User-Grasp-Skill-Driven Robotic Grasping Using Demo-Conditioned Implicit Models

Christian Nicolas Milianti, David Hinwood, Thilina Tharanga Malagalage Don,
Amirali Khodadadian Gostar, and Ehsan Asadi
Royal Melbourne Institute of Technology

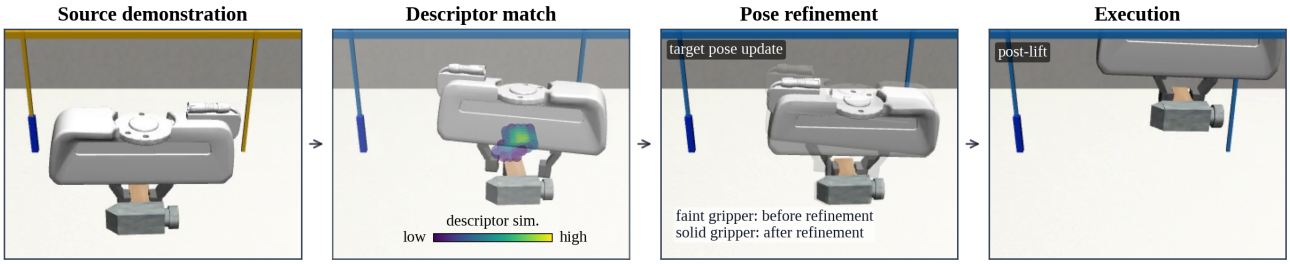

Fig. 1. We present DCIGS, a method for demonstration-conditioned grasp-skill localisation and pose synthesis using a fixed pretrained implicit grasp representation. DCIGS does not require object-specific models, fine-tuning, known object translations, mean-centred TSDFs, or a separate object-tracking module. Given a source demonstration, it queries learned descriptor features at the demonstrated grasp location and searches the target TSDF voxel domain for the most similar descriptor match using cosine similarity. It then initialises a target grasp pose at that matched point using an implicit pretrained grasp rotation model, refines the pose by descriptor optimisation, and executes the resulting grasp. The example shows hammer cleanup demo 8, translation condition 2, using the pile-trained checkpoint and one-view TSDF reconstruction.

***Abstract*—Task-oriented grasping requires selecting object-relative contact regions that support the intended downstream action, not only grasps that are mechanically stable. Existing grasp detectors typically rank target-scene affordance directly, while many demonstration-transfer methods train separate descriptors, leaving open whether a fixed implicit grasp representation already contains a reusable demonstration-conditioned skill signal. We test whether a pretrained implicit grasp representation can serve as a demonstration-conditioned local descriptor for grasp-skill localisation and pose synthesis. Demonstration-Conditioned Implicit Grasp Skills (DCIGS) queries learned descriptor features at a demonstrated grasp, searches the target TSDF voxel domain for the closest descriptor match using cosine similarity, initialises a target grasp pose at the matched point with a pretrained rotation prediction, and refines the pose with a symmetry-aware gripper descriptor objective. The study tests 6 different objects, ten demonstrations per object, nine yaw-zero target translations per demonstration, two pretrained model checkpoints, and one- versus four-view TSDF reconstruction. In the strongest execution configuration, DCIGS reaches a 4.34 mm median localisation error, localises 540/540 target translations within three TSDF voxels (100.0%), reaches 491/540 demonstration-aligned grasp successes (90.9%), and a direct no-refinement ablation improves from 240/540 (44.4%) to 491/540 (90.9%) demonstration-aligned successes. These results support representation reuse under controlled grasp-skill localisation and pose-synthesis conditions, while showing that local descriptor refinement is a major part of pose synthesis.**

I. Introduction

In task-oriented manipulation, grasping is not only about holding an object; it is about choosing a contact region that supports the intended next action. A source demonstration therefore specifies more than a successful grasp pose: it identifies an object-relative, and potentially task-relevant grasp region that should be recovered when the object appears in a target scene. We refer to this as the *grasp skill* demonstrated by the source grasp. This makes representation choice central, because a model that only ranks generic target-scene grasp affordance can select a mechanically stable grasp that may not be usable for downstream manipulation tasks.

Implicit grasp representations are a natural place to look for such a signal. The pretrained implicit grasp model of Jiang et al. couples six-degree-of-freedom (6-DoF) grasp affordance prediction with implicit three-dimensional (3D) reconstruction, producing queryable local features shaped by both geometry and graspability [2]. In the original model, those features are inputs to downstream prediction heads used for grasp affordance and reconstruction. The missing capability tested here is whether the same frozen features can also act as a demonstration-conditioned descriptor: given a source demonstration grasp pose, can they be used to synthesise a grasp-skill aware grasp pose in a target grasp scene?

We propose Demonstration-Conditioned Implicit Grasp Skills (DCIGS), a two-stage grasp-skill localisation and synthesis pipeline. Figure 1 illustrates the pipeline: DCIGS matches learned descriptor features from a demonstrated grasp into the target TSDF, initialises a target grasp pose, and refines it by gripper-attached descriptor optimisation. We evaluate localisation and floating-end-effector execution without object-centred TSDFs, known target translations, or a separate object detector.

Rather than learning a new model from demonstrations, we keep the pretrained implicit grasp model fixed and reuse its internal features for demonstration-conditioned grasp-skill localisation and pose synthesis. This differs from target-scene grasp detectors that directly rank grasps by affordance, because the selected target grasp is conditioned on a demonstrated source grasp. It also differs from local implicit grasp-transfer

methods that learn or align a transfer representation, because DCIGS tests whether an existing affordance-geometry representation already contains a usable demonstration-conditioned skill signal. The gap is therefore representation-centric: existing methods commonly use grasp representations for scoring or transfer, while this work probes whether a fixed implicit grasp representation can act as a reusable task descriptor.

The study is deliberately scoped to voxel-level grasp-skill localisation, local test-time pose refinement initialised from the affordance-model orientation prediction, and floating-end-effector execution of the refined pose. We do not yet evaluate full task execution and leave that for future work. In a controlled yaw-zero translation study over six objects, the strongest execution configuration reaches a 4.34 mm median localisation error, localises all 540 target translations within three TSDF voxels, reaches 491/540 demonstration-aligned grasp successes, and achieves all-nine demonstration-aligned success for 46/60 demonstrations.

The contributions of this paper are:

1) A representation-reuse formulation of *demonstration-conditioned grasp-skill localisation*, where a fixed implicit affordance-geometry feature field maps a source grasp query to a target TSDF voxel.
2) A two-stage DCIGS pipeline that connects voxel localisation to pose synthesis through pretrained orientation initialisation and symmetry-aware gripper-attached descriptor refinement.
3) A controlled evaluation showing 4.34 mm median localisation error, 540/540 localisations within three TSDF voxels, and 491/540 demonstration-aligned grasp successes in the strongest execution configuration, with refinement improving demonstration-aligned success from 240/540 to 491/540.

