Scalable Unseen Objects 6-DoF Absolute Pose Estimation with Robotic Integration

(Responses to Editor and Reviewers)

IEEE Transactions on Robotics,

Manuscript ID: 25-1276.R1

Dear Editor and Reviewers:

We wish you all a great start to the New Year!

Thank you very much for carefully reviewing our manuscript and for reconsidering its revision. The comments from the editor and reviewers were highly constructive and very helpful for improving the quality of our manuscript. We have carefully revised the manuscript considering these suggestions. Below we give point-by-point responses to each comment made by the editor and reviewers, within the 10-page limit as required by the journal. Comments are in **text box** and our responses are in *italics*. Modifications in the manuscript are highlighted with blue color so they can be easily located and matched to the revised manuscript.

Yours sincerely,

All the Authors

Editor Comments

The reviewers concur that this is a clearly written, technically solid paper introducing SinRef-6D, a single-reference-view 6-DoF pose estimation framework based on an SSM backbone with iterative refinement, supported by extensive experiments on multiple benchmarks. The method is seen as a meaningful, though not radically novel, step toward scalable pose estimation without CAD models or dense reference views, with well-chosen baselines and thorough ablations.

The main reservations, e.g., from Reviewer 3, concern the positioning and validation of the core contribution. The paper does not yet provide direct quantitative comparisons to existing single-reference/single-view methods discussed in Sec. II-C, even though the same user-driven absolute pose labeling could be applied to them. This leaves the incremental benefit of the proposed backbone and pipeline insufficiently established. In addition, the downstream grasping evaluation is not fully convincing as a showcase of accurate 6-DoF pose use. In particular, the grasping policy is simple, the objects are relatively easy, and, as Reviewer 2 requests, a more systematic real-world study with greater variation in objects, poses, and clutter is required. Further issues include clarifying the claims around “human–robot collaboration” and the integrated hardware–software system, a more principled discussion and ablation of the two-GeoTransformer design, and some targeted robustness and terminology refinements. Addressing these points in a substantial revision would significantly strengthen the work for any resubmission.

Response to Editor:

We appreciate the efforts of the Editor in handling the review process, which has certainly helped us to improve the quality of our manuscript. All the comments have been thoroughly addressed. ***First****, following the comments of Reviewer 3, we substantially revised the related work and experimental sections by adding direct quantitative comparisons to representative single-reference methods discussed in Sec. II-C, under unified segmentation and evaluation protocols. These new results explicitly demonstrate the incremental benefit of the proposed SSM-based backbone and pipeline design.* ***Second****, we significantly expanded the real-world robotic grasping evaluation, increasing both the number and diversity of unseen objects, the number of grasping trials, and the difficulty of tested scenarios (including non-planar placements, large reference–query pose discrepancies, clutter, dark, and occlusions). We also clarified how the estimated 6-DoF object poses are explicitly utilized by the downstream grasping policy, thereby better showcasing the role of accurate absolute pose estimation.* ***In addition****, we provided a more principled discussion and ablation of the two-GeoTransformer design, added targeted robustness analyses and failure case studies, and clarified the claims related to human–robot collaboration and the integrated hardware–software system.* ***Collectively****, these revisions result in a more clearly positioned, thoroughly validated, and methodologically grounded presentation of SinRef-6D. We believe the revised manuscript substantially addresses the Editor’s and Reviewers’ concerns and hope the revised manuscript can convince the Editor.*

Reviewer 1

Summary Comments to the Author

SinRef-6D addresses the task of 6DoF pose estimation from RGB-D images. The paper overcomes key challenges related to the scalability of prior learning-based pose estimation methods. Namely, the authors train a state space model to iteratively refine object pose estimates from a single RGB-D image. This overcomes key limitations of prior work that rely on detailed CAD reference models, which are difficult to acquire from in-the-wild datasets. The key strengths in this paper are the thorough baselines and ablations that justify the necessity of each contribution, strong quantitative results against similar methods, and detailed experimental methodology. Several minor critiques would improve clarity regarding why this approach performs better and some details regarding the approach’s sensitivity to scene and object variations.

Response to Comments:

We sincerely thank you for the positive and constructive summary. We have carefully considered all the comments and have revised the *manuscript accordingly to further improve clarity and completeness. In particular, we have expanded the discussion to more explicitly contrast our method with the most directly comparable single-reference-view approaches and added additional experiments and discussions on the sensitivity of our model to object and scene variations. These revisions are intended to make both the technical contributions and the experimental insights clearer to readers.*

The Reviewer's Comment 1.1

Overall, the benefits and limitations addressed by this paper were clear and well-written. One minor nitpick is that the last contribution in the introduction section, "We develop an integrated hardware-software robotic system…" is not adequately justified as to how it improves upon prior object-centric manipulation setups. If the authors feel that their setup has made key innovations to improve upon existing solutions, it is important to explicitly highlight these improvements in Section IV and the introduction. In the current manuscript, it seems like many of the improvements mentioned in section IV are relevant for supporting contribution 1 (more efficient 6DoF pose estimation), but not sufficient for differentiating from prior hardware-software systems for annotating object poses. One way to improve this is to separate the "Extensive experiments demonstrate…" section in the last contribution to a separate paragraph and merge the "We develop an integrated hardware-software…" section with the first contribution. Unless their hardware-software setup introduces a novel contribution, it would be sufficient to state that they validated the efficacy of their task setup and annotation method on an integrated hardware-software system to highlight that the authors have corroborated their claim in the real world.

Response to Comments:

We sincerely thank you for this insightful comment and fully agree that the current description may over-emphasize our integrated hardware–software system as a standalone contribution. Our main novelty indeed lies in the single-reference task formulation and the SinRef-6D framework, while the robotic system primarily serves as a practical instantiation to validate scalability in real-world grasping. In the revised manuscript, we have accordingly softened and reorganized our contribution statements: In the Introduction, we now explicitly present the hardware–software system as an experimental validation platform rather than an independent contribution. Specifically, we follow your suggestion to merge the previous statement "We develop an integrated hardware–software robotic system..." with the first contribution, and clarify that the system is used to validate the proposed task setup, pose estimation framework, and annotation method in real-world grasping scenarios, rather than to introduce a novel hardware architecture. We believe these changes better align the contribution claims with the actual technical novelty.

Changes Made in the Manuscript:

Section I-Contribution-First Paragraph: ...We further develop an integrated hardware-software robotic system tailored to the proposed task setup and framework, validating their efficacy in real-world scenarios.

Section I-Contribution-Last Paragraph: Delete: We develop an integrated hardware-software robotic system tailored to the proposed task setup and corresponding framework in real-world scenarios.

