BiGraspFormer: End-to-End Bimanual Grasp Transformer

Kangmin Kim[1], Seunghyeok Back[2], Geonhyup Lee[1], Sangbeom Lee[1], Sangjun Noh[1], Kyoobin Lee[1†]

***Abstract*— Bimanual grasping is essential for robots to handle large and complex objects. However, existing methods either focus solely on single-arm grasping or employ separate grasp generation and bimanual evaluation stages, leading to coordination problems including collision risks and unbalanced force distribution. To address these limitations, we propose BiGraspFormer, a unified end-to-end transformer framework that directly generates coordinated bimanual grasps from object point clouds. Our key idea is the Single-Guided Bimanual (SGB) strategy, which first generates diverse single grasp candidates using a transformer decoder, then leverages their learned features through specialized attention mechanisms to jointly predict bimanual poses and quality scores. This conditioning strategy reduces the complexity of the 12-DoF search space while ensuring coordinated bimanual manipulation. Comprehensive simulation experiments and real-world validation demonstrate that BiGraspFormer consistently outperforms existing methods while maintaining efficient inference speed (<0.05s), confirming the effectiveness of our framework. Code and supplementary materials are available at https://sites.google.com/bigraspformer**

I. INTRODUCTION

Bimanual grasping enables robots to manipulate large, heavy, or unwieldy objects beyond single-arm capabilities, making it essential for tasks such as lifting furniture, carrying long boards, or moving large boxes [1], [2]. However, most robotic grasping research has focused on single-arm systems, primarily on learning to detect 6-DoF grasp poses from point clouds [3]–[8]. While effective for single-arm tasks, these approaches cannot be directly extended to bimanual scenarios. First, bimanual grasping expands the action space to 12-DoF, doubling the computational complexity. Second, it introduces new challenges, including collision avoidance, balanced force/torque distribution, and dual-arm coordination for post-grasp manipulation.

For bimanual grasping, only a few methods have been proposed so far. The DA2 dataset [9] introduced the first benchmark by extending single-arm datasets [3], [10], [11] with dual-arm-specific metrics such as force closure, dexterity, and torque balance [9], [12]. However, most existing approaches adopt modular architectures that separate grasp generation and evaluation. For example, Dual-PointNetGPD [9] evaluates the quality of grasp pairs from given candidates, requiring external single-arm grasp generators. Similarly, CGDF [13] directly generates bimanual grasps but lacks integrated quality prediction, instead relying on additional scoring modules or heuristic pairing strategies [14], [15].

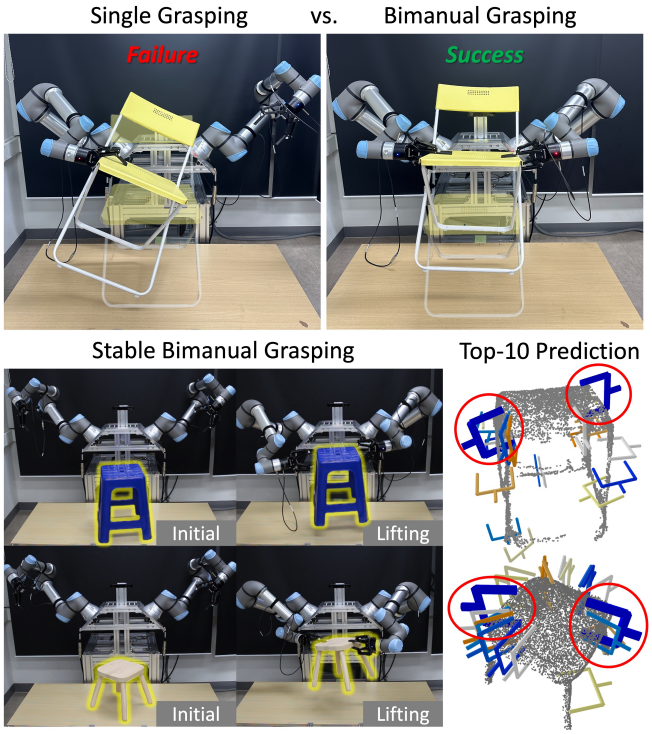

Fig. 1. **BiGraspFormer for coordinated bimanual grasping.** (Top) Comparison between single and bimanual grasping for large objects. (Bottom) BiGraspFormer successfully grasps and lifts diverse objects in real-world environments, demonstrating stable grasps across various geometries. We visualize the point cloud with the top-10 grasp predictions; the highest-scoring pair is highlighted in thick blue.

As a result, current methods yield limited diversity, poor coordination, and high computation due to modular pipelines.

In this paper, we propose **BiGraspFormer**, the first unified end-to-end framework that directly generates coordinated bimanual grasps from object point clouds (Fig. 1). The key insight is that single-grasp features can effectively guide bimanual grasp generation, rather than treating dual-arm coordination as two independent problems. BiGraspFormer introduces a novel Single-Guided Bimanual (SGB) strategy: it first generates diverse single-arm grasp candidates, then leverages their learned features through specialized attention mechanisms to jointly predict bimanual poses and quality scores. This unified approach eliminates separate modules and explicitly models coordination between grasps, enabling stable and efficient dual-arm manipulation. Comprehensive experiments in both simulation and real-world environments demonstrate that BiGraspFormer achieves superior success, diversity, and speed compared to existing methods.

[1] Department of AI Convergence, Gwangju Institute of Science and Technology (GIST), Gwangju 61005, Republic of Korea.
[2] Department of AI Machinery, Korea Institute of Machinery & Materials (KIMM), Daejeon 34103, Republic of Korea.
[†] Corresponding author: Kyoobin Lee (kyoobinlee@gist.ac.kr).