II. Related Work

We position DCIGS relative to three uses of learned spatial representations: descriptor fields for demonstration transfer, local implicit representations for grasp transfer, and learned 3D feature fields for manipulation correspondence.

Descriptor-field methods show that learned continuous representations can transfer demonstrated manipulation interactions by matching object-relative descriptors. Neural Descriptor Fields use learned descriptor fields to transfer manipulation demonstrations across object instances by matching descriptors associated with demonstrated object-relative interactions [5]. Local Neural Descriptor Fields extend this idea by emphasising locally conditioned descriptors for generalising manipulation skills to novel object categories [1]. Related region-descriptor formulations similarly treat dense representation learning as a mechanism for manipulation-relevant correspondence [4]. DCIGS uses the same broad principle of descriptor matching, but asks a narrower question: whether a representation trained for grasp prediction already contains a usable demonstration-conditioned matching signal.

Local implicit grasp-transfer methods are most directly related because they connect local shape, implicit representations, and grasp transfer. Tekden et al. study grasp transfer using self-aligning implicit representations of local surfaces, transferring demonstrated grasps by identifying geometrically similar local surface regions [6]. Our setting differs in what is learned and what is tested. We do not learn or align a local transfer representation; instead, we keep a pretrained implicit grasp model fixed and probe whether its queried affordance-geometry features already support demonstration-conditioned grasp-skill localisation. Other work has also shown that learned 3D feature fields can support manipulation transfer by matching task-relevant points across scenes [7].

DCIGS differs in both representation and scope: the feature field comes from an implicit grasp detector trained for grasp affordance prediction and 3D reconstruction [2], rather than a visual foundation model, and the experiment isolates grasp-skill localisation and floating-end-effector execution rather than full policy transfer. This positioning matches the representation-centric question of whether a learned robot representation can support downstream behaviour beyond its original prediction head.

III. Problem Formulation

Let $S \in \mathcal{S}$ denote a valid scene. In our setting, we consider a source scene $S_{\mathrm{src}} \in \mathcal{S}$ containing an expert demonstration and a target scene $S_{\mathrm{tgt}} \in \mathcal{S}$ in which the same task object appears under a changed configuration.

We represent the demonstrated grasp as

$$g^D = (p^D, R^D), \tag{1}$$

where $p^D \in \mathbb{R}^3$ is the demonstrated grasp position and $R^D \in SO(3)$ is the demonstrated grasp orientation.

To evaluate localisation in the controlled translation study, we use the known yaw rotation $R_z(\psi) \in SO(3)$ and translation $t \in \mathbb{R}^3$ that generated the target scene. These define the expected grasp-skill location in the target scene

$$p^* = R_z(\psi)p^D + t. \tag{2}$$

Here, p^* is used only as a geometric reference point for validation. DCIGS does not receive $R_z(\psi)$, t, or p^* during descriptor search or pose synthesis.

We denote the pretrained implicit feature field by

$$\Phi_\theta : \mathcal{S} \times \mathbb{R}^3 \to \mathbb{R}^d, \tag{3}$$

where θ denotes the frozen pretrained weights of the implicit feature field from the pretrained grasp model [2]. The model maps an input TSDF to structured XY, XZ, and YZ feature planes, queries those planes at continuous spatial locations, concatenates the three plane samples, and feeds the resulting local descriptors to implicit affordance and occupancy heads. In this paper, $\Phi_\theta(S, p)$ denotes the queried local descriptor at spatial location p: the concatenated tri-plane feature sampled at that query point, before any downstream affordance or occupancy decoding.

The demonstration-conditioned query descriptor is

$$q = \Phi_\theta(S_{\mathrm{src}}, p^D). \tag{4}$$

We denote the target search domain by

$$\Omega(S_{\mathrm{tgt}}) \subset \mathbb{R}^3. \tag{5}$$

In this work, $\Omega(S_{\mathrm{tgt}})$ is the set of voxel centres considered in the target TSDF scene.

The complete problem studied in this paper is therefore twofold:

> Given a demonstrated grasp $g^D = (p^D, R^D)$ in the source scene, can a frozen pretrained implicit grasp representation be used to (i) localise the target-scene voxel at which the same local grasp skill is most likely to apply, and (ii) synthesise a grasp-skill-aware target grasp pose that is locally consistent with the demonstrated gripper–object interaction?

The desired outputs are a localised grasp-skill target point $\hat{p} \in \Omega(S_{\mathrm{tgt}})$ and a refined target grasp pose $\hat{g}_{\mathrm{refined}}$. The location output identifies where the demonstrated local grasp skill is predicted to transfer in the target TSDF. The pose output specifies how the gripper should be placed at that location so that the target grasp remains consistent with the demonstrated gripper–object interaction. The next section describes how these two outputs are computed using single-point descriptor matching followed by symmetry-aware multi-point pose refinement.

IV. Method

DCIGS consists of a localisation stage followed by a synthesis stage. The localisation stage queries the frozen implicit grasp feature field at the demonstrated grasp and searches the target TSDF for the most similar local feature. The synthesis stage initialises a target grasp pose at the localised voxel using the pretrained orientation prediction, then refines that pose by matching gripper-attached descriptor queries with a symmetry-aware objective. This refinement treats the pretrained orientation as an initial condition rather than a final discrete proposal, allowing continuous test-time pose optimisation without retraining the frozen representation.

A. Single-Point Descriptor Localisation

Let S_{src} be the source scene with demonstrated grasp

$$g^D = (p^D, R^D),$$

and let S_{tgt} be the target scene. We denote the frozen implicit grasp feature field from Jiang et al. [2] by

$$\Phi_\theta(S, p) \in \mathbb{R}^d,$$

where p is a spatial query point in scene S. Concretely, $\Phi_\theta(S, p)$ is the concatenation of the pretrained model's XY, XZ, and YZ feature-plane samples at p.

The source demonstration is represented by the local descriptor at the demonstrated grasp position,

$$q = \Phi_\theta(S_{\mathrm{src}}, p^D). \tag{6}$$

For each voxel centre $p \in \Omega(S_{\mathrm{tgt}})$ in the target TSDF search domain, we query the corresponding target descriptor,

$$f(p) = \Phi_\theta(S_{\mathrm{tgt}}, p). \tag{7}$$

The target location is selected by cosine-similarity matching,

$$\hat{p} = \arg\max_{p \in \Omega(S_{\mathrm{tgt}})} \cos(q, f(p)). \tag{8}$$

This stage tests whether the frozen feature queried at a demonstrated grasp is discriminative enough to recover the corresponding target grasp-skill region without training a new descriptor.