The Reviewer's Comment 1.2

Additionally, the introduction section claims that the authors design. Novel task setup grounded in human-robot collaboration. This terminology is usually used to indicate some interleaving between an autonomous robot process combined with human-in-the-loop corrections or refinement. However, in the proposed task setup, it seems like the human is simply teleoperating the robot to collect sample views for annotation. To avoid confusion regarding this term, the reviewer recommends reframing this contribution as a scalable label collection pipeline rather than a novel task setup.

Response to Comments:

We appreciate your clarification regarding the terminology. We agree that in our current implementation, the human operator mostly teleoperates the robot and uses a semi-automatic pose annotator, which is different from more interactive human-in-the-loop control paradigms. To avoid confusion and better reflect what we actually implement, we have revised the related text in the Introduction from "a novel task setup" to "a scalable label collection pipeline" as you suggested.

Changes Made in the Manuscript:

Section I-Paragraph 4:

"...we design a novel task setup where each unseen object..." -> "...we design a scalable label collection pipeline where each unseen object..."

The Reviewer's Comment 1.3

In the related works section, it would be helpful for the reader to select the most directly comparable prior work and expand briefly on the key differences with this past work. Doing this would also help the reader understand the rationale behind their choice of baselines more effectively as it seems like there are many prior works that investigate this problem. It would also help the authors justify their intuition for why their application of SSMs should enable better performance compared to other methods like FoundationPose, which also iteratively refines the object pose estimate using a transformer-based architecture instead.

Response to Comments:

We thank you for this valuable suggestion. Following this comment, we have expanded the discussion in Sec. II.C to more explicitly contrast *our method with the most directly comparable single-reference-view approaches, including UNOPose, One2Any, and Any6D (also added the quantitative comparison in Section V-C-3) and Table V). We clarify that while these methods significantly reduce onboarding cost, they primarily estimate relative pose between the reference and query views, which is insufficient for robotic manipulation that requires absolute object poses for action execution. Our work differs in problem formulation and design goal, focusing on single-reference-view absolute pose estimation under robotic manipulation constraints. We also emphasize that, unlike transformer-based frameworks such as FoundationPose that rely on CAD models or dense reference views, our SSMs-based backbone is specifically designed to model long-range spatial dependencies under limited geometric information from a single reference view, which is critical in our task setting.*

Changes Made in the Manuscript:

Section II-C-Paragraph 2: More recently, some works [76-78] have explored pose estimation using a single RGB-D reference view to reduce onboarding cost for unseen objects. UNOPose [76] incorporates depth data and proposes a one-reference-based pose estimation framework that constructs an SE(3)-invariant reference representation and adaptively weights correspondences to handle low viewpoint overlap. One2Any [77] further introduces a category-agnostic method for 6-DoF object pose estimation that leverages a reference-query RGB-D pair to generate pose embeddings and decode object coordinates. Any6D [78] estimates both object pose and size from an RGB-D anchor image by leveraging joint object alignment and a render-and-compare strategy. Despite their effectiveness, these methods primarily focus on relative pose estimation between the reference and query views, which is insufficient for robotic manipulation scenarios where absolute object poses in a common coordinate system are required for action execution. In contrast, our work targets single-reference 6-DoF absolute pose estimation under robotic manipulation settings. To this end, we introduce a human-robot collaborative reference acquisition and annotation pipeline, a single reference view-based point cloud focalization strategy to establish a common coordinate system, and SSMs-based feature extraction networks tailored for the limited geometric and spatial information available from a single view. This problem-driven design enables direct deployment in manipulation pipelines while maintaining scalability to unseen objects.

Section II-C-Paragraph 3: Related works such as FoundationPose [61] also employ transformer-based architectures for iterative pose refinement; however, our SSM-based backbone is explicitly designed to model long-range spatial dependencies under severely limited geometric information, which is particularly critical in our single-reference setting.

The Reviewer's Comment 1.4

Overall, the experiments are well designed and support this method's design choices. However, one large design choice that was not supported quantitatively is why two GeoTransformer models are needed. This has clear disadvantages and would be helpful to add an offline ablation study comparing the performance with and without two models. The authors can add this to Table V. Lastly, it seems like most of the objects in the qualitative results are on a flat planar surface (not necessarily planar with the camera). It seems that if the object was not stably resting on a flat planar surface, estimating the pose would be more challenging, as the normal priors from the table are less helpful for inferring the pose. It would be helpful for practitioners interested in deploying this model to provide an additional qualitative analysis in Figure 8, examining SinRef's robustness to non-planar aligned objects, as this is common during actual robot deployments. Furthermore, it would be helpful to provide a more detailed explanation regarding how different the randomly selected query views are. One concern is that if the viewing angles or occlusions are similar across the query views, then the variance is more attributed to the model's sensitivity to minor scene differences (high-frequency visual details) versus large scene differences (occlusions).

Response to Comments:

We thank you for the constructive and detailed comments, which helped us further strengthen the empirical support and clarity of the manuscript. We respond to each point below.
(1) On the necessity of using two GeoTransformer models*: Following your suggestion, we have added an explicit ablation study in Table V (now Table VI) that evaluates the variant trained with a single GeoTransformer. In addition, the first row of Table VII reports results when only one GeoTransformer is trained and repeatedly applied with different numbers of inference iterations. These results consistently show that using a single model leads to degraded performance in later refinement stages, whereas training two GeoTransformers specialized for large initial pose discrepancies and small residual errors yields superior accuracy. Together, these experiments quantitatively justify our design choice of employing two GeoTransformers for coarse and fine alignment.* ***(2) On robustness to non-planar aligned objects****: We agree that robustness to non-planar object placements is critical for real-world robotic deployments. Accordingly, we have substantially expanded the experimental evaluation by introducing a large number of objects placed in inverted or non-planar configurations. We further increased the number of unseen objects and grasping trials to provide a more comprehensive assessment. Representative results of these newly added experiments are shown in the bottom two rows of Fig. 8 and Fig. 9, and additional grasping demonstrations have been uploaded to our anonymous project page. Moreover, we included an explicit analysis of real-world failure cases—covering non-planar alignment, large reference–query viewpoint discrepancies, and severe occlusions—as illustrated in the last two rows of Fig. 7. These additions demonstrate that the proposed method remains robust beyond planar object place scenarios, while also transparently characterizing its limitations.* ***(3) On the diversity of query views and viewpoint differences****: For evaluations on the five large public datasets, reference views are typically captured around canonical frontal viewpoints with moderate variations, while the full test sets include query views with substantially different viewpoints and occlusion patterns. All compared methods are evaluated under same test sets. In real-world robotic experiments, reference views are captured from the robot's default manipulation viewpoint (slightly elevated frontal views), while query objects are randomly placed in the workspace, naturally producing a wide range of viewpoint differences. Following your suggestion, we further added experiments that intentionally increase the viewpoint gap between reference and query views, with qualitative results shown in the bottom two rows of Fig. 8. These results reflect the robustness of the method to large scene differences, rather than sensitivity to minor visual perturbations.*

