. Through this, our framework overcomes the shortcomings of prior methods [12], [18] that treat dual-arm grasping as a combination of independent single-arm grasp proposals.

IV. Experiments

In this section, we describe the dataset, evaluation metrics and the performance of our method compared with different baselines. First, we uniformly sample 1000 points from each object mesh to construct the input point cloud used during inference. For each run, a batch of B dual-arm grasps $\{H_i \in SE(3) \times SE(3)\}_{i=1}^{B}$ are randomly initialized , and

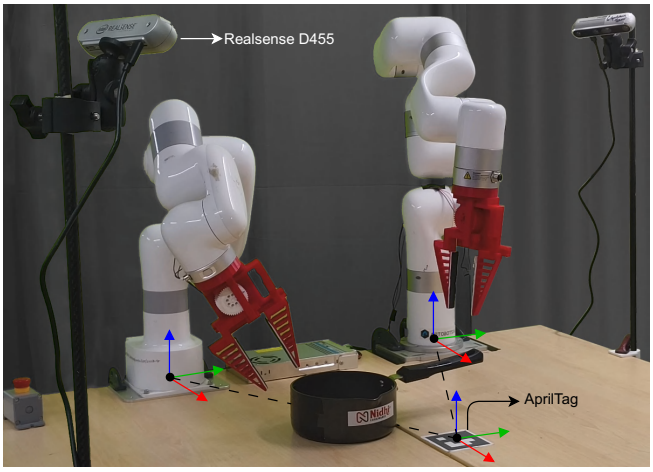

Fig. 4: **Real-world experimental setup.** Dual-arm system with an *XArm7* and *XArm6 Lite*, calibrated using an AprilTag and observed by two RealSense D455 cameras.

the diffusion process is applied for $T = 250$ denoising steps using the formulation defined in Section III. We use the last 50 steps for collision refinement, since by this stage most grasps have converged to feasible regions, with only a small subset intersecting slightly with the object surface. The importance of this refinement stage is further supported by the ablation study in Section V. After denoising and refinement, the final set of generated dual-arm grasps are evaluated using the metrics discussed later in this section.

Dataset: We train and evaluate our model and baselines on the DG16M dataset [17]. The dataset contains 4,143 objects, each consisting approximately 2,000 positive and negative dual-arm grasps validated under improved force-closure evaluation. For our experiments, we adopt a random split in which 400 objects are reserved for testing and the remaining objects are used for training. All reported quantitative results are evaluated on this unseen test split. In addition, we construct a synthetic dataset of colliding and non-colliding dual-arm grasps by sampling grasp poses and checking for collisions with the object. This is used exclusively to train the Collision Head.

Baselines. To assess the performance of our framework, we compare against three categories of baselines for dual-arm grasping: (i) *Farthest-region grasping*, (ii) *VLM-region based grasping*, and (iii) *Affordance-based grasping*.

Farthest-region grasping. **CGDF** [12] applies a part-guided strategy to generate grasps in the two farthest regions of the point cloud and combines them into dual-arm pairs. Additionally, **VCGS** [32] uses a variational grasp generation model to generate single-arm grasps in constrained regions and we use it in the same setup as used in [12] for evaluation.

VLM-based region grasping. **UniDiffGrasp** [18] uses GPT-4V [33] and VL-Part [34] to predict two graspable regions on the object and then applies the CGDF part-guided strategy to form dual-arm pairs. To also study the effect of different VLMs, we adapt this baseline by predicting the regions through bounding-box and keypoint prediction using RoboBrain 2.0 7B [35], a model trained on embodied reasoning tasks. We call these adapted baselines as **RoboBrainGrasp-BB** and **RoboBrainGrasp-KP** respectively.

Affordance-based grasping. **DualAfford** [24] predicts functional object regions for dual-arm interaction and generates the grasp pairs by sequentially conditioning the second gripper's pose on the first. We did not retrain this baseline on our evaluation dataset due to its complex and object category-specific training pipeline. Instead, we evaluated our framework in a zero-shot manner on the test category objects used in DualAfford's pickup task.