The localised target point defines an initial grasp pose

$$g_0 = (\hat{p}, R_0), \tag{9}$$

where R_0 is the grasp-orientation output of the pretrained implicit affordance model [2]. In the main pipeline, this pose is used as the initial condition for local pose refinement. In the direct refinement ablation, g_0 is executed without descriptor refinement under the same floating-end-effector protocol. Using R_0 as an initial condition allows the refined pose to move beyond the fixed finite set of pretrained rotation proposals while still reusing the pretrained model without additional training.

B. Symmetry-Aware Gripper Pose Refinement

The refinement stage replaces the single spatial query with a set of gripper-attached queries. Let

$$\mathcal{X} = \{x_i\}_{i=1}^N, \qquad x_i \in \mathbb{R}^3,$$

be a sparse wireframe descriptor support expressed in the local gripper frame. Appendix A specifies the support used in the experiments. Under the demonstrated grasp, these points define a gripper-relative descriptor pattern in the source scene. Starting from $g_0 = (\hat{p}, R_0)$, we optimise a local pose increment $\Delta g = (\Delta t, \Delta R)$. For a candidate update, the same support points are placed in the target scene at $\hat{p} + \Delta t + R_0 \Delta R x_i$. Let d_i^D denote the source descriptor sampled at $p^D + R^D x_i$, and let $d_i^T(\Delta g)$ denote the target descriptor sampled at $\hat{p} + \Delta t + R_0 \Delta R x_i$. Because a parallel-jaw gripper is symmetric under a left-right finger swap, we score both the identity correspondence and a fixed mirrored permutation π of the support points. The symmetry-aware score is

$$S_{\mathrm{sym}}(\Delta g) = \max_{\sigma \in \{\mathrm{direct}, \pi\}} \frac{1}{N} \sum_{i=1}^{N} \cos\Big(d_i^D, d_{\sigma(i)}^T(\Delta g)\Big). \tag{10}$$

The direct case compares source query point i with target query point i. The π case repeats the comparison after swapping the left- and right-finger query points. We then choose the highest-scoring pose update found during test-time Adam optimisation, $\Delta g^\star = (\Delta t^\star, \Delta R^\star)$, and execute

$$\hat{g}_{\mathrm{refined}} = (\hat{p} + \Delta t^\star, R_0 \Delta R^\star). \tag{11}$$

In practice, Adam evaluates a sequence of candidate updates, including a final post-step candidate, and DCIGS executes the best-scoring evaluated update rather than necessarily the last iterate. The descriptor score reported for the refined pose is recomputed at this selected execution pose. The source descriptors d_i^D remain fixed from the demonstration, and the

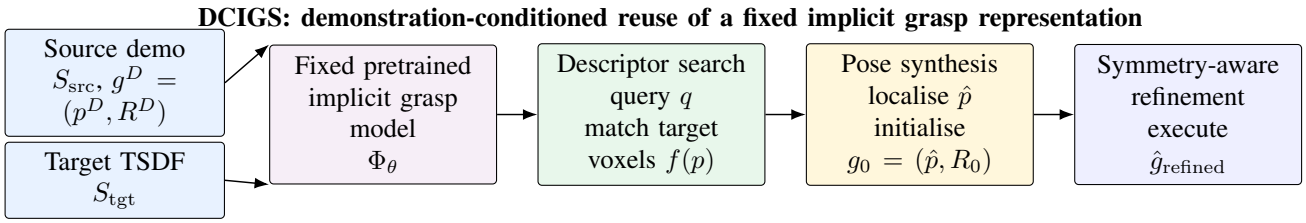

Fig. 2. DCIGS method overview. A source demonstration provides a TSDF and demonstrated grasp. The fixed implicit grasp model is queried at the demonstrated grasp, the closest descriptor match is found in the target TSDF, and the matched voxel is converted into an initial grasp pose. A symmetry-aware refinement step then optimises gripper-attached descriptor queries before execution.

target descriptors $d_i^T(\Delta g)$ change as the pose is refined. The optimisation changes only the test-time grasp pose; it does not update the frozen representation or train a policy.

V. Experimental Protocol

a) Trials.: The experiment evaluates whether demonstration-conditioned representation reuse produces target grasps that preserve the demonstrated grasp region and remain stable under execution. A source demonstration is a MimicGen-derived scene [3] and one demonstrated grasp g^D. We use six scenes: coffee, hammer cleanup, mug cleanup, square, stack, and threading. Each scene contributes ten source demonstrations. A target is the same task object re-anchored in the table-centred workspace under a 3×3 workspace-safe fractional translation grid with yaw fixed at $0°$. Appendix A gives the exact grid construction. The varied factors are the six scenes, ten demonstrations per scene, nine target translations, two pretrained implicit grasp checkpoints, and one- or four-view TSDF reconstruction. Figure 4 visualises the camera layouts used for the one- and four-view reconstructions. The checkpoint labels follow the released pretrained variants from Jiang et al. [2]. In the original data generation for those pretrained variants that we re-utilise, *pile* scenes are formed by dropping objects into a workspace-sized box and removing the box to leave a cluttered pile, while *packed* scenes place objects at random tabletop locations in canonical poses. These labels describe the pretrained checkpoint distributions; DCIGS keeps both checkpoints fixed.

b) Inputs.: DCIGS receives the source demonstration, source TSDF, target TSDF, and frozen pretrained model. The target TSDF is the table-centred workspace TSDF; DCIGS is not given an object-centred crop, detector-provided object mask, or target mean-centring step, and it does not receive the known target translation or expected grasp-skill location in the target TSDF. The top-grasp baseline is a diagnostic, affordance-only baseline: it receives the same target TSDF and same pretrained model, but does not receive the source demonstration descriptor. Its role is not to imitate DCIGS; it tests whether demonstration-conditioned representation reuse adds value beyond selecting a high-affordance target grasp from the original detector.

c) Opening width.: For DCIGS execution, both the pre-refinement ablation and the refined pose use the same source-demo-conditioned approach opening. The opening is the closed gripper width measured in the source demonstration plus four TSDF voxels, i.e., a 30 mm margin in the reported $0.3\,\mathrm{m}$, 40^3-voxel setting. After approach, the floating gripper closes during execution. This keeps the gripper opening fixed between the pre-refinement and refined DCIGS rows, so the ablation isolates pose refinement rather than changing both pose and gripper width opening. The pretrained detector's width prediction is not used for DCIGS execution; it is only used by the affordance-only top-grasp baseline.