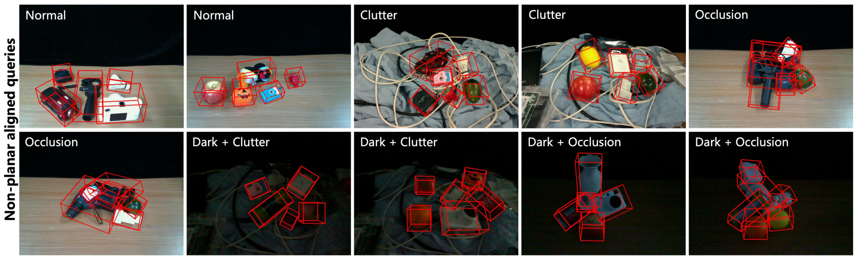
Changes Made in the Manuscript:

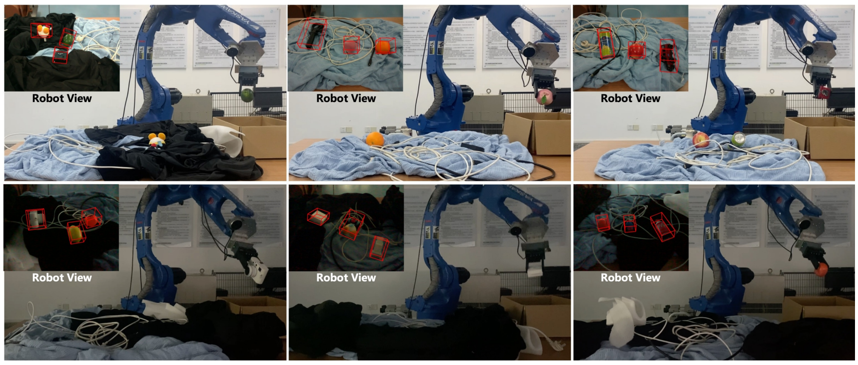
New added ablation experiment in Table VI:

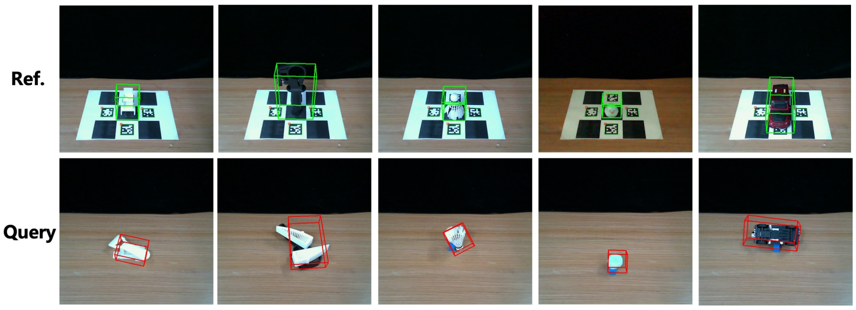
Row	Method	AR ↑	Param. (M) ↓
C	only one GeoTransformer	36.5	643.6
G	Full Model	62.2	691.8

Section V-F-1): In addition, we only train a single GeoTransformer for point cloud iterative alignment (row C of Tab. VI). Although the parameter count is slightly reduced, the performance degrades significantly. This is because a single alignment model struggles to simultaneously specialize in handling large initial pose discrepancies and accurately refining small residual errors.

New added non-planar visualizations in Fig. 8 (top), grasping experiments in Fig. 9 (middle), failure cases analysis in Fig. 7 (bottom):







Section V-E-1): Since the robustness to non-planar object placements is critical for real-world robotic deployments, we substantially introduce some objects placed in inverted or non-planar configurations (bottom two rows). In general, these experiments validate the effectiveness of SinRef-6D and demonstrate its potential for downstream robotic grasping tasks.

Caption of Fig. 8: Unseen object 6-DoF pose estimation in non-planar aligned query views. These include some challenging scenes commonly encountered in robotic grasping, including clutter, occlusion, low light, and dark conditions.

Section V-E-2): To further assess robustness under more challenging conditions, we additionally include grasping demonstrations involving geometrically irregular unseen objects placed in non-planar configurations and under large reference-query viewpoint discrepancies, as shown in the bottom two rows of Fig. 9.

Section V-D-3): Moreover, we provide an explicit analysis of real-world failure cases and observe a performance degradation under non-planar object placements, large reference–query viewpoint discrepancies, and severe occlusions, as illustrated in the last two rows of Fig. 7. Our future work will focus on enhancing the robustness of SinRef-6D in such challenging scenes and objects.

Caption of Fig. 7: The bottom two rows illustrate the labeled reference views and representative failure cases arising from particularly challenging scenarios, such as severe occlusion/self-occlusion and large viewpoint gaps between the reference and query views.

Reviewer 2

Summary Comments to the Author

This paper presents SinRef-6D, a novel framework for estimating the 6-DoF absolute pose of unseen objects for robotic manipulation. The key novelty is that it requires only a single RGB-D reference view without CAD models or a dense set of reference views. The paper is well-writtened, well-motivated, and technically strong, which makes a solid contribution to the field of 6D pose estimation.

Response to Comments:

We sincerely appreciate your positive and encouraging review. We have carefully considered all comments and revised the manuscript accordingly to further improve clarity and technical presentation.

The Reviewer's Comment 2.1

The proposed method is also technically sound and elegantly designed. The concept of "Points Focalization" is intuitive and the use of SSMs is an effective choice for capturing long-range spatial dependencies with linear complexity. While the paper is excellent, there are some areas where further work in a revision could significantly strengthen its claims and impact.

Response to Comments:

We thank your positive evaluation and constructive suggestion. We are glad that the intuition behind Points Focalization and the use of SSMs for efficient long-range spatial modeling are well recognized. We have carefully considered your suggestions for improvement and revised the manuscript accordingly to further strengthen the technical clarity and experimental validation.