Our contributions are summarized as follows:

- We propose **BiGraspFormer**, the first unified end-to-end transformer for diverse, stable bimanual grasp generation.
- We introduce the **Single-Guided Bimanual (SGB)** strategy, which leverages single-arm grasp features to guide bimanual generation, reducing computational complexity and enhancing dual-arm coordination.
- We achieve **state-of-the-art bimanual grasping performance** while maintaining fast inference suitable for real-world deployment.

II. Related Works

A. Learning-based Single-arm Grasping

Most existing methods have focused on single-arm grasping, evolving from discriminative approaches [4], [6] that score grasp candidates to generative methods [5], [7], [8], [16], [17] that directly synthesize grasps from visual inputs. While these single-arm approaches achieve strong performance on diverse objects, they face fundamental limitations for bimanual scenarios. The action space expands from 6-DoF to 12-DoF, and coordinated dual-arm manipulation introduces critical requirements including collision avoidance, force balance, and post-grasp manipulability for task execution. Simply combining two independent single-arm solutions cannot address these coordination challenges. Our SGB strategy bridges this gap by leveraging single grasp features as guidance for coordinated bimanual generation, rather than simply pairing independent single-arm solutions.

B. Model-based Bimanual Grasping

Model-based bimanual grasping methods [18]–[21] have traditionally required explicit object knowledge such as 3D CAD models or predefined shape categories. These methods employ techniques including genetic algorithms for rod-shaped objects [18], medial axis transforms for shape analysis [19], grasp matrix computations for stability evaluation [20], [21]. While these approaches successfully generated stable grasps for known objects, they cannot handle unknown objects in unstructured environments due to their dependence on explicit geometric models. This fundamental limitation has motivated the development of learning-based approaches that can generalize to diverse unseen objects.

C. Learning-based Bimanual Grasping

Learning-based approaches have addressed model-based limitations through specialized datasets and neural methods. The DA2 dataset [9] introduced dual-arm-specific metrics including force closure, dexterity, and torque balance, along with Dual-PointNetGPD for grasp quality evaluation. However, this approach only evaluates pre-generated candidates, requiring separate grasp generators for practical use. Generative methods have since emerged to address this limitation. Dual-Afford [22] generates task-specific grasps for trained affordances but lacks generalization and diversity. CGDF [13] employs diffusion models to generate single grasps but requires additional pairing modules to form bimanual candidates, causing computational overhead and grasp pairs with potential arm collisions and unbalanced force distribution due to independent single-arm generation. BiGraspFormer addresses these limitations through unified end-to-end generation and evaluation, eliminating the need for separate modules while ensuring coordinated bimanual planning.

III. Method

A. Motivation

Our goal is to predict bimanual grasps $\mathcal{B} = \{\mathbf{g}^1, \mathbf{g}^2, q\}$ from an object point cloud P, where $\mathbf{g}^1$ and $\mathbf{g}^2$ represent the 6-DoF grasp poses of the two arms and q denotes grasp quality. This task is challenging as it involves a 12-DoF action space, doubling the complexity of single-arm grasping. Each grasp needs to achieve force-closure stability while both arms coordinate to avoid collisions, maintain torque balance, and ensure overall stability.

To tackle this challenge, we introduce the **Single-Guided Bimanual (SGB)** grasp generation scheme, which decomposes the prediction of $\mathcal{B}$ into three structured stages. First, generate diverse single grasp candidates under basic stability constraints. Second, select feasible grasp pairs by discarding collisions and ensuring balanced force distribution. Finally, refine these pairs into stable bimanual grasps using learned features from both the object and single grasps. This formulation explicitly enforces both individual grasp quality and dual-arm coordination, decomposing the complex 12-DoF search space into a sequence of more tractable subproblems.

B. BiGraspFormer

BiGraspFormer employs the proposed SGB strategy as an end-to-end framework for bimanual grasp generation. As shown in Fig. 2, the network consists of four modules: 1) object encoder that extracts both local and global geometric features from the input point cloud, 2) Single Grasp Proposer (SGP) that generates diverse 6-DoF grasp candidates, 3) Bimanual Pair Matcher (BPM) that selects feasible grasp pairs based on collision checks and bimanual quality scores, and 4) Bimanual Grasp Generator (BGG) that refines these pairs into final 12-DoF bimanual grasps with scores.

Object Encoder. Bimanual grasping requires not only capturing fine-grained local geometry (e.g., contact points for force closure) but also modeling global object structure for coordinated bimanual grasps. We designed our object encoder to capture both local detail and global context within an integrated representation. We use PointNet++ [23] to extract local geometry. It uses Set Abstraction (SA) layers to get multi-scale features from the input point cloud $P \in \mathbb{R}^{N\times 3}$. This keeps the detailed shape information important for checking if single grasps are stable. To complement these local features with global context, a transformer encoder is employed, consisting of self-attention layers. In this encoder, queries q, keys k, and values v are derived from the local features from PointNet++. This enables the network to capture relationships between distant regions and spatial relationships across the object. Specifically, the