d) Metrics.: Localisation is measured by the demonstration-conditioned localisation error: the Euclidean distance in millimetres between $\hat{p}$ and the evaluation-only expected grasp-skill location p^*. We report continuous localisation error and counts within one, two, and three TSDF voxels; one voxel is 7.5 mm in this setup because the 0.3 m TSDF workspace is discretised at 40^3 voxels. The main task metric is *demonstration-aligned grasp success*: the executed grasp must both remain stable under the deterministic shake test and match the demonstrated grasp region within 20 mm and 30 degrees under the symmetry-aware gripper rotation metric. We separately report *stable grasp execution*, which uses only the deterministic shake-test criterion detailed in the appendix. Stable execution is secondary context for comparing mechanical grasping. We also report a diagnostic demonstration-aligned top-grasp score by retrospectively checking whether the affordance-only baseline happened to satisfy the demonstrated grasp-skill criterion.

e) Settings.: Table I reports localisation separately, because descriptor localisation is not itself a task-success metric. Table II reports scene-level and overall task execution for the primary execution configuration, the *pile* checkpoint with one-view TSDF reconstruction, including the DCIGS pre-refinement ablation and the diagnostic top-grasp baseline. We use this checkpoint/view setting as the primary execution setting because several evaluated objects appear in non-canonical tabletop orientations and because it gives the strongest post-refinement grasping result, even though the *packed* one-view setting gives a slightly smaller median localisation error. Appendix A gives the corresponding per-demonstration rows.

VI. Results

Localisation performance is consistently high across the tested checkpoint/view settings. The *packed* one-view setting

TABLE I
DESCRIPTOR LOCALISATION UNDER TARGET TRANSLATIONS.

Counts report target matches within one, two, and three TSDF voxels of the expected grasp-skill location. One voxel is 7.5 mm.

Setting	Median error (mm)	90th pct. error (mm)	Within 1 voxel	Within 2 voxels	Within 3 voxels
pile, 1 view	4.34	7.48	487/540 (90.2%)	539/540 (99.8%)	**540/540 (100.0%)**
packed, 1 view	**4.23**	**7.07**	**496/540 (91.9%)**	**540/540 (100.0%)**	**540/540 (100.0%)**
packed, 4 views	4.46	8.18	471/540 (87.2%)	538/540 (99.6%)	539/540 (99.8%)
pile, 4 views	4.60	8.66	453/540 (83.9%)	537/540 (99.4%)	**540/540 (100.0%)**

TABLE II
PRIMARY EXECUTION RESULTS.

Rows aggregate ten demonstrations and nine translations per scene. Demo-aligned success is the main DCIGS metric; stable execution is mechanical-grasp context. Top grasp is an affordance-only diagnostic baseline with no source descriptor.

	DCIGS without refinement		DCIGS with refinement			Top-grasp baseline		DCIGS gain	
Scene	Stable[a]	Aligned[b]	Aligned[b]	Δ aligned[c]	Stable[a]	Stable[a]	Aligned[b]	Δ stable[d]	Δ aligned[e]
coffee	46/90 (51.1%)	0/90 (0.0%)	81/90 (90.0%)	+90.0 pp	81/90 (90.0%)	59/90 (65.6%)	3/90 (3.3%)	+24.4 pp	+86.7 pp
hammer	86/90 (95.6%)	86/90 (95.6%)	87/90 (96.7%)	+1.1 pp	87/90 (96.7%)	11/90 (12.2%)	0/90 (0.0%)	+84.4 pp	+96.7 pp
mug	26/90 (28.9%)	13/90 (14.4%)	64/90 (71.1%)	+56.7 pp	65/90 (72.2%)	13/90 (14.4%)	0/90 (0.0%)	+57.8 pp	+71.1 pp
square	43/90 (47.8%)	43/90 (47.8%)	79/90 (87.8%)	+40.0 pp	79/90 (87.8%)	1/90 (1.1%)	0/90 (0.0%)	+86.7 pp	+87.8 pp
stack	88/90 (97.8%)	38/90 (42.2%)	90/90 (100.0%)	+57.8 pp	90/90 (100.0%)	74/90 (82.2%)	0/90 (0.0%)	+17.8 pp	+100.0 pp
threading	84/90 (93.3%)	60/90 (66.7%)	90/90 (100.0%)	+33.3 pp	90/90 (100.0%)	40/90 (44.4%)	2/90 (2.2%)	+55.6 pp	+97.8 pp
overall	373/540 (69.1%)	240/540 (44.4%)	**491/540 (90.9%)**	**+46.5 pp**	492/540 (91.1%)	198/540 (36.7%)	5/540 (0.9%)	**+54.4 pp**	**+90.0 pp**

[a]Shake-test success only. [b]Stable execution plus demonstration-relative alignment within 20 mm and 30 degrees. [c]Refined DCIGS aligned success minus DCIGS without refinement aligned success, in percentage points. [d]Refined DCIGS stable execution minus top-grasp stable execution. [e]Refined DCIGS aligned success minus top-grasp aligned success.

has the smallest median localisation error, at 4.23 mm, while the primary *pile* one-view setting is slightly higher, at 4.34 mm. At three-voxel tolerance, both one-view settings localise all 540 target translations. This supports the localisation component of our central claim: a frozen feature queried at the demonstrated grasp can recover the corresponding target grasp-skill region in this controlled translation setting.

The main task result is the demonstration-aligned grasp success of the primary configuration. DCIGS reaches 491/540 demonstration-aligned grasp successes, which requires both stable execution and a demonstration-relative grasp match. Stable grasp execution is reported as secondary mechanical-grasping context: DCIGS reaches 492/540 stable executions versus 198/540 for the diagnostic top-grasp baseline, giving a +54.4 percentage-point DCIGS gain. When the same executed top-grasp baseline poses are retrospectively checked against the demonstrated grasp skill, they reach only 5/540 demonstration-aligned successes, compared with 491/540 for DCIGS. This diagnostic comparison shows that blindly selecting the highest-affordance target grasp rarely recovers the demonstrated grasp skill, even when it sometimes produces a stable grasp. The following comparison explains why *pile* one-view is used as the primary execution setting despite the slightly smaller median localisation error of *packed* one-view. After refinement, *pile* one-view reaches 491/540 demonstration-aligned successes and 492/540 stable executions, compared with 478/540 and 485/540 for *packed* one-view. This choice is also consistent with the evaluated task poses: hammer cleanup and square include side-lying or otherwise non-canonical tabletop object orientations, whereas the *packed* checkpoint label corresponds to canonical placements in the original pretrained-data distribution. Across the remaining aggregate settings, demonstration-aligned success is 440/540 for *packed* four-view and 439/540 for *pile* four-view.