Metrics. We evaluate the performance of the proposed method and baselines using *Force Closure Evaluation (FCE)*, *Grasp Success Rate (GSR)* and *Grasp Collision Rate (GCR)*. These metrics together capture analytical grasp stability, physical robustness, and geometric feasibility.

Force Closure Evaluation (FCE): Following the same formulation in DG16M [17], this analytical metric verifies whether a grasp achieves force closure by testing if the applied contact forces can resist arbitrary external wrenches under friction and gripper force constraints. FCE provides a theoretical guarantee of stability and is the primary analytic measure of our grasp quality.

Grasp Success Rate (GSR): This simulation-based metric evaluates the physical robustness of generated grasps in Isaac Gym [21]. Each grasp is executed by initializing floating grippers at the predicted grasp pose, closing the fingers, and then enabling gravity. A trial is marked as successful if the object is lifted to a defined height and remains stably grasped. GSR captures whether grasps that satisfy analytic criteria also translate to positive executions under dynamics.

Grasp Collision Rate (GCR): This metric measures the percentage of generated grasps whose final pose intersects with the object geometry. A lower collision rate indicates better geometric validity and effectiveness of the collision-guidance mechanism.

V. Results and Ablation

Qualitative Results. Figure 3 illustrates qualitative comparisons across methods. For CGDF and VCGS (Figure 3 (a), (b)), the farthest-region heuristic often selects physically incompatible grasp regions, such as on the bucket where one gripper is forced to an unstable corner. Even when the chosen regions happen to be graspable, like in the case of the chair and stool, the resulting configurations require excessive or unbalanced contact forces. For the VLM-based baselines, as seen in the stool and ironing-table, RoboBrain-based methods (Figure 3(d)) predict coarse bounding boxes and keypoints that either lie in free space or fail to align with usable regions. Similarly, UniDiffGrasp (Figure 3(c)) often defaults to arbitrary splits (e.g., the bucket), which result in unstable grasps. In contrast, DAGDiff (Figure 3(e)) consistently produces dual-arm grasps that are both stable and collision-free, without relying on heuristics or semantic labels and performs twice as good compared to previous methods in all three metrics. Our classifier-guided diffusion

Method	FCE(%)↑	GSR(%)↑	GCR(%)↓
DAGDiff (ours)	**60.14**	**72.50**	**15.10**
CGDF [12]	35.14	56.25	30.55
VCGS [32]	16.85	23.36	74.73
UniDiffGrasp [18]	10.10	31.68	59.90
RoboBrainGrasp-KP [35]	9.80	27.85	66.30
RoboBrainGrasp-BB [35]	7.12	27.81	70.26
Ours-DA†	**56.45**	**68.80**	**18.59**
Dual-Afford†† [24]	–	54.33	–

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

††Values reported directly from the DualAfford paper.

TABLE I: Comparison of dual-arm grasp generation methods. Higher is better for FCE and GSR; lower is better for GCR.

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

TABLE II: Ablation study with three variants: without force-closure head, without collision head, and with post-hoc FC Head. Metrics reported are FCE, GSR, and GCR.

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

TABLE III: Real-world dual-arm grasp execution results. Each entry shows the number of successful grasps over total attempted grasps for the corresponding object.

discovers physically consistent grasp regions directly, enabling reliable dual-arm grasp generation across diverse and complex object geometries.

Quantitative Results. Table I summarizes the performance of DAGDiff compared to other baselines. Farthest-region heuristics (CGDF, VCGS) perform poorly in FCE and GSR because they decouple grasp selection from stability. VCGS further suffers from its reliance on a global shape representation, which explains its very high GCR. VLM-based methods (UniDiffGrasp, RoboBrainGrasp) also show poor performance in all metrics, as their semantically predicted regions are frequently coarse or misplaced due to limited 3D and physical reasoning which leads to grasps generated in the unsuitable object regions. Finally, compared to DualAfford [24], which reports results from category-specific models, our single unified model achieves higher GSR in a zero-shot setting (under the same evaluation protocol), indicating stronger generalization to unseen categories. Overall, by coupling diffusion with explicit stability and collision guidance, DAGDiff overcomes the core limitations of heuristic, semantic, and affordance-based baselines, producing grasps that are jointly stable and generalizable. As shown in Table I, it achieves 60.1% FCE, 72.5% GSR, and 15.1% GCR, roughly twice the stability and success of the baselines, while reducing collisions by more than half.