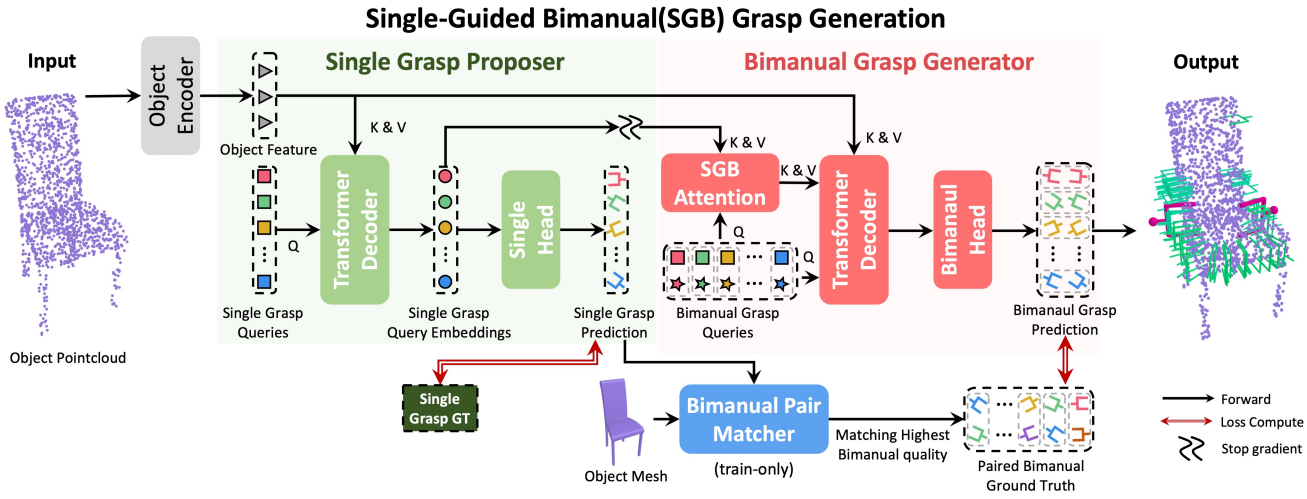

Fig. 2. **Overview of our BiGraspFormer framework.** An object encoder processes point cloud P to extract geometric features. Single Grasp Proposer generates force-closure single grasps, Bimanual Pair Matcher matches them using bimanual quality metrics (force stability, torque balance, dexterity) to create ground truth, and Bimanual Grasp Generator employs SGB attention for final bimanual grasp generation.

encoder consists of two SA layers from PointNet++ followed by six transformer encoder blocks. The final output is a set of object global features $\mathcal{F}'_g \in \mathbb{R}^{N' \times C}$, which serve as the backbone representation for the subsequent stages of SGB grasp generation.

Single Grasp Proposer. Following the SGB scheme, the SGP generates diverse 6-DoF single-grasp candidates that satisfy force-closure constraints only. These candidates are subsequently paired and refined by the BPM and BGG modules into final bimanual grasps. To this end, a DETR-style set prediction formulation is employed, where grasp poses are predicted from learnable queries. Specifically, K' grasp queries $\{q_i^s\}_{i=1}^{K'}$ are fed into a transformer decoder that attends to the encoded object features $\mathcal{F}'_g$. To enable query-based feature extraction, we employ cross-attention, where the grasp queries serve as q, while the encoded features $\mathcal{F}'_g$ serve as both keys k and values v. This enables each learnable query to specialize in different object regions and capture distinct grasp poses. The decoder consists of six transformer decoder blocks, followed by a prediction head composed of three MLP layers. This module outputs a set of single grasp candidates $\hat{\mathcal{G}} = \{\hat{\mathbf{g}}_i\}_{i=1}^{K'}$, which provide the input for the subsequent pairing stage.

Bimanual Pair Matcher. BPM bridges single grasp proposals and bimanual grasp generation by constructing reliable supervision pairs from the set $\hat{\mathcal{G}}$. Although the DA2 dataset [9] contains substantial grasp annotations (2,001 per object), the large object sizes make grasp coverage relatively sparse compared to single-arm datasets, making direct training insufficient for learning diverse bimanual coordination. BPM addresses this limitation by matching high-quality bimanual ground-truth pairs from the diverse single grasps proposed by the SGP. BPM operates in three steps: First, it evaluates all possible pairs from single grasps generated by SGP using the established bimanual quality metric from the DA2 dataset [9], which combines force closure, dexterity, and torque balance. Second, collision checking eliminates infeasible pairs that would cause gripper-gripper or gripper-object collision using the object CAD model. Finally, for each anchor grasp, BPM selects the highest-scoring collision-free counterpart, forming reliable bimanual supervision pairs $\mathcal{B}^+ = (\mathbf{g}_i^{1+}, \mathbf{g}_i^{2+}, q_i^+)_{i=1}^{K'}$ as *ground truth* for BGG. Note that BPM is used only during *training* since it requires the object mesh. The process is efficiently parallelized on GPUs, introducing negligible computational overhead to the training pipeline. Since BPM requires object CAD models for collision detection and ground truth quality evaluation, it cannot be deployed at inference time where only point cloud observations are available. Instead, during inference, the BGG module directly predicts both bimanual grasp poses and their quality scores in an end-to-end manner.

Bimanual Grasp Generator. As the final stage of the SGB framework, BGG generates 12-DoF bimanual grasps conditioned on single grasp features. It jointly refines them into complete bimanual poses with a quality that accounts for force closure, dexterity, and torque balance. BGG employs a DETR-style set prediction with learnable bimanual grasp queries $\{\mathbf{q}_i^b\}_{i=1}^{M'}$ processed through a transformer decoder equipped with the proposed **SGB attention layer**. In this layer, bimanual grasp queries serve as queries while single grasp features $\mathbf{F}_{sgp}$ from SGP serve as both keys and values, enabling bimanual generation to be conditioned on single grasp information:

$$\begin{aligned} &\text{SGB-Attention}(\mathbf{Q}^b, \mathbf{F}_{sgp}) \\ &\quad = \text{softmax}\left(\frac{\mathbf{Q}^b(\mathbf{F}_{sgp})^T}{\sqrt{d}}\right)\mathbf{F}_{sgp} \end{aligned} \tag{1}$$

where $\mathbf{Q}^b$ represents the bimanual grasp queries $\{\mathbf{q}_i^b\}_{i=1}^{M'}$ and d is the embedding dimension. This allows each bimanual query to selectively attend to relevant single grasp informa-

tion for coordinated bimanual grasp generation.