The refinement ablation shows that pose synthesis is not just the pretrained orientation head followed by execution. Localisation is unchanged by construction, but executing g_0 directly gives 240/540 demonstration-aligned successes, while the full refined pipeline gives 491/540. Figure 3 compares the source demonstration, the affordance-only top-grasp baseline, direct execution of g_0, and the refined DCIGS pose for five selected target translations. It shows both selected early pre-lift grasp-pose frames and post-lift frames across the mug, square, stack, threading, and hammer scenes. Within each row, the two DCIGS columns keep the demonstration, checkpoint, TSDF view setting, localised point, and execution opening policy fixed; descriptor refinement is the component that changes the executed pose. Appendix Table IV gives a compact descriptor-score diagnostic for the same paired trials. Coffee shows why stable execution and demonstration-aligned success must be separated: DCIGS pre-refinement execution of g_0 gives 46/90 stable executions but 0/90 demonstration-aligned successes, while refinement raises demonstration-aligned success to 81/90. The improvement is also large where pose alignment matters most: stack improves from 38/90 to 90/90 demonstration-aligned successes, and threading improves from

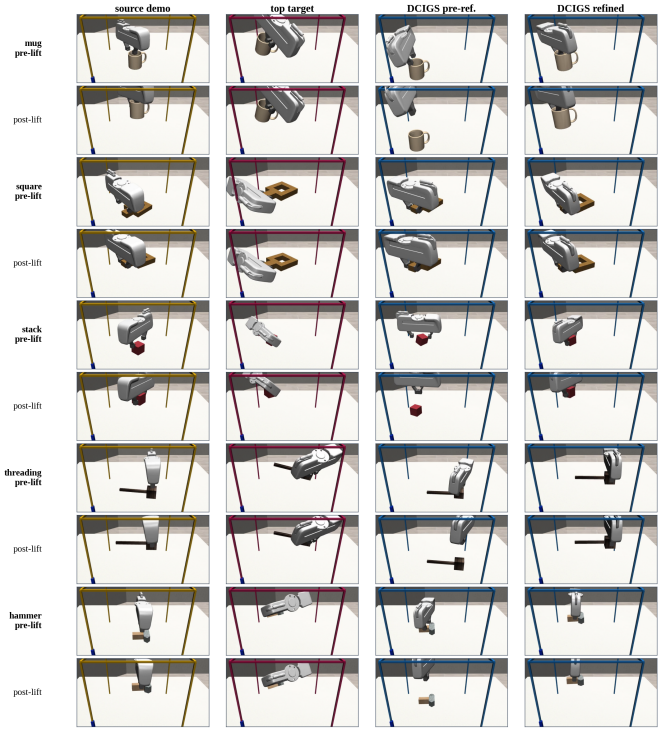

Fig. 3. Qualitative comparison of source demonstrations, target-scene top grasps, and DCIGS executions. Each scene shows pre-lift and post-lift frames for the same selected target translation. The columns compare the source demonstration, the affordance-only top-grasp baseline, DCIGS before refinement, and DCIGS after refinement. The wireframe boxes denote the fixed TSDF workspace for each rollout.

60/90 to 90/90. Stable grasp execution also adds 119/540 successful executions, but the demonstration-aligned success gain is the key evidence that continuous descriptor refinement contributes to grasp-skill-aware pose synthesis. The per-demonstration rows in Appendix A also separate descriptor localisation from execution quality. Mug cleanup demo 2 is a representative contact-sensitive failure case: it localises correctly but fails stable execution, as detailed in Appendix A.

VII. Discussion and Limitations

The main implication is representation-centric: a frozen implicit grasp model trained for affordance and geometry prediction can provide a reusable local descriptor for a task-suitable demonstration-conditioned grasping objective. In this controlled source-target setting, querying the representation with a demonstrated grasp exposes enough local structure to localise corresponding grasp regions and synthesise executable grasp poses more often than affordance-only target ranking. The refinement ablation further shows that the target voxel and pretrained orientation are not sufficient on their own; continuous descriptor optimisation is a major part of converting localisation into pose synthesis. The checkpoint comparison also shows that median localisation error alone is not the right selection criterion for the final method setting. The *packed* one-view checkpoint is slightly tighter for descriptor localisation, but the *pile* one-view checkpoint gives better post-refinement execution overall.

The failure modes are also informative. Source contact quality affects execution outcomes: a well-localised contact region can still fail if the demonstrated grasp has little tolerance to small pose errors. Descriptor refinement improves the aggregate distribution, but some stable grasps can still be inconsistent with the demonstrated contact geometry.

Several limitations constrain the generality of these findings. The evaluation is restricted to controlled yaw-zero translations within MimicGen-derived scenes, leaving open questions about broader pose variation, category generalisation, and real-world applicability. The controlled setting is nevertheless the intended test: it isolates whether the representation contains a grasp-skill correspondence signal before adding harder perception confounds such as distractor objects in the same scene. The method depends on the frozen representation's learned grasp manifold; if the pretrained model does not encode the relevant local affordance structure, descriptor matching can select geometrically similar but task-inconsistent regions. The baseline set is limited to affordance-only top grasping and the DCIGS pre-refinement ablation, and the evaluation does not yet include full task execution or real-robot transfer. Future work should test richer baselines, non-yaw-zero target changes, full task policies, and real-robot execution.

VIII. Conclusion

This paper tests whether a frozen implicit grasp representation can be reused as a demonstration-conditioned local skill descriptor. To our knowledge, DCIGS is the first method to use frozen internal features of a pretrained TSDF-input implicit 6-DoF grasp detector as a demonstration-conditioned dense correspondence field for grasp-skill localisation and pose refinement. DCIGS localises a target grasp region by matching queried features from a source demonstration, then synthesises an execution pose through pretrained orientation initialisation and symmetry-aware gripper-attached descriptor refinement. In the strongest controlled yaw-zero execution configuration, DCIGS reaches a 4.34 mm median localisation error, localises all 540 target translations within three TSDF voxels (100.0%), reaches 491/540 demonstration-aligned grasp successes (90.9%), and 46/60 demonstrations achieve demonstration-aligned success on all nine translations. A direct ablation shows that DCIGS pre-refinement execution of g_0 reaches only 240/540 demonstration-aligned successes (44.4%), while refinement reaches 491/540 (90.9%). The result supports the view that representation design can affect downstream robot behaviour beyond the original grasp-scoring objective, while leaving broader policy learning, broader pose generalisation, and real-robot transfer as future work.

Acknowledgements

This research was supported by the Commonwealth through an Australian Government Research Training Program Scholarship [DOI: https://doi.org/10.82133/C42F-K220].

