Click-to-Model: Real-Time Interactive Object Modeling and Robust 6D Pose Tracking

Abstract— **Accurate perception of target objects is fundamental to robotic manipulation, yet current perception methods rely on known object models and perform poorly in tracking unseen objects. Simultaneously, computational costs during real-time robotic arm grasping hinder the pursuit of real-time performance. To address this challenge, we propose CLM, an interactive tracking framework based on learn-match. For object modeling, it introduces a shape-color-driven particle size matching method that effectively captures surface feature points. Combined with depth maps, this approach reconstructs models while ensuring consistency with the original object's dimensions. Furthermore, we incorporate a learning-matching-based tracking mechanism that balances accuracy with real-time performance. Experiments conducted on YCB-V, and our self-built dataset demonstrate that our method not only matches the ADD(-S) metric of current SOTA approaches, but more importantly, achieves stable tracking of unseen objects and resolves re-localization issues when targets exit the field of view. Ablation studies further confirm that the scale-matching strategy based on joint shape-color particles is crucial for scale recovery.**

I. INTRODUCTION

Robotic systems deployed in dynamic and unstructured environments must frequently interact with previously unseen objects. Reliable 6D pose estimation and tracking is therefore essential for perception, manipulation, and task execution, as shown in Fig. 1(a). However, as illustrated in Fig. 1(b), most existing pose tracking methods assume access to pre-defined CAD models or require offline object scanning prior to deployment. These assumptions limit scalability and significantly slow down real-world operation, particularly in human-in-the-loop scenarios where rapid object onboarding is required.

Meta AI's SAM-3D [1] introduces a novel paradigm for interactive object modeling. This method leverages appearance and texture information from RGB images to generate three-dimensional approximations of target objects through user interaction, enabling geometric reconstruction of unseen objects without requiring pre-existing CAD models. This modeling approach demonstrates strong flexibility in open scenes, laying the groundwork for subsequent pose estimation and tracking. However, as the method primarily relies on texture and appearance cues from two-dimensional images, it lacks constraints on real depth and physical scale. Consequently, the generated 3D models typically reside in a normalized scale space, exhibiting significant dimensional discrepancies compared to real objects. This scale mismatch directly impacts the accuracy of model-based 6D pose estimation [2], imposing clear limitations in scale-sensitive applications like robotic grasping [3–5].

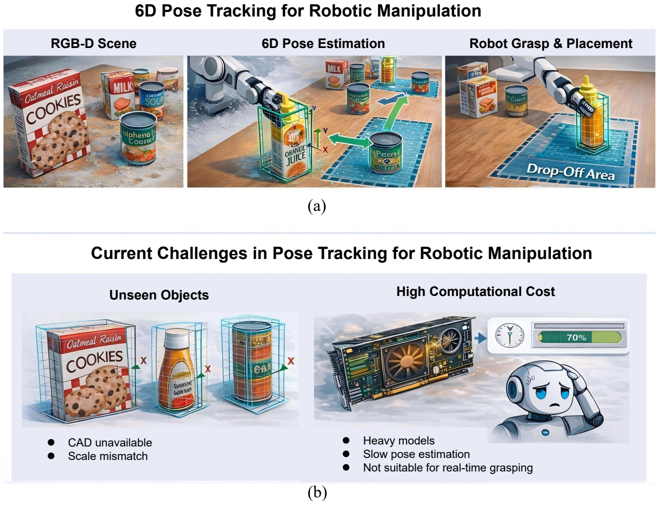

Fig. 1 (a) RGB-D based 6D pose estimation enabling robotic grasping and placement. (b) Fundamental obstacles for accurate and real-time pose tracking in robotic tasks.

Conversely, NVIDIA's FoundationPose [6] achieves high-precision 6D pose estimation and continuous tracking for novel objects through a learning-driven feature matching and pose regression mechanism. It demonstrates superior robustness on public datasets and in complex occlusion scenarios. However, this approach relies on relatively precise object models and initial mask inputs. Additionally, its feature matching and optimization processes incur substantial computational overhead, leading to high inference latency. In real-time robotic arm grasping or human-robot interaction scenarios, this computational burden struggles to meet the demands for low latency and high responsiveness.

Based on this analysis, existing approaches have made progress in "model acquisition" and "objects tracking" separately, but no unified framework has yet been established that balances scale consistency, real-time performance, and interactivity. To address this, we propose an interactive object pose estimation and tracking framework based on a learn-match mechanism. This framework minimizes user interaction costs by requiring clicks to identify the target object and establishes a tightly coupled relationship between online modeling and pose estimation, as illustrated in Fig. 2. By integrating learning-driven feature representation with efficient matching strategies, it significantly reduces inference latency while maintaining pose estimation accuracy, enabling real-time 6D pose output for unseen objects. This approach enhances scale consistency and system responsiveness while