The SGB attention output $\mathcal{F}_{\text{sgb}}$ and global object features $\mathcal{F}'_g$ both serve as key-value pairs for subsequent transformer decoder blocks. In each block, bimanual queries perform cross-attention to these features separately, fusing single-grasp evidence with global context. The decoder consists of six transformer blocks followed by a prediction head consisting of seven MLP layers, outputting bimanual candidates $\hat{\mathcal{B}} = \{(\hat{\mathbf{g}}_i^1, \hat{\mathbf{g}}_i^2, \hat{q}_i)\}_{i=1}^{M'}$ where $\hat{\mathbf{g}}_i^1$ and $\hat{\mathbf{g}}_i^2$ are predicted 6-DoF grasp poses for both arms and $\hat{q}_i$ is the quality score.

C. Implementation Details

Loss Functions. We utilize $\mathcal{L}_{dist}$ to regress grasp representation point $\mathbf{v}$, which is suggested in [5]. The regression loss for single grasp:

$$\begin{aligned}\mathcal{L}_{single}\left(\mathcal{G}_i, \hat{\mathcal{G}}_j\right) &= \mathcal{L}_{dist}\left(\mathbf{g}_i, \hat{\mathbf{g}}_j\right) \\ &= \left\|\mathbf{v}_i - \hat{\mathbf{v}}_j\right\|_2,\end{aligned} \tag{2}$$

For bimanual grasps, we incorporate an L1 loss as $\mathcal{L}_{quality}$ to regress the bimanual quality. The overall bimanual grasp loss is:

$$\begin{aligned}\mathcal{L}_{bimanual}\left(\mathcal{B}_i, \hat{\mathcal{B}}_j\right) = \mathcal{L}_{dist}\left(\mathbf{g}_i^1, \hat{\mathbf{g}}_j^1\right) + \mathcal{L}_{dist}\left(\mathbf{g}_i^2, \hat{\mathbf{g}}_j^2\right) \\ + \mathcal{L}_{quality}\left(q_i, \hat{q}_j\right)\end{aligned} \tag{3}$$

To enable training with these losses, we employ bipartite matching [8] between predictions and ground truths. We use the Hungarian method [24] to find optimal assignments, where the cost function for single grasps equals $\mathcal{L}_{single}$, while for bimanual grasps we exclude quality terms from $\mathcal{L}_{bimanual}$ during matching to focus on geometric alignment. The overall grasp loss for BiGraspFormer training is formulated as:

$$\mathcal{L}_{grasp} = \mathcal{L}_{single}\left(\mathcal{G}_i, \hat{\mathcal{G}}_{\hat{\rho}_s i}\right) + \mathcal{L}_{bimanual}\left(\mathcal{B}_i, \hat{\mathcal{B}}_{\hat{\rho}_{bi} i}\right) \tag{4}$$

where, $\hat{\rho}_s$ and $\hat{\rho}_{bi}$ are optimal assignments.

Training Details. The input point cloud consists of $N = 2048$ points, with PointNet++ [23] encoding them into $N' = 512$ center points. The embedding dimension of global object features C is set to 512. For training data preparation, we sample $K = 128$ single grasps using the spatial grid-based method from [16] to form uniformly distributed ground truth set $\mathcal{G}_s$. We set the number of single grasp queries K', bimanual grasp queries M', and ground truth bimanual grasps M to 512 each. The model is implemented in PyTorch 2.1 with CUDA 11.8 and trained on four NVIDIA Tesla A100 GPUs (40GB) using AdamW optimizer with learning rate 5×10^{-4} and batch size 24 until validation loss convergence. Inference and real-world experiments use a single NVIDIA RTX 3090 GPU.

IV. Experiments

A. Comparison with State-of-the-Art Methods

Datasets. We train and evaluate our BiGraspFormer using the DA2 dataset [9], which contains 6,327 CAD models from ShapeNetSem [25]. We use the main files subset containing 3,417 objects, each with annotations of 2,001 bimanual grasps and corresponding quality scores. Since no official data splits are provided, we divide the dataset into a training set with 2,823 objects, a validation set with 494 objects, and a test set of 100 objects. The test set consists of objects not included in the training or validation sets, used specifically for simulation-based evaluation. To train the SGP, we convert bimanual grasps into single grasps and remove duplicates. For the BGG, we generate ground truth bimanual grasps using BPM on the single grasps from the SGP, as described in Section III-B.

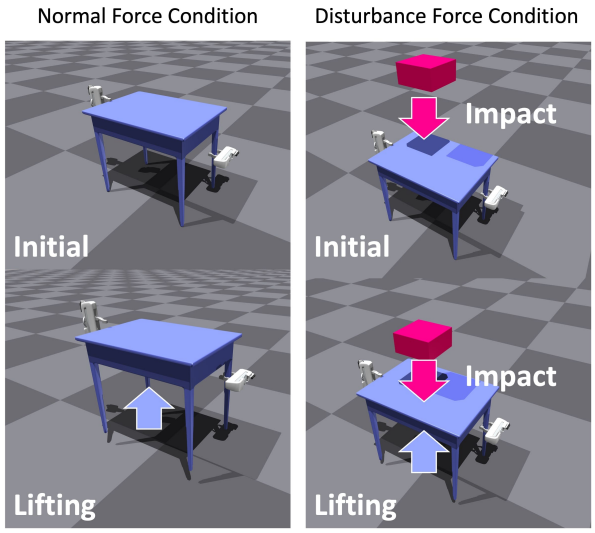

Fig. 3. **Simulation experiment settings.** Left shows normal force condition where objects are grasped and lifted under gravity only. Right shows disturbance force condition where a weighted cube is dropped onto the object during lifting to apply additional external forces and test grasp robustness.