REFERENCES

[1] Ethan Chun, Yilun Du, Anthony Simeonov, Tomas Lozano-Perez, and Leslie Kaelbling. Local Neural Descriptor Fields: Locally Conditioned Object Representations for Manipulation, March 2023. URL http://arxiv.org/abs/2302.03573. arXiv:2302.03573 [cs].

[2] Zhenyu Jiang, Yifeng Zhu, Maxwell Svetlik, Kuan Fang, and Yuke Zhu. Synergies Between Affordance and Geometry: 6-DoF Grasp Detection via Implicit Representations. In *Robotics: Science and Systems XVII*. Robotics: Science and Systems Foundation, July 2021. ISBN 978-0-9923747-7-8. doi: 10.15607/RSS.2021.XVII.024. URL http://www.roboticsproceedings.org/rss17/p024.pdf.

[3] Ajay Mandlekar, Soroush Nasiriany, Bowen Wen, Iretiayo Akinola, Yashraj Narang, Linxi Fan, Yuke Zhu, and Dieter Fox. MimicGen: A Data Generation System for Scalable Robot Learning using Human Demonstrations, October 2023. URL http://arxiv.org/abs/2310.17596. arXiv:2310.17596 [cs].

[4] Anish Pratheepkumar, Markus Ikeda, Michael Hofmann, Fabian Widmoser, Andreas Pichler, and Markus Vincze. NRDF - Neural Region Descriptor Fields as Implicit ROI Representation for Robotic 3D Surface Processing. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 12955–12962, Abu Dhabi, United Arab Emirates, October 2024. IEEE. ISBN 979-8-3503-7770-5. doi: 10.1109/IROS58592.2024.10802862. URL https://ieeexplore.ieee.org/document/10802862/.

[5] Anthony Simeonov, Yilun Du, Andrea Tagliasacchi, Joshua B. Tenenbaum, Alberto Rodriguez, Pulkit Agrawal, and Vincent Sitzmann. Neural Descriptor Fields: SE(3)-Equivariant Object Representations for Manipulation, December 2021. URL http://arxiv.org/abs/2112.05124. arXiv:2112.05124 [cs].

[6] Ahmet Tekden, Marc Peter Deisenroth, and Yasemin Bekiroglu. Grasp Transfer based on Self-Aligning Implicit Representations of Local Surfaces. *IEEE Robotics and Automation Letters*, 8(10):6315–6322, October 2023. ISSN 2377-3766, 2377-3774. doi: 10.1109/LRA.2023.3306272. URL http://arxiv.org/abs/2308.07807. arXiv:2308.07807 [cs].

[7] Qianxu Wang, Haotong Zhang, Congyue Deng, Yang You, Hao Dong, Yixin Zhu, and Leonidas Guibas. SparseDFF: Sparse-View Feature Distillation for One-Shot Dexterous Manipulation, March 2024. URL http://arxiv.org/abs/2310.16838. arXiv:2310.16838 [cs].

APPENDIX A
PER-DEMONSTRATION PRIMARY RESULTS

Table V reports the per-demonstration results for the primary pile/one-view setting. Each row contains the nine yaw-zero target translations generated from one source demonstration. Demonstration-aligned grasp success is the main demo-conditioned task metric; stable grasp execution is reported as secondary stability context.

a) Mug cleanup demo 2.: Mug cleanup demo 2 is retained for balanced reporting and illustrates a contact-sensitive failure mode. It localises all 9/9 target translations and the refined poses satisfy the demonstration-relative alignment test in all 9/9 cases, but all nine executions fail stable execution because the object does not lift sufficiently before the shake test. Manual inspection shows that this source demonstration grasps the mug using a narrow contact involving the mug interior and handle, rather than a more forgiving exterior contact region. This makes the transferred grasp sensitive to millimetre-scale optimisation error: a pose can be close to the demonstration under the geometric test but still miss the contact needed for a stable lift. Its 0/9 demonstration-aligned success is therefore an execution/contact-quality failure rather than a descriptor-localisation failure.

b) Translations.: Target placements are generated from a 3×3 fractional grid over the valid tabletop/TSDF workspace region. The valid region is computed per source demonstration so that the target object remains inside both the table bounds and the TSDF workspace, using a conservative object footprint estimate and a 2 cm safety margin. The grid fractions are $\{0.08, 0.5, 0.92\}$ along each axis, yaw is fixed at $0°$, and the z offset is zero. Fractions are applied inside this valid region, so $(0.08, 0.08)$ means 8% along the valid x range and 8% along the valid y range. The stored metric translation is the selected target centre minus the source object centre, so the translation scale varies slightly by object and demonstration. Table III summarises the resulting signed offset range by scene.

TABLE III
TARGET TRANSLATION RANGES.

Values are millimetres. The same signed range is used for x and y.

Scene	Signed x/y offset range
coffee	± 80.3 mm
hammer cleanup	± 60.8 mm
mug cleanup	± 67.8 mm
square	± 44.2 mm
stack	± 84.0 mm
threading	± 39.6 mm

c) Camera views.: Figure 4 shows the deterministic side-view camera geometry used to build the one-view and four-view TSDF reconstructions. The geometry uses a $0.3\,\mathrm{m}$ workspace, a camera radius scale of 2.0, and $30°$ elevation above horizontal. Here radius scale means the multiplier applied to the workspace size to set the camera distance from the look-at point, so the camera radius is $2.0 \times 0.3\,\mathrm{m} = 0.6\,\mathrm{m}$. The one-view setting uses azimuth $-90°$, while the four-view setting uses azimuths $0°$, $90°$, $180°$, and $-90°$.

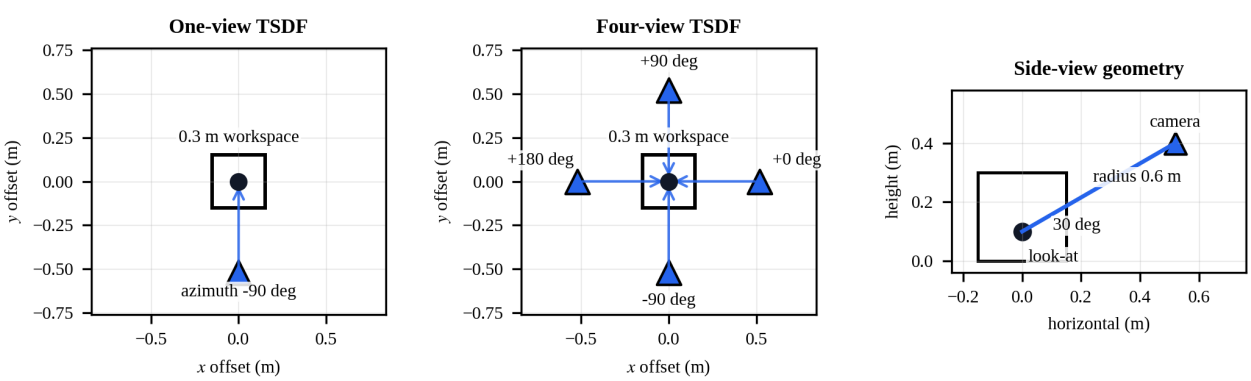

Fig. 4. TSDF camera geometry used for the reported reconstructions. The one-view setting uses a single side-view camera at azimuth $-90°$. The four-view setting uses cameras at $0°$, $90°$, $180°$, and $-90°$. All views use a $0.3\,\mathrm{m}$ workspace, $0.6\,\mathrm{m}$ camera radius, and $30°$ elevation.