The Reviewer's Comment 2.2

The real-world grapsing demonstration is a fantastic addition, but its evaluation could be made more rigorous to fully support the claims of robustness. The paper reports success rates from 50 trials on a limited set of graspable objects and a more extensive quantitative evaluation is expected. I would strongly recommend the authors expand this section by increasing the number and variety of unseen objects used for grasping and providing a more detailed description of the experimental protocol, particularly concerning the randomization of object poses and positions, as well as the complexity of the clutter. Further, there are some questions remained to be cleared: how are the "randomly placed unseen objects" distributed? Are challenging initial poses (e.g., large rotations relative to the reference view) explicitly tested? A more systematic study of the system's performance under varying conditions would be highly valuable.

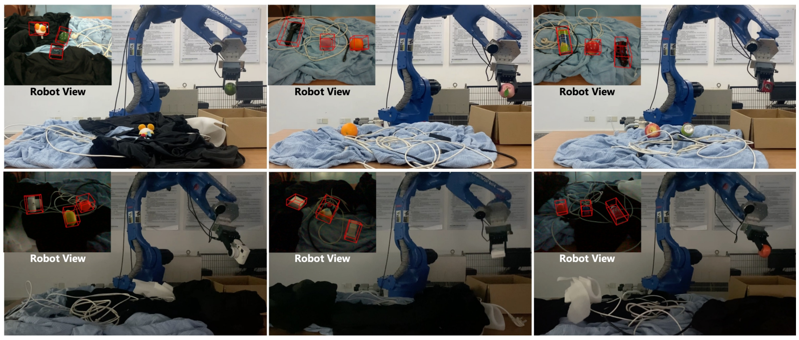
Response to Comments:

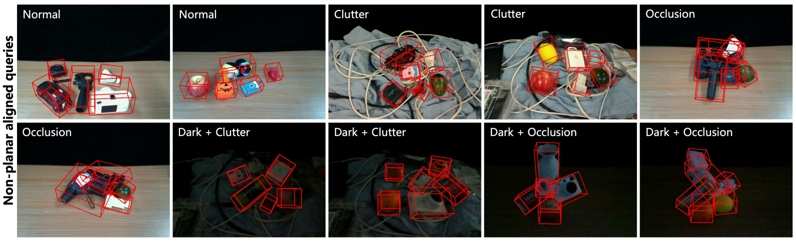
*We sincerely thank you for the positive assessment of the real-world grasping demonstrations and for the constructive suggestions on strengthening their rigor. Following this feedback, we have substantially expanded both the scale and the diversity of our real-world evaluation, and revised the manuscript accordingly. **First**, we increased the number and variety of unseen objects used in the grasping experiments. In addition to the originally reported objects, we introduced a broader set of household objects with more diverse geometries, aspect ratios, and shape complexities, better reflecting real-world manipulation scenarios. Correspondingly, the total number of grasping trials was increased to 200 executions, providing a more statistically meaningful evaluation. **Second**, we enhanced the difficulty and coverage of object pose configurations. Beyond planar placements, we explicitly evaluated challenging initial conditions where objects are tilted, rotated, or placed upside down, resulting in large pose discrepancies between the reference and query views. These settings directly reflect non-planar alignment and large viewpoint differences, which commonly arise in real-world robotic deployments. Qualitative results under such challenging configurations, including cluttered, dark, and heavily occluded scenes, have been added to the last two rows of Fig. 8. In these experiments, object poses are randomly placed on a grasping table, while scene clutter is generated by randomly moving non-target, non-rigid items such as blankets and cables. **Third**, to further improve transparency and diagnostic value, we added a dedicated real-world failure case analysis in the last two rows of Fig. 7, covering representative failure modes observed in real-world scenarios, including severe occlusion, large reference–query viewpoint gaps, and non-planar object placements. This analysis provides a more nuanced understanding of the limitations of the proposed method. **In addition**, several grasping demonstrations on newly added non-planar placed unseen objects have been uploaded to our anonymous project page, allowing readers to visually inspect the system behavior under diverse and challenging conditions. Also, we have added these visualizations to the last two rows of Fig. 9.*

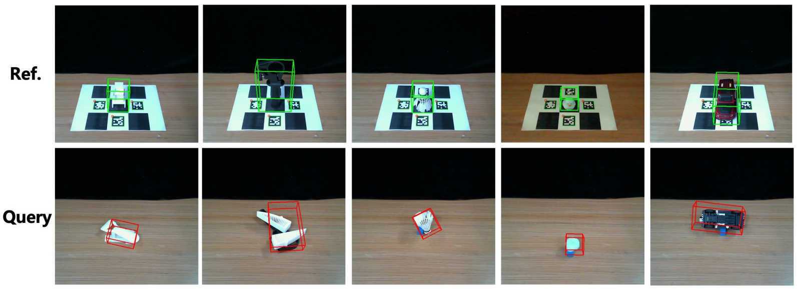
***Overall**, by increasing the number and diversity of unseen objects, expanding grasping trials, explicitly testing challenging initial poses, and providing qualitative analysis of real-world failures, we believe the revised evaluation offers a more systematic and comprehensive assessment of the robustness and practical applicability of our method.*

Changes Made in the Manuscript:

New added non-planar grasping experiments in Fig. 9 (top), visualizations in Fig. 8 (middle), failure cases analysis in Fig. 7 (bottom):







Section V-E-2): To further assess robustness under more challenging conditions, we additionally include grasping demonstrations involving geometrically irregular unseen objects placed in non-planar configurations and under large reference-query viewpoint discrepancies, as shown in the bottom two rows of Fig. 9. In these experiments, object poses are randomly placed, while scene clutter is generated by randomly moving non-target, non-rigid items such as blankets and cables. Quantitatively, we conduct a total of 200 real-world grasping trials, evenly divided between planar placements and challenging non-planar object configurations (100 trials each). Each scene contains two or three randomly placed unseen objects, achieving overall success rates of 85% and 74%, respectively, where a trial is considered successful only if all objects in the scene are successfully grasped.

Section V-E-1): Since the robustness to non-planar object placements is critical for real-world robotic deployments, we substantially introduce some objects placed in inverted or non-planar configurations (bottom two rows). In general, these experiments validate the effectiveness of SinRef-6D and demonstrate its potential for downstream robotic grasping tasks.

Caption of Fig. 8: Unseen object 6-DoF pose estimation in non-planar aligned query views. These include some challenging scenes commonly encountered in robotic grasping, including clutter, occlusion, low light, and dark conditions.

Section V-D-3): We provide an explicit analysis of real-world failure cases and observe a performance degradation under non-planar object placements, large reference–query viewpoint discrepancies, and severe occlusions, as illustrated in the last two rows of Fig. 7. Our future work will focus on enhancing the robustness of SinRef-6D in such challenging scenes and objects.