Metrics. For performance evaluation, we use two metrics: simulation-based grasp success rate and diversity. The success rate evaluation follows prior works [5], [7], [8], [13] under both normal force conditions and disturbance conditions where external forces are applied to test grasp robustness. We utilize the Isaac Gym simulator [26] with two free-floating Franka Emika Panda grippers with extended 6-cm fingers. Using grippers without arms ensures fair evaluation independent of robot reachability or approach direction [5], [7], [8], [13], [16]. In the simulation, we place the grippers at the predicted bimanual grasp poses and close both simultaneously without gravity. After enabling gravity, both grippers lift the object upward. A grasp succeeds if all fingers maintain contact and the object remains firmly held. For disturbance conditions, a 1kg weighted cube is dropped from 60cm height onto the center of the object during lifting to introduce additional external forces.

For diversity evaluation, we measure how well successful grasps cover the object surface. We approximate each gripper with a simplified geometric model [4] and position them at successful grasp poses from our predictions. For each object, we calculate diversity as the percentage of object point cloud points that lie within any gripper's geometric bounds,

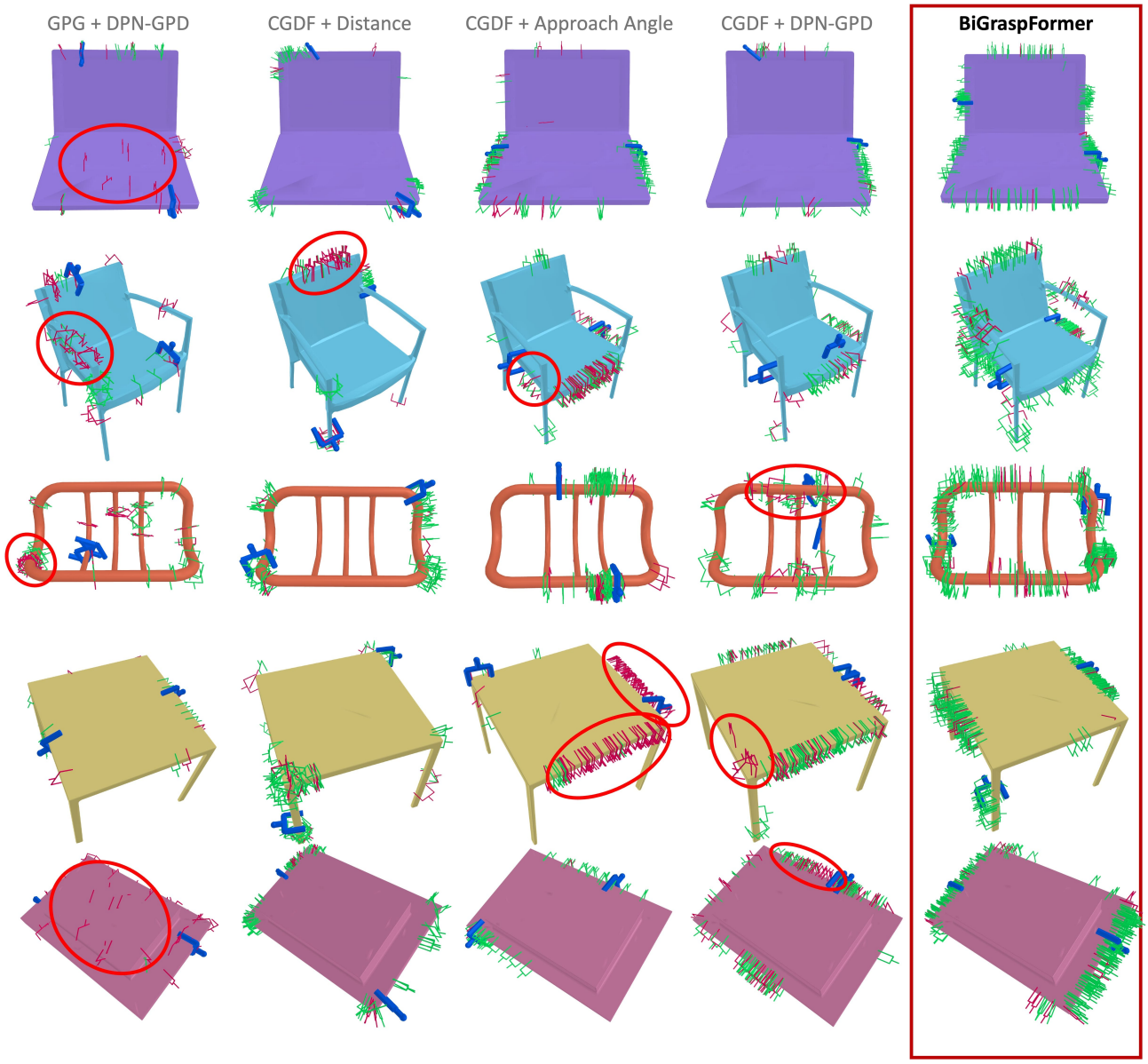

Fig. 4. **Visualization of predicted bimanual grasp poses.** The top 100 bimanual grasp poses predicted by DPN-GPD, CGDF, and BiGraspFormer are shown for each object in the test set, based on simulation outcomes. The top-1 grasp pair is highlighted in blue. Green grasps represent successful grasps, while red grasps indicate failures due to instability, object collisions, or torque imbalance during grasping or lifting. Red circles highlight notable failure cases of baseline methods.

ensuring each point is counted only once regardless of how many grippers cover it. Higher diversity scores indicate that grasps are distributed across a larger portion of the object surface rather than concentrated in specific regions. For these two metrics, we consider the predictions ranked in the top 1%, 30% and 50% for grasp success rate, top 30% and 50% for diversity.