Ablations. The key design choice in our framework is the use of classifier gradients to guide grasp generation, ensuring that the final dual-arm grasps are physically valid and stable. To assess the necessity of this design, we conduct ablation studies to study its contribution, as shown in Table II.

(A) Generation without the Force-Closure Head: In this variant, we remove the Force-Closure head and rely solely on the Energy head's denoising objective for grasp generation. This leads to a significant drop in both FCE and GSR, indicating that the Energy head alone cannot reliably capture dual-arm stability. Because the FC head is explicitly trained to discriminate between stable and unstable dual-arm grasp pairs, its gradients provide precise and meaningful guidance during generation. This highlights the necessity of explicit force-closure guidance for producing stable dual-arm grasps.

(B) Generation without Collision Head: Next, we remove the Collision Head and evaluate the grasps without collision refinement, and notice that the collision rate increases from 15.10% to 23.50%. This causes the grasps to make incorrect contact with the object surface and this affects FCE and GSR negatively too. The Collision Head provides explicit signals to avoid collisions, which the other heads cannot enforce.

(C) Post-hoc FC Head training: To test if the Force-Closure (FC) head could be added post-hoc, we first trained the vision encoder and Energy head, then froze them and trained the FC head separately. This makes FCE and GSR drop substantially, as the encoder could not adapt to the classification objective, leaving the FC head with weak features and poor guidance. This shows that the FC head is not a plug-and-play module but must be jointly trained with the encoder to learn stability-aware representations that effectively guide grasp generation.

VI. Real Life Experiments

We validated our framework on a heterogeneous dual-arm setup (Figure 4) with an *XArm7* and an *XArm6 Lite*, using two Intel RealSense D455 cameras for point cloud fusion via ICP. The fused point cloud P was passed to our trained model to generate dual-arm grasps, pruned to retain only kinematically reachable ones. In 10 real-world trials (Table III), most executions were successful, while failures were mainly from the loss of detail in point cloud reconstruction, which led to incorrect generation of grasps. These results show that our method achieves zero-shot transfer to real sensor data, handling previously unseen objects such as drones and kitchen utensils like saucepans and trays.

VII. Conclusion

In this work, we present DAGDiff, a novel diffusion-based framework for generating stable and collision-free dual-arm grasps directly in the $SE(3) \times SE(3)$ space. By guiding the generative process with force-closure and collision-aware signals, our method outperforms heuristic or region detection-based pipelines and demonstrates reliable zero-shot transfer to previously unseen objects in real-world trials.

While effective, the framework currently assumes complete, segmented point clouds and does not account for closed chain kinematics. Moreover, its inference speed is limited by the iterative nature of diffusion. As future work, we aim to address these limitations by extending the approach to partial observations, and conducting comprehensive real-world evaluations, paving the way toward more scalable and practical dual-arm manipulation.

REFERENCES

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

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

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

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

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

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

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

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

[9] H.-S. Fang, C. Wang, H. Fang, M. Gou, J. Liu, H. Yan, W. Liu, Y. Xie, and C. Lu, "Anygrasp: Robust and efficient grasp perception in spatial and temporal domains," *IEEE Transactions on Robotics*, vol. 39, no. 5, pp. 3929–3945, 2023.

[10] A. D. Vuong, M. N. Vu, H. Le, B. Huang, H. T. T. Binh, T. Vo, A. Kugi, and A. Nguyen, "Grasp-anything: Large-scale grasp dataset from foundation models," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 14 030–14 037.

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

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

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

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

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

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

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

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

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

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

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

[22] R. M. Murray, S. S. Sastry, and L. Zexiang, *A Mathematical Introduction to Robotic Manipulation*, 1st ed. USA: CRC Press, Inc., 1994.

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

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

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

[26] J. Carvalho, A. T. Le, P. Jahr, Q. Sun, J. Urain, D. Koert, and J. Peters, "Grasp diffusion network: Learning grasp generators from partial point clouds with diffusion models in so (3) xr3," *arXiv preprint arXiv:2412.08398*, 2024.

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

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

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

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

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

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

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

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

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