Caption of Fig. 7: The bottom two rows illustrate the labeled reference views and representative failure cases arising from particularly challenging scenarios, such as severe occlusion/self-occlusion and large viewpoint gaps between the reference and query views.

The Reviewer's Comment 2.3

The paper proposes using two separate GeoTransformer models with unshared weights: one for the initial alignment and another for all subsequent refinement iterations (Section III-F). The motivation is to have specialized models for the large initial pose discrepancy and the smaller subsequent ones. While this is empirically shown to be effective in the ablation study (Table VI), this design choice feels somewhat like an engineering solution rather than a generalizable principle. It raises the question of whether a single, more robust model could be designed to handle this dynamic range. I would encourage the authors to discuss this design decision more thoroughly. Acknowledging the trade-offs of the current approach and briefly discussing alternatives would strengthen the paper's methodological contribution.

Response to Comments:

We thank you for this thoughtful comment and fully agree that this design choice warrants a clearer methodological discussion. Our decision to use two GeoTransformer models with unshared weights is motivated by the distinct alignment regimes encountered during iterative point-wise alignment under a single-reference-view setting. Specifically, the first point-wise alignment stage typically faces a large pose discrepancy between the reference and query views, effectively corresponding to a coarse alignment problem. After this initial alignment, the pose discrepancy is significantly reduced, and subsequent iterations operate in a local refinement regime. These two regimes exhibit substantially different input distributions in terms of overlap ratio, correspondence noise, and geometric consistency. As demonstrated in Section V-F-1) (new added row C of Tab. VI) and Section V-F-2) (first row of Tab. VII), training a single GeoTransformer to handle both regimes leads to degraded performance during later iterations, where additional refinement unexpectedly reduces pose accuracy. We argue that this behavior arises from a distribution mismatch: the inputs encountered during refinement differ markedly from those seen during training, which is dominated by large-displacement alignment cases. Introducing a second GeoTransformer specialized for refinement alleviates this issue by explicitly modeling the local alignment regime. We acknowledge that alternative designs are possible. For example, a single model could be conditioned on iteration index or estimated pose discrepancy, or trained using curriculum or mixture-of-experts strategies to cover a broader dynamic range. While these directions are promising, our current design represents a practical and effective trade-off, balancing robustness, simplicity, and performance for generalized pose estimation–guided robotic manipulation. According to your comment, we revised the manuscript to explicitly discuss this design rationale, its trade-offs, and potential alternatives (future work).

Changes Made in the Manuscript:

Section III-F-First Paragraph: Therefore, we employ two GeoTransformer models with unshared weights to explicitly handle different alignment regimes. The first model is responsible for the initial point-wise alignment, where the pose discrepancy between the reference and query views is typically large and corresponds to a coarse alignment problem. After this stage, the pose discrepancy is significantly reduced, and a second GeoTransformer is used for subsequent iterative refinement under a local alignment regime. This separation allows each model to specialize in a distinct input distribution, improving stability and accuracy during refinement.

New added ablation in Table VI:

Row	Method	AR ↑	Param. (M) ↓
C	only one GeoTransformer	36.5	643.6
G	Full Model	62.2	691.8

Section V-F-1): In addition, we only train a single GeoTransformer for point cloud iterative alignment (row C of Tab. VI). Although the parameter count is slightly reduced, the performance degrades significantly. This is because a single alignment model struggles to simultaneously specialize in handling large initial pose discrepancies and accurately refining small residual errors.

Section V-F-2): As shown in the first row, using a single GeoTransformer for all alignment iterations leads to degraded performance in later refinement stages. We attribute this to a distribution mismatch between the training data, which is dominated by large pose discrepancies, and the inputs encountered during refinement, where the alignment error is much smaller. This observation motivates the use of a separate refinement model specialized for local alignment. Alternative designs, such as conditioning a single alignment model on the iteration index or estimated pose discrepancy, or adopting mixture-of-experts or curriculum learning strategies to handle different alignment regimes, could potentially address this dynamic range challenge and be explored in future work.

Reviewer 3

Summary Comments to the Author

This work introduces SinRef-6D, a 6-DoF pose estimation algorithm that utilizes a single reference view of a previously unseen object to estimate that object in a new pose. This enables effective pose estimation without requiring CAD models or dense views of the object first, which are laborious to acquire. SinRef-6D has two key components: first, an iterative point-wise alignment procedure, and second, a novel State Space Model (SSM) based feature extraction pipeline. This enables rich feature extraction in both point and RGB space which enables accurate point-to-point alignment and subsequent pose estimation. The authors propose a user-labeling method for estimating the absolute pose of the single reference view. The resulting pose estimates is then utilized downstream for robotic grasping. Extensive quantitative and qualitative comparisons are demonstrated over 6 pose estimation benchmarks showing favorable performance to baseline methods, even including methods requiring dense views or CAD models, an impressive result. Grasping achieves roughly 80-90% success across varying conditions.

Response to Comments:

We sincerely thank you for the thoughtful review and positive summary of our work. We have carefully considered all comments and revised the manuscript accordingly to further strengthen the technical presentation, clarify key design choices, and improve experimental analysis.

The Reviewer's Comment 3.1

The main novelty of this work lies in the feature extraction network. Iterative alignment is of course an established methodology, but this work presents a (to the reviewer's knowledge) novel and high performing feature backbone which seems to enable impressive performance. It is worth noting that alternative single-view reference methods are already existing, so the core problem statement is not novel in and of itself.

Response to Comments:

We thank you for the insightful and balanced assessment. We would like to clarify that our work is explicitly problem-driven, with the central question being: how to enable scalable 6-DoF absolute pose estimation of training-time unseen objects for robotic manipulation, using the minimum amount of prior information. Under this objective, we deliberately adopt a single reference view as the only object prior, which aligns with several recent concurrent efforts. However, our focus differs in that we target absolute object pose estimation that can be directly used for robotic manipulation, rather than relative pose estimation or purely perception-oriented evaluation. This distinction introduces several practical challenges (e.g. efficiently obtaining reference view with object absolute pose, handling significant pose discrepancies between reference and query views, and dealing with limited geometric and spatial information from sparse observations) that have not been jointly addressed in prior single-reference approaches. To this end, we design the entire pipeline around the requirements of generalized pose estimation–guided robotic manipulation. Specifically, we (1) develop a human–robot collaborative reference view acquisition and annotation tool tailored for single-reference absolute pose labeling in robotic manipulation scenarios, which is not considered in previous single-reference relative pose estimation methods; (2) propose a single reference view-based point cloud focalization strategy to focalize both reference and query observations into a common coordinate system, facilitating more stable alignment learning; (3) introduce the SSMs-based feature extraction network to effectively handle the limited geometric and spatial information contained in a single reference view; and (4) employ point-wise iterative alignment to robustly cope with potentially large pose discrepancies between the reference and query views. Each individual component is integrated under a unified, problem-oriented design enables a scalable and effective solution to unseen object 6-DoF absolute pose estimation for real-world robotic manipulation. We believe this task-driven, system-level design constitutes the primary contribution of our work.