Baselines. We used the following baselines:

- **GPG + DPN-GPD**: GPG [27] for single grasp generation with DPN-GPD [9] for quality-based bimanual pairing.
- **CGDF + Distance**: Unconstrained CGDF [13] for single grasp generation with distance-based pairing that selects two grasps located farthest apart on the object [14], [15].
- **CGDF + Approach Angle**: Unconstrained CGDF for single grasp generation with approach angle-based pairing, selecting grasps with opposing approach directions similar to antipodal selection [28].
- **CGDF + DPN-GPD**: Unconstrained CGDF for single grasp generation with DPN-GPD for quality-based bimanual pairing.

For fair comparison, we use 256 single grasp candidates for all methods to balance computational efficiency with grasp diversity. DPN-GPD is trained to score grasp quality for bimanual scenarios, while unconstrained CGDF generates grasps across the entire object surface without restricting to specific regions. We select the top 512 bimanual pairs based

TABLE I

SIMULATION BASED GRASP SUCCESS AND DIVERSITY PERFORMANCE OF BIGRASPFORMER AND STATE-OF-THE-ART METHOD ON DISTURBANCE FORCE CONDITION

Method	Success Rate			Diversity	
	Top 1%	Top 30%	Top 50%	Top 30%	Top 50%
GPG [27]+DPN-GPD [9]	20.96	19.72	19.97	12.20	14.91
CGDF [13]+Distance [15]	23.98	24.02	23.42	8.28	9.12
CGDF [13]+Approach Angle [28]	34.70	31.34	31.77	15.51	18.17
CGDF [13]+DPN-GPD [9]	36.46	32.14	32.86	14.39	17.95
BiGraspFormer(Ours)	**59.72**	**57.16**	**55.66**	**29.40**	**37.99**

TABLE II

SIMULATION BASED GRASP SUCCESS AND DIVERSITY PERFORMANCE OF BIGRASPFORMER AND STATE-OF-THE-ART METHOD ON NORMAL FORCE CONDITION

Method	Success Rate			Diversity		Time
	Top 1%	Top 30%	Top 50%	Top 30%	Top 50%	(s)
GPG [27]+DPN-GPD [9]	53.05	53.32	53.37	17.23	19.65	10.99
CGDF [13]+Distance [15]	76.09	78.01	77.75	12.66	13.18	4.39
CGDF [13]+Approach Angle [28]	71.38	71.14	71.04	23.29	26.13	4.82
CGDF [13]+DPN-GPD [9]	71.61	71.79	72.67	20.74	24.26	18.18
BiGraspFormer(Ours)	**89.67**	**84.86**	**83.65**	**36.99**	**46.47**	**0.04**

on their predicted quality scores for evaluation.

Result Table I and Table II compare grasp success rates and diversity with state-of-the-art methods in simulation. BiGraspFormer consistently outperforms all baselines under both normal and disturbance conditions. In particular, BiGraspFormer achieves superior top 1% success rates across all conditions. Our method also demonstrates superior diversity compared to baselines, generating diverse yet stable bimanual grasp poses. Under external disturbances (Table I), BiGraspFormer consistently surpasses baselines in both success and diversity, confirming robustness to disturbances.

Table II also reports computational efficiency measured on an RTX 3090 and Intel i7-12700K. For fair comparison, we measure total inference time including model forward pass for all methods, plus pairing stage time for baseline approaches that require separate pairing modules. Times are averaged across multiple runs. BiGraspFormer achieves inference times under 0.05 seconds, demonstrating efficient grasp generation. Fig. 4 shows qualitative comparisons with state-of-the-art methods. BiGraspFormer generates more diverse and stable bimanual grasp candidates across various object geometries.

TABLE III

ABLATION STUDY OF THE MODULES IN THE SGB FRAMEWORK ON NORMAL FORCE CONDITION

BPM	SGB attention	Success Rate			Diversity	
		Top 1%	Top 30%	Top 50%	Top 30%	Top 50%
✗	✗	65.64	60.15	60.37	32.69	38.74
✓	✗	80.61	77.40	76.30	36.92	46.18
✓	✓	**89.67**	**84.86**	**83.65**	**36.99**	**46.47**

B. Ablation Study

Effect of Single-Guided Bimanual Grasp Generation. We conduct ablation studies to verify the effectiveness of the proposed SGB framework by comparing BiGraspFormer with variants lacking the BPM and SGB attention layer. As a baseline, we use a model that directly generates bimanual grasps. This baseline consists of only the object encoder and BGG module without SGB attention layer, trained on the original DA2 dataset [9] instead of BPM-generated pairs. Table III shows that directly generating bimanual grasps fails to achieve state-of-the-art performance. Without the SGB attention layer (using only BPM), the model achieves competitive top-1% success rates but degrades at top-30% and top-50%. In contrast, BiGraspFormer with the complete SGB framework outperforms all variants across all metrics, demonstrating the effectiveness of our SGB grasp generation scheme for producing stable bimanual grasp candidates.

TABLE IV

ANALYSIS OF DIFFERENT NUMBERS OF BIMANUAL GRASP QUERIES ON NORMAL FORCE CONDITION

Number of Queries	Success Rate			Diversity	
	Top 1%	Top 30%	Top 50%	Top 30%	Top 50%
128	76.40	70.28	68.78	18.10	24.06
256	82.50	79.10	78.37	27.39	35.49
512(Ours)	**89.67**	**84.86**	**83.65**	**36.99**	**46.47**

Effect of number of Bimanual Grasp Queries To investigate the impact of query quantity, we compare BiGraspFormer models with varying numbers of learnable bimanual grasp queries (128, 256, and 512). Table IV shows that increasing the number of queries consistently improves both metrics, indicating that more queries better capture the complexity of the bimanual grasp space. With only 128 queries, the model performs worse than state-of-the-art baselines, suggesting that insufficient queries cannot adequately represent the large bimanual grasp space. Notably, BiGraspFormer achieves state-of-the-art performance with just 256 queries. Even with half the candidates, it demonstrates higher diversity than other methods generating 512 bimanual pairs. This validates our SGB framework's effectiveness in decomposing the complex 12-DoF search space into simpler subproblems, enabling the model to learn more efficiently and generate more diverse grasps with fewer queries.