d) Shake test.: For reproducibility, the deterministic shake test uses two cycles over the world x and y axes, 48 simulation steps per segment, and an 80-step final hold. A trial passes if the object still satisfies the 50 mm lift threshold, remains at least 5 mm table-clear, translates by no more than 30 mm, and rotates by no more than 35 degrees.

e) Search domain.: In the experiments, $\Omega(S_{\mathrm{tgt}})$ includes all target TSDF voxel centres. The same formulation could be restricted to narrower search domains, such as surface-near voxels, occupied voxels, voxels inside a local crop, or voxels passing a validity mask.

f) Refinement support.: The refinement support $\mathcal{X}$ is generated from the loaded robosuite parallel-jaw gripper geometry in the local gripper frame. The code derives four template anchors from the gripper visual mesh: rear and front points on the left finger, and rear and front points on the right finger. For each source demonstration, the left-right separation is scaled to the closed demonstration gripper width used for descriptor sampling. The support is then sampled along four line segments: the left finger, the right finger, the rear bridge between the fingers, and a short approach-depth segment extending 55 mm backward from the bridge midpoint along the local gripper z axis. Points are sampled at approximately one source TSDF voxel spacing, which is 7.5 mm in the reported $0.3\,\mathrm{m}$, 40^3-voxel runs. Duplicate samples are removed, and the exact bridge-midpoint query anchor used for single-point localisation is appended so that the refinement objective contains the same anchor as the localisation stage. The finger-swap symmetry π is the nearest-neighbour permutation induced by reflecting these local support points across the gripper's left-right x axis.

TABLE IV
DESCRIPTOR-SCORE CHANGE AFTER REFINEMENT.

Scores use the internal symmetry-aware refinement objective. The after-refinement score is measured at the selected execution pose. Higher is better.

Quantity	Before refinement g_0	After selected pose	Score increase
Mean descriptor score	0.8990	0.9803	+0.0812
Median descriptor score	0.9197	0.9821	+0.0624

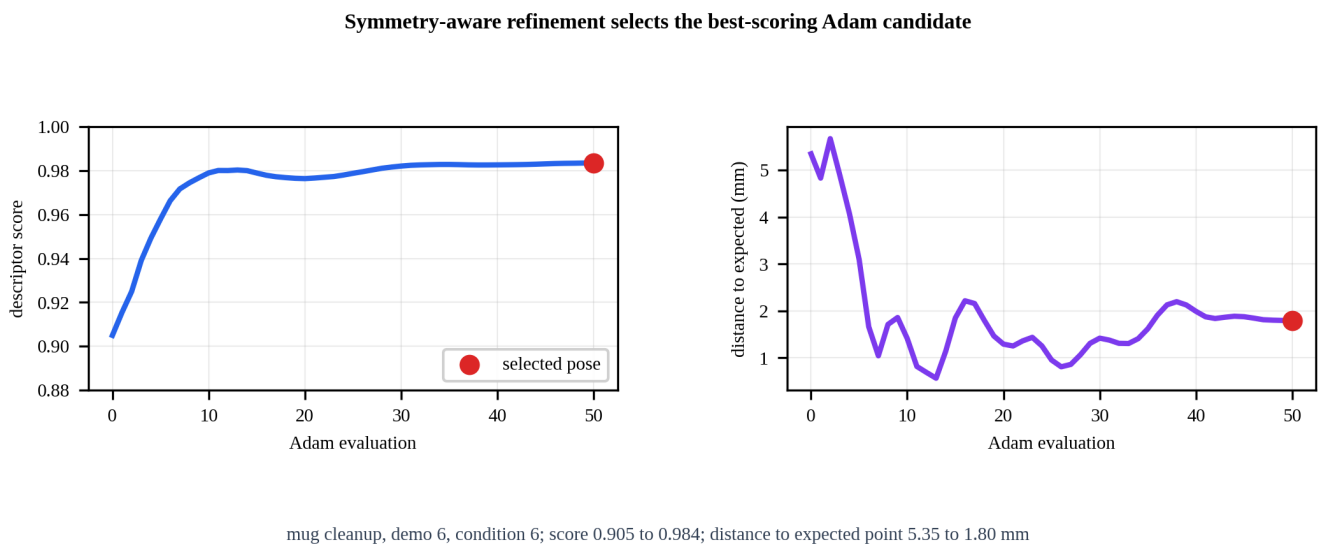

Fig. 5. Example descriptor-refinement trace for a mug-cleanup trial. The optimiser evaluates candidate gripper poses and selects the best-scoring descriptor match for execution. In this example, refinement increases the descriptor score from 0.905 to 0.984 and reduces the distance to the expected grasp-skill location from 5.35 mm to 1.80 mm.

TABLE V

PER-DEMONSTRATION PRIMARY PILE/ONE-VIEW RESULTS.

Counts are out of nine target translations per source demonstration. Localisation is reported within 20 mm. Top grasp is an affordance-only diagnostic baseline.