Changes Made in the Manuscript:

Section III-A-Paragraph 1: Overall, our work is problem-driven, aiming to enable scalable 6-DoF absolute pose estimation of unseen objects for robotic manipulation with minimal prior information. To this end, we present a unified system that operates with only a single reference view, where each component is explicitly designed to address the challenges arising from single-reference, manipulation-oriented absolute pose estimation.

The Reviewer's Comment 3.2

A key weakness of the paper is a lack of comparison to the other single reference view methods (i.e., those outlined in Sec. II.C). In the related work discussion, the authors seem to suggest that relative pose estimation is insufficient for robotics, however, their method is subject to the same limitation, resolved only by including a user-driven labeling technique. This can easily be applied to the methods referred in Sec.II.C. Proper comparison to these works would highlight the importance of the contributed single-reference estimation method. Additionally, the reliance on a calibration board for rotation for the reference view seems quite challenging to overcome, though is likely a challenge for all methods requiring a reference pose.

Response to Comments:

We thank you for this constructive comment and fully agree that a direct and fair comparison with other single-reference methods is essential to clarify the strengths and limitations of the proposed approach. Following your suggestion, we have added explicit comparisons with the three most closely related single-reference methods discussed in Sec. II.C (One2Any, Any6D, and UNOPose). The quantitative results are reported in the newly added Table V. To ensure a fair comparison, all methods use the same object segmentation strategy and evaluation protocols: ground-truth segmentation on the LM dataset, and CNOS segmentation on the LM-O and TUD-L datasets; ADD-0.1d is used for LM, while AR is used for LM-O and TUD-L, consistent with prior work. The results of One2Any and Any6D are taken directly from their original papers, while UNOPose is evaluated using its pretrained model under CNOS segmentation. Existing single-reference methods primarily focus on estimating relative transformations between reference and query views, whereas our work explicitly targets 6-DoF absolute object poses that can be directly used for robotic manipulation. This requirement fundamentally guides our pipeline design, including reference view acquisition, point cloud focalization into a common coordinate system, and robust iterative alignment under large reference–query pose discrepancies. The added comparisons demonstrate that, under identical segmentation and evaluation protocols, SinRef-6D achieves more consistent and accurate absolute pose estimation. We acknowledge your concern regarding the use of a calibration board for reference view rotation annotation; in practice, this procedure is lightweight in robotic setups and constitutes a common challenge for all reference-based methods. Reducing this dependency remains an important direction for future work.

Changes Made in the Manuscript:

Table V. Comparison with other single-reference methods.

Method	LM	LM-O	TUD-L
	ADD-0.1d	AR	
One2Any [77]	52.6	-	-
Any6D [78]	-	28.6	-
UNOPose [76]	-	56.0	67.1
Ours	**90.3**	**56.5**	**77.4**

Section V-C-3): Table V shows explicit quantitative comparisons with the three most closely related single-reference methods discussed in Sec. II.C (One2Any [77], Any6D [78], and UNOPose [76]). To ensure a fair comparison, all methods use the same segmentation and evaluation protocols: ground-truth segmentation on the LM dataset, and CNOS segmentation [80] on the LM-O and TUD-L datasets; ADD-0.1d is used for LM, while AR is used for LM-O and TUD-L, consistent with prior work. The results of One2Any and Any6D are taken directly from their original papers, while UNOPose is evaluated using its pretrained model under CNOS segmentation. These results show that SinRef-6D achieves more accurate pose estimation performance.

The Reviewer's Comment 3.3

While the grasping experiments do show high performance, it is unclear to me that the experiment effectively utilizes the estimated object pose. First, the grasping strategy is a simplistic heuristic that seems to barely utilizes the object pose. Second, the examples shown in Fig.9 show only cylindrical objects being grasped which can likely be achieved by a heuristic and the segmented point clouds alone. A more fleshed out downstream task that requires precise object poses would more clearly highlight the pose estimation.

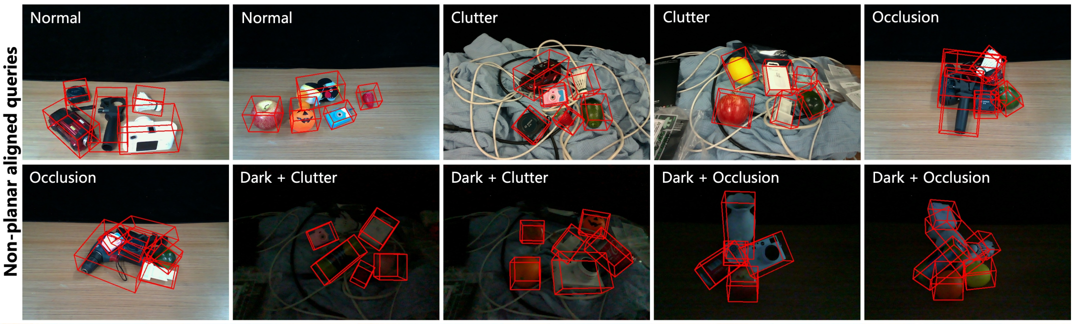
Response to Comments:

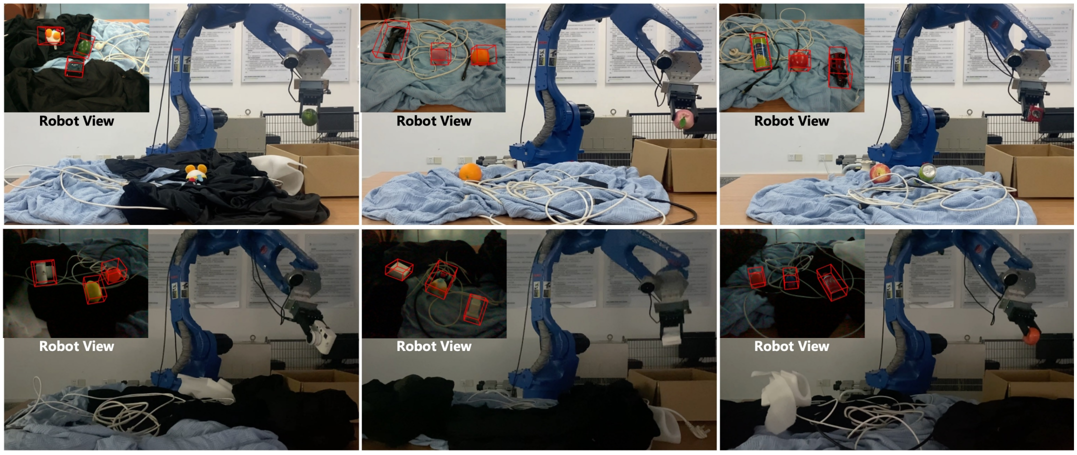
We thank you for the constructive feedback and for highlighting the importance of clearly demonstrating how the estimated 6-DoF object pose is utilized in downstream robotic grasping. First, we clarify that the grasping policy explicitly relies on the estimated 6-DoF object pose. The estimated 3-DoF translation determines the target grasp point, i.e., the position to which the end-effector center is moved, while the 3-DoF rotation determines the grasping orientation. Specifically, for each object, the grasping direction is defined by a vector from its closest visible surface point projected onto the z-axis toward the object center in the estimated object coordinate system, and the gripper closes along the x-axis. This design directly couples grasp execution to both the estimated object position and orientation, rather than relying solely on segmented point clouds. Second, we agree with you that grasping cylindrical objects—especially when they are planar-aligned—can be relatively straightforward and may not fully reflect the necessity of accurate 6-DoF pose estimation. To address this concern, we have significantly expanded the real-world evaluation by introducing multiple non-cylindrical, geometrically irregular unseen objects, which are deliberately placed in tilted, inverted, and non-planar configurations. Representative qualitative results are shown in the last two rows of Fig. 8 and Fig. 9, and additional grasping demonstrations have been uploaded to our anonymous project page. These scenarios are considerably more challenging and clearly demonstrate the necessity of precise 6-DoF pose estimation for successful grasp execution. In future work, we plan to further explore downstream applications that demand even higher pose accuracy, such as augmented reality and fine-grained hand-object interaction.

Changes Made in the Manuscript:

Section IV-B-2): For each object, the estimated 3-DoF translation determines the target grasp point, i.e., the position to which the end-effector center is moved, the grasping direction is defined by a vector from its closest visible point projected onto the z-axis to the center point of the object coordinate system, while the gripper closes along the x-axis.

New added visualizations of non-planar aligned unseen objects pose estimation in Fig. 8 (top) and grasping experiments in Fig. 9 (bottom):





Section V-E-1): Since the robustness to non-planar object placements is critical for real-world robotic deployments, we substantially introduce some objects placed in inverted or non-planar configurations (bottom two rows). In general, these experiments validate the effectiveness of SinRef-6D and demonstrate its potential for downstream robotic grasping tasks.

Caption of Fig. 8: Unseen object 6-DoF pose estimation in non-planar aligned query views. These include some challenging scenes commonly encountered in robotic grasping, including clutter, occlusion, low light, and dark conditions.

Section V-E-2): To further assess robustness under more challenging conditions, we additionally include grasping demonstrations involving geometrically irregular unseen objects placed in non-planar configurations and under large reference-query viewpoint discrepancies, as shown in the bottom two rows of Fig. 9.

The Reviewer's Comment 3.4

The paper is clearly written, is well organized, and with excellent related work. The methodology is largely clear. The only exception is perhaps the details of the SSM models. The experiments are thorough and the ablations help detail the important components of the method.

Response to Comments:

We sincerely thank you for the positive evaluation of the clarity, organization, and experimental thoroughness of our paper. We also appreciate the comment regarding the presentation of the SSM models. In response, we have carefully revised the manuscript to further clarify the design, role, and operation of the SSM components used in our framework. Specifically, we expanded the explanation of how SSMs are instantiated for point-wise and image-based feature extraction, clarified the notion of "sequence" in this context (i.e., spatial tokens rather than temporal signals), and improved the consistency of notation used to describe intermediate feature representations. In addition, we provided a more detailed, figure-grounded explanation of the network architecture by closely integrating the textual description with Fig. 3 and Fig. 4, enabling a clearer and more intuitive understanding of the overall design. These revisions aim to make the SSM-based design more accessible to readers who may be less familiar with state space models, while preserving the technical rigor and conciseness expected by the community.

Changes Made in the Manuscript:

Section III-D: The sequence modeled by the SSMs refers to an ordered collection of spatial tokens rather than physical time steps. For the RGB branch, the input image is partitioned into patches and flattened into sequences following two fixed raster-scan orders (as shown in part (C) of Fig. 2), which are consistently used during training and inference. This ordering enables the SSM to capture long-range spatial dependencies across image regions without encoding temporal information. For the point cloud branch, due to the intrinsic unordered nature of point sets, the input sequence is constructed by iterating over all points, and the resulting order is neither fixed nor semantically meaningful. For the Point SSM, as shown in Fig. 3, we first perform a point-wise scan and use K-Nearest Neighbor (KNN) to sample a set of points for each scanned point to form a token. Then, we compute all token embeddings and add a position embedding to them. Subsequently, the token embeddings are concatenated and passed into the points state space (PSS) blocks to obtain the point-wise feature $F^P \in {}^{2048 \times 256}$. The details of the selective SSM in PSS blocks can be found in the S6 model [81]. For RGB image feature extraction, we propose an RGB SSM based on the cross-scan manner and multi-scale feature fusion, as depicted in Fig. 4. The architecture consists of four stages, where each stage employs visual state space (VSS) blocks [83] to extract image features at different scales. These multi-scale features are then fused, reshaped, and chosen by using the image mask to obtain the final image feature representation $F^I \in {}^{2048 \times 256}$.

Specifically, F^P and F^I denote the generic point-wise and image feature representations, which are instantiated as reference and query features. The complete process is illustrated in part (C) of Fig. 2, where F_r^P and $F_q^{P^i}$ represent the extracted point-wise reference and query features (at the *i-th* iteration), while F_r^I and F_q^I denote the extracted features from the reference and query RGB images.

The Reviewer's Comment 3.5 (other notes/questions)

3.5.1: Please provide more detail on how the "sequence" modeling of the SSMs are applied here. What is the "time" dimension? Is the sequence determined simply by the order of the tokenized point/image patches? Does the order have to remain fixed?
3.5.2: In Sec. III.B. and Sec. V.F., the "50th to 120th in its rendering sequence" is referenced. This seems arbitrary and over-specified. Why can we not treat the single reference view as either random or manually selected for clarity?
3.5.3: In Sec. III.D., the F^P and F^I terms are missing notation in some places.
3.5.4: In Eq.8, why is i only 1 or 2? In the ablations the number of iterations is set as high as 4. This should likely be change to a hyperparameter term.
3.5.5: In Sec V.C.1, is it fair to remove the language prompt from the Oryon method? If it is trained with language and simply removed at test time, will that not cause the query to be necessarily out of distribution and thus cause the network to fail? Are the authors re-training to account for distribution shift of any networks being supplied with only one image who may have been trained on more?
3.5.6: In Sec V.F., how does the random reference view ablation differ from the variations already tested in Tables 1 and 4?