TABLE V

GRASP PERFORMANCE IN THE REAL-WORLD

Yellow chair	Wood stool	Green bin	White shelf	Toy box	Blue chair	White frame	Yellow stair
8/10	10/10	10/10	8/10	10/10	9/10	9/10	7/10

C. Real Robot Experiments

To validate that BiGraspFormer generalizes beyond simulation, we conducted real-robot experiments evaluating grasp success on large, complex-shaped objects placed in diverse poses. Our setup consisted of two UR5e robotic arms, each equipped with a Robotiq 2F-140 gripper, and two Azure Kinect cameras mounted on opposite sides of the workspace

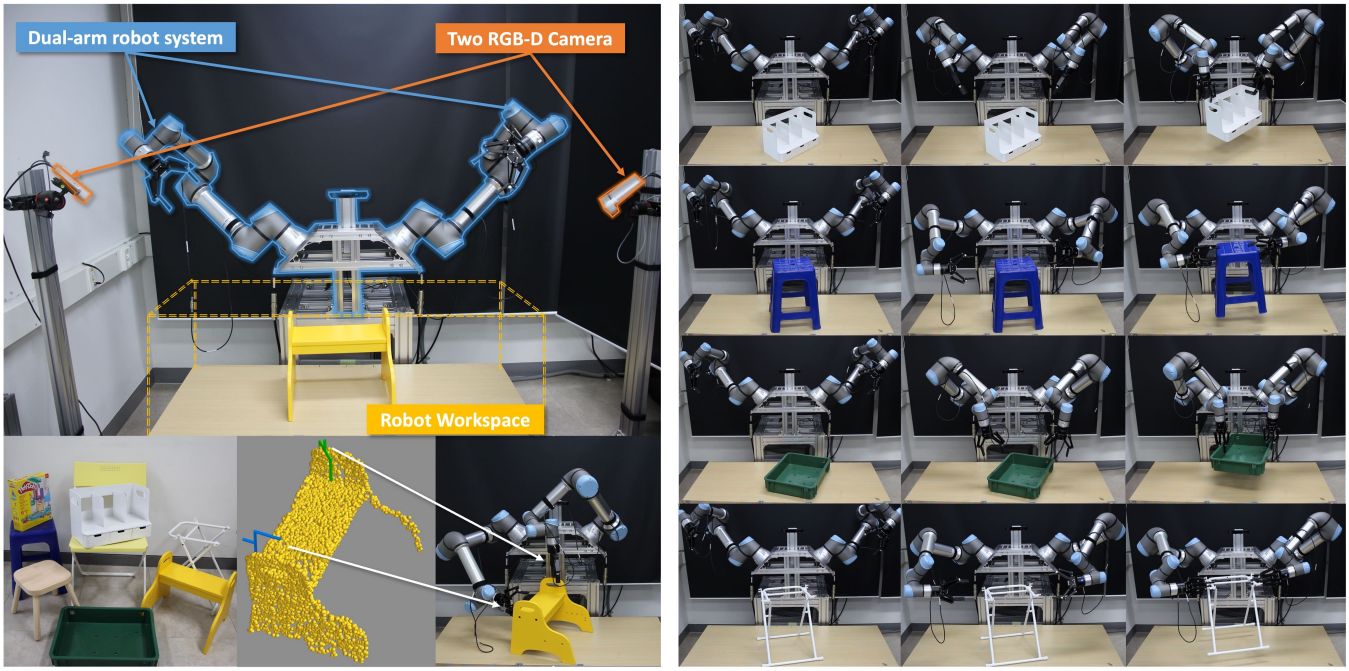

Fig. 5. **Real-World Experimental Setup.** (Left) Dual-arm robotic system with two UR5e arms and two Azure Kinect RGB-D cameras, along with test objects and visualization of feasible bimanual grasps overlaid on object point clouds. (Right) Collision-free trajectory execution of selected grasp poses for grasping and lifting diverse objects.

(Fig. 5). Point clouds captured from both cameras were fused into a unified point cloud, which was provided to the model as input. In each trial, the target object was placed on a table in a different orientation and position. We employed curobo [29] for collision-free motion planning, using the real object point cloud. Grasp candidates predicted by BiGraspFormer were ranked by their quality scores, and the predicted grasps were evaluated in descending order of quality score. For each candidate, motion planning was attempted, and if a collision-free trajectory was found, the corresponding grasp was executed. The robot then lifted the object upward and placed it back down; the attempt was deemed successful if the object remained securely held throughout the lift.

We conducted 10 trials per object, each with a distinct object pose, across a total of eight test objects (Fig. 5). As summarized in Table V, BiGraspFormer successfully grasped and lifted most objects. For simpler large objects such as the toy box and green bin, all trials succeeded. For more challenging shapes such as the yellow chair, white frame, and blue chair, the model still achieved high success rates. Failures mainly stemmed from unstable torque distribution, occasionally leading to slippage during lifting. The yellow stair was the most difficult object due to its heavy weight, where even minor imbalances caused grasp failures. Despite these challenges, BiGraspFormer consistently demonstrated strong performance across all tested objects, confirming its ability to generate reliable bimanual grasps for large and complex objects in real-world environments.