Demo	Localised $\leq$ 20 mm	DCIGS aligned[a]	DCIGS stable[b]	Top grasp stable[b]	Δ stable[c]
coffee 0	9/9 (100%)	9/9 (100%)	9/9 (100%)	6/9 (67%)	+33 pp
coffee 1	9/9 (100%)	9/9 (100%)	9/9 (100%)	7/9 (78%)	+22 pp
coffee 2	9/9 (100%)	9/9 (100%)	9/9 (100%)	6/9 (67%)	+33 pp
coffee 3	9/9 (100%)	9/9 (100%)	9/9 (100%)	8/9 (89%)	+11 pp
coffee 4	9/9 (100%)	9/9 (100%)	9/9 (100%)	7/9 (78%)	+22 pp
coffee 5	9/9 (100%)	6/9 (67%)	6/9 (67%)	5/9 (56%)	+11 pp
coffee 6	9/9 (100%)	6/9 (67%)	6/9 (67%)	4/9 (44%)	+22 pp
coffee 7	9/9 (100%)	9/9 (100%)	9/9 (100%)	6/9 (67%)	+33 pp
coffee 8	9/9 (100%)	9/9 (100%)	9/9 (100%)	4/9 (44%)	+56 pp
coffee 9	9/9 (100%)	6/9 (67%)	6/9 (67%)	6/9 (67%)	+0 pp
hammer cleanup 0	9/9 (100%)	9/9 (100%)	9/9 (100%)	0/9 (0%)	+100 pp
hammer cleanup 1	9/9 (100%)	9/9 (100%)	9/9 (100%)	1/9 (11%)	+89 pp
hammer cleanup 2	9/9 (100%)	9/9 (100%)	9/9 (100%)	2/9 (22%)	+78 pp
hammer cleanup 3	9/9 (100%)	9/9 (100%)	9/9 (100%)	1/9 (11%)	+89 pp
hammer cleanup 4	9/9 (100%)	9/9 (100%)	9/9 (100%)	1/9 (11%)	+89 pp
hammer cleanup 5	9/9 (100%)	9/9 (100%)	9/9 (100%)	0/9 (0%)	+100 pp
hammer cleanup 6	9/9 (100%)	9/9 (100%)	9/9 (100%)	1/9 (11%)	+89 pp
hammer cleanup 7	9/9 (100%)	9/9 (100%)	9/9 (100%)	4/9 (44%)	+56 pp
hammer cleanup 8	9/9 (100%)	9/9 (100%)	9/9 (100%)	1/9 (11%)	+89 pp
hammer cleanup 9	9/9 (100%)	6/9 (67%)	6/9 (67%)	0/9 (0%)	+67 pp
mug cleanup 0	9/9 (100%)	9/9 (100%)	9/9 (100%)	3/9 (33%)	+67 pp
mug cleanup 1	9/9 (100%)	6/9 (67%)	6/9 (67%)	0/9 (0%)	+67 pp
mug cleanup 2	9/9 (100%)	0/9 (0%)	0/9 (0%)	1/9 (11%)	-11 pp
mug cleanup 3	9/9 (100%)	9/9 (100%)	9/9 (100%)	0/9 (0%)	+100 pp
mug cleanup 4	9/9 (100%)	9/9 (100%)	9/9 (100%)	2/9 (22%)	+78 pp
mug cleanup 5	9/9 (100%)	4/9 (44%)	4/9 (44%)	1/9 (11%)	+33 pp
mug cleanup 6	9/9 (100%)	7/9 (78%)	8/9 (89%)	0/9 (0%)	+89 pp
mug cleanup 7	9/9 (100%)	3/9 (33%)	3/9 (33%)	5/9 (56%)	-22 pp
mug cleanup 8	9/9 (100%)	9/9 (100%)	9/9 (100%)	0/9 (0%)	+100 pp
mug cleanup 9	9/9 (100%)	8/9 (89%)	8/9 (89%)	1/9 (11%)	+78 pp
square 0	9/9 (100%)	9/9 (100%)	9/9 (100%)	1/9 (11%)	+89 pp
square 1	9/9 (100%)	9/9 (100%)	9/9 (100%)	0/9 (0%)	+100 pp
square 2	9/9 (100%)	9/9 (100%)	9/9 (100%)	0/9 (0%)	+100 pp
square 3	9/9 (100%)	6/9 (67%)	6/9 (67%)	0/9 (0%)	+67 pp
square 4	9/9 (100%)	9/9 (100%)	9/9 (100%)	0/9 (0%)	+100 pp
square 5	9/9 (100%)	8/9 (89%)	8/9 (89%)	0/9 (0%)	+89 pp
square 6	9/9 (100%)	9/9 (100%)	9/9 (100%)	0/9 (0%)	+100 pp
square 7	9/9 (100%)	6/9 (67%)	6/9 (67%)	0/9 (0%)	+67 pp
square 8	9/9 (100%)	5/9 (56%)	5/9 (56%)	0/9 (0%)	+56 pp
square 9	9/9 (100%)	9/9 (100%)	9/9 (100%)	0/9 (0%)	+100 pp
stack 0	9/9 (100%)	9/9 (100%)	9/9 (100%)	6/9 (67%)	+33 pp
stack 1	9/9 (100%)	9/9 (100%)	9/9 (100%)	7/9 (78%)	+22 pp
stack 2	9/9 (100%)	9/9 (100%)	9/9 (100%)	7/9 (78%)	+22 pp
stack 3	9/9 (100%)	9/9 (100%)	9/9 (100%)	7/9 (78%)	+22 pp
stack 4	9/9 (100%)	9/9 (100%)	9/9 (100%)	8/9 (89%)	+11 pp
stack 5	9/9 (100%)	9/9 (100%)	9/9 (100%)	8/9 (89%)	+11 pp
stack 6	9/9 (100%)	9/9 (100%)	9/9 (100%)	9/9 (100%)	+0 pp
stack 7	9/9 (100%)	9/9 (100%)	9/9 (100%)	9/9 (100%)	+0 pp
stack 8	9/9 (100%)	9/9 (100%)	9/9 (100%)	6/9 (67%)	+33 pp
stack 9	9/9 (100%)	9/9 (100%)	9/9 (100%)	7/9 (78%)	+22 pp
threading 0	9/9 (100%)	9/9 (100%)	9/9 (100%)	2/9 (22%)	+78 pp
threading 1	9/9 (100%)	9/9 (100%)	9/9 (100%)	7/9 (78%)	+22 pp
threading 2	9/9 (100%)	9/9 (100%)	9/9 (100%)	7/9 (78%)	+22 pp
threading 3	9/9 (100%)	9/9 (100%)	9/9 (100%)	5/9 (56%)	+44 pp
threading 4	9/9 (100%)	9/9 (100%)	9/9 (100%)	3/9 (33%)	+67 pp
threading 5	9/9 (100%)	9/9 (100%)	9/9 (100%)	1/9 (11%)	+89 pp
threading 6	9/9 (100%)	9/9 (100%)	9/9 (100%)	1/9 (11%)	+89 pp
threading 7	9/9 (100%)	9/9 (100%)	9/9 (100%)	2/9 (22%)	+78 pp
threading 8	9/9 (100%)	9/9 (100%)	9/9 (100%)	7/9 (78%)	+22 pp
threading 9	9/9 (100%)	9/9 (100%)	9/9 (100%)	5/9 (56%)	+44 pp

[a]Stable execution plus demonstration-relative alignment within 20 mm and 30 degrees. [b]Shake-test success only. [c]DCIGS stable execution minus top-grasp stable execution.