Response to Comments 3.5.1:

We thank you for these questions. In our implementation, the "sequence" of the SSMs corresponds to the tokenized point and image patches, not to physical time steps. For the RGB SSM, we tokenize the image into patches and flatten them in raster-scan order. The order is kept fixed during training and inference for consistency, but the SSM itself does not encode any physical time; it acts as a sequence model over spatial tokens. For the Point SSM, since point clouds are inherently unordered, the input sequence is formed by iterating over all points, and the resulting order does not carry any physical or semantic meaning. We have added this explanation to Section III-D to clarify the notion of "sequence" in our SSMs.

Changes Made in the Manuscript:

Section III-D-Paragraph 2: The sequence modeled by the SSMs refers to an ordered collection of spatial tokens rather than physical time steps. For the RGB branch, the input image is partitioned into patches and flattened into sequences following two fixed raster-scan orders (as shown in part (C) of Fig. 2), which are consistently used during training and inference. This ordering enables the SSM to capture long-range spatial dependencies across image regions without encoding temporal information. For the point cloud branch, due to the intrinsic unordered nature of point sets, the input sequence is constructed by iterating over all points, and the resulting order is neither fixed nor semantically meaningful.

Response to Comments 3.5.2:

We appreciate your comment. During training on the synthetic dataset, the synthetic reference view is sampled from a range that approximates the robotic manipulation viewpoint while introducing natural perturbations. Specifically, we follow the same rendering protocol as GigaPose and randomly sample one viewpoint from the 50th to the 120th in its rendering sequence, which simulates manual reference view acquisition with natural pose variations. During evaluation in real-world robotic scenarios, the reference view for each unseen object is captured by *the robot from an occlusion-free manipulation viewpoint and annotated using our custom-developed semi-automatic annotator. Randomly sampling arbitrary rendering viewpoints may introduce extreme perspectives, such as near top-down or overhead views, which differ significantly from real-world reference acquisition. Although manually selecting the viewpoints of all objects could alleviate this issue, it is labor-intensive and may reduce the robustness of the model to viewpoint variations.*

Changes Made in the Manuscript:

Section III-B-Paragraph 1: Notably, randomly sampling arbitrary rendering viewpoints may introduce extreme perspectives (e.g., near top-down views) that deviate significantly from real-world reference acquisition, whereas manually selecting viewpoints for all objects would be labor-intensive and may reduce robustness to viewpoint variations. From a practical real-world application perspective, during training, the synthetic reference view is sampled from a viewpoint range that approximates the robotic manipulation viewpoint while introducing natural perturbations.

Response to Comments 3.5.3:

You are correct. In the original version, F^P and F^I were used as generic feature representations, while their instantiations as reference and query features (F^P_r, F^P_q, F^I_r, F^I_q) were introduced later without being explicitly linked. We have added a clarifying sentence in Sec. III-D to make this relationship explicit and ensure notational consistency.

Changes Made in the Manuscript:

Section III-D-Last Paragraph: Specifically, F^P and F^I denote the generic point-wise and image feature representations, which are instantiated as reference and query features. The complete process is illustrated in part (C) of Fig. 2, where F^P_r and $F^{P^i}_q$ represent the extracted point-wise reference and query features (at the *i-th* iteration), while F^I_r and F^I_q denote the extracted features from the reference and query RGB images.

Response to Comments 3.5.4:

Thank you for the insightful comment. We clarify that the index i in Eq. (8) denotes the number of non-weight-sharing GeoTransformers employed during training, i.e., the training iterations, which is consistent with the first column of Table VII. In contrast, the first row of Table VII corresponds to the inference iterations, where 1 $\sim$ 4 rounds of point cloud iterative alignment are performed at test time to study the trade-off between accuracy and efficiency. Following your suggestion, we have revised Eq. (8) by replacing i with an explicit hyperparameter to avoid potential ambiguity.

Changes Made in the Manuscript:

Section III-F-Last Paragraph:

$$Loss = \sum_{i=1,\ldots,K} \Big(CE(A^i, \bar{p}_q) + CE(A^{i^\top}, \bar{p}_r) \Big), \tag{1}$$

where K denotes the number of point-wise alignment iterations used during training.

Ablation Study: Section V-F-2): The training iterations refer to the number of times the GeoTransformer weights are updated during training (*i.e.*, K in Eq. (8))

Response to Comments 3.5.5:

We apologize for the confusion caused by an inaccurate description in the previous version. In fact, the reported results of Oryon are directly taken from Table 3 of the One2Any paper, where Oryon is evaluated under its original experimental setting, including language input, and using its official pretrained model. We did not remove the language prompt at test time, nor did we retrain or modify the Oryon model in any way. To further ensure correctness, we have confirmed this evaluation protocol through direct email communication with the authors of One2Any and Any6D. Therefore, the reported Oryon results do not suffer from distribution shift caused by missing language input and are fair comparisons under the same settings as reported in the original literature. We have corrected the corresponding description in the manuscript to avoid further misunderstanding.

Changes Made in the Manuscript:

Section V-C-First Paragraph: For a fair comparison with Oryon [47], we adopt the results reported in One2Any [77], which directly evaluate Oryon using its pretrained model under the original setting including language input.

Response to Comments 3.5.6:

In Tables 1 and 4, we adopt two reference view acquisition strategies. The first follows the same protocol as used during training (randomly sampled and rendered from the viewpoints 50th to 120th defined in GigaPose), while the second selects, for each object, a reference view that is manually chosen from the corresponding dataset to be closest to the robotic manipulation viewpoint. In Sec. V-F-3), our primary goal is to evaluate the robustness of the proposed model to random reference view selection. Therefore, we employ the same reference view sampling strategy as in training and conduct 20 independent random trials, reporting the mean and variance of the results.

Changes Made in the Manuscript:

Ablation Study: Section V-F-3): Unlike the reference view variations evaluated in Tabs. I and IV, which compare different reference acquisition strategies, the objective here is to assess the robustness of the proposed model under stochastic reference view selection. Specifically, we conduct 20 tests on both the YCB-V and LM-O datasets, where in each test the reference view is randomly sampled and rendered from the viewpoints *50th* to *120th* defined in GigaPose, following the same reference view sampling protocol used during training.