V. Conclusions

We presented BiGraspFormer, the first unified end-to-end framework that directly generates coordinated bimanual grasps from object point clouds using our novel Single-Guided Bimanual (SGB) strategy. The key insight of SGB is to first generate diverse single grasp candidates then leverage their learned features through specialized attention mechanisms to jointly predict bimanual poses and quality scores, effectively reducing the complexity of the 12-DoF search space while ensuring coordinated dual-arm manipulation. This unified approach eliminates the need for separate grasp generation and evaluation modules, enabling explicit modeling of coordination between grasps for stable bimanual manipulation. Extensive simulation and real-world experiments demonstrate that BiGraspFormer consistently outperforms existing methods in both success rate and diversity under normal and disturbance conditions while maintaining efficient inference speed suitable for real-time deployment. Future work will focus on generating action-policy-aware bimanual grasps across diverse objects and poses, integrating our framework with downstream manipulation policies to enable comprehensive bimanual manipulation tasks.

References

[1] K. F. Gbagbe, M. A. Cabrera, A. Alabbas, O. Alyunes, A. Lykov, and D. Tsetserukou, "Bi-vla: Vision-language-action model-based system for bimanual robotic dexterous manipulations," in *2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2024, pp. 2864–2869.

[2] A. Billard and D. Kragic, "Trends and challenges in robot manipulation," *Science*, vol. 364, no. 6446, p. eaat8414, 2019.

[3] H.-S. Fang, C. Wang, M. Gou, and C. Lu, "Graspnet-1billion: A large-scale benchmark for general object grasping," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 444–11 453.

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

[5] 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)*. IEEE, 2021, pp. 13 438–13 444.

[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," *arXiv preprint arXiv:1703.09312*, 2017.

[7] C. Wu, J. Chen, Q. Cao, J. Zhang, Y. Tai, L. Sun, and K. Jia, "Grasp proposal networks: An end-to-end solution for visual learning of robotic grasps," *Advances in Neural Information Processing Systems*, vol. 33, pp. 13 174–13 184, 2020.

[8] A. Alliegro, M. Rudorfer, F. Frattin, A. Leonardis, and T. Tommasi, "End-to-end learning to grasp via sampling from object point clouds," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 9865–9872, 2022.

[9] 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, pp. 8941–8948, 2022.

[10] S. Back, J. Lee, K. Kim, H. Rho, G. Lee, R. Kang, S. Lee, S. Noh, Y. Lee, T. Lee *et al.*, "Graspclutter6d: A large-scale real-world dataset for robust perception and grasping in cluttered scenes," *arXiv preprint arXiv:2504.06866*, 2025.

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

[12] 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.

[13] 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.

[14] C. Borst, M. Fischer, and G. Hirzinger, "Grasp planning: How to choose a suitable task wrench space," in *IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA'04. 2004*, vol. 1. IEEE, 2004, pp. 319–325.

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

[16] 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.

[17] X.-M. Wu, J.-F. Cai, J.-J. Jiang, D. Zheng, Y.-L. Wei, and W.-S. Zheng, "An economic framework for 6-dof grasp detection," in *European Conference on Computer Vision*. Springer, 2024, pp. 357–375.

[18] Y. Tao, J. Wan, H. Liu, H. Gao, and Y. Wen, "Optimal grasping pose selection method for dual-arm robot based on improved genetic algorithm," in *2022 WRC Symposium on Advanced Robotics and Automation (WRC SARA)*. IEEE, 2022, pp. 120–126.

[19] N. Vahrenkamp, M. Przybylski, T. Asfour, and R. Dillmann, "Bimanual grasp planning," in *2011 11th IEEE-RAS International Conference on Humanoid Robots*. IEEE, 2011, pp. 493–499.

[20] F. Cheraghpour, S. A. A. Moosavian, and A. Nahvi, "Multiple aspect grasp performance index for cooperative object manipulation tasks," in *2009 IEEE/ASME International Conference on Advanced Intelligent Mechatronics*. IEEE, 2009, pp. 386–391.

[21] F. Caccavale, V. Lippiello, G. Muscio, F. Pierri, F. Ruggiero, and L. Villani, "Grasp planning and parallel control of a redundant dual-arm/hand manipulation system," *Robotica*, vol. 31, no. 7, pp. 1169–1194, 2013.

[22] Y. Zhao, R. Wu, Z. Chen, Y. Zhang, Q. Fan, K. Mo, and H. Dong, "Dualafford: Learning collaborative visual affordance for dual-gripper manipulation," *arXiv preprint arXiv:2207.01971*, 2022.

[23] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *Advances in neural information processing systems*, vol. 30, 2017.

[24] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[25] M. Savva, A. X. Chang, and P. Hanrahan, "Semantically-enriched 3d models for common-sense knowledge. cvpr 2015 workshop on functionality," *Physics, Intentionality and Causality*, vol. 7, 2015.

[26] 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.

[27] A. Ten Pas and R. Platt, "Using geometry to detect grasp poses in 3d point clouds," *Robotics Research: Volume 1*, pp. 307–324, 2018.

[28] C. Eppner, A. Mousavian, and D. Fox, "A billion ways to grasp: An evaluation of grasp sampling schemes on a dense, physics-based grasp data set," in *The International Symposium of Robotics Research*. Springer, 2019, pp. 890–905.

[29] B. Sundaralingam, S. K. S. Hari, A. Fishman, C. Garrett, K. Van Wyk, V. Blukis, A. Millane, H. Oleynikova, A. Handa, F. Ramos *et al.*, "curobo: Parallelized collision-free minimum-jerk robot motion generation," *arXiv preprint arXiv:2310.17274*, 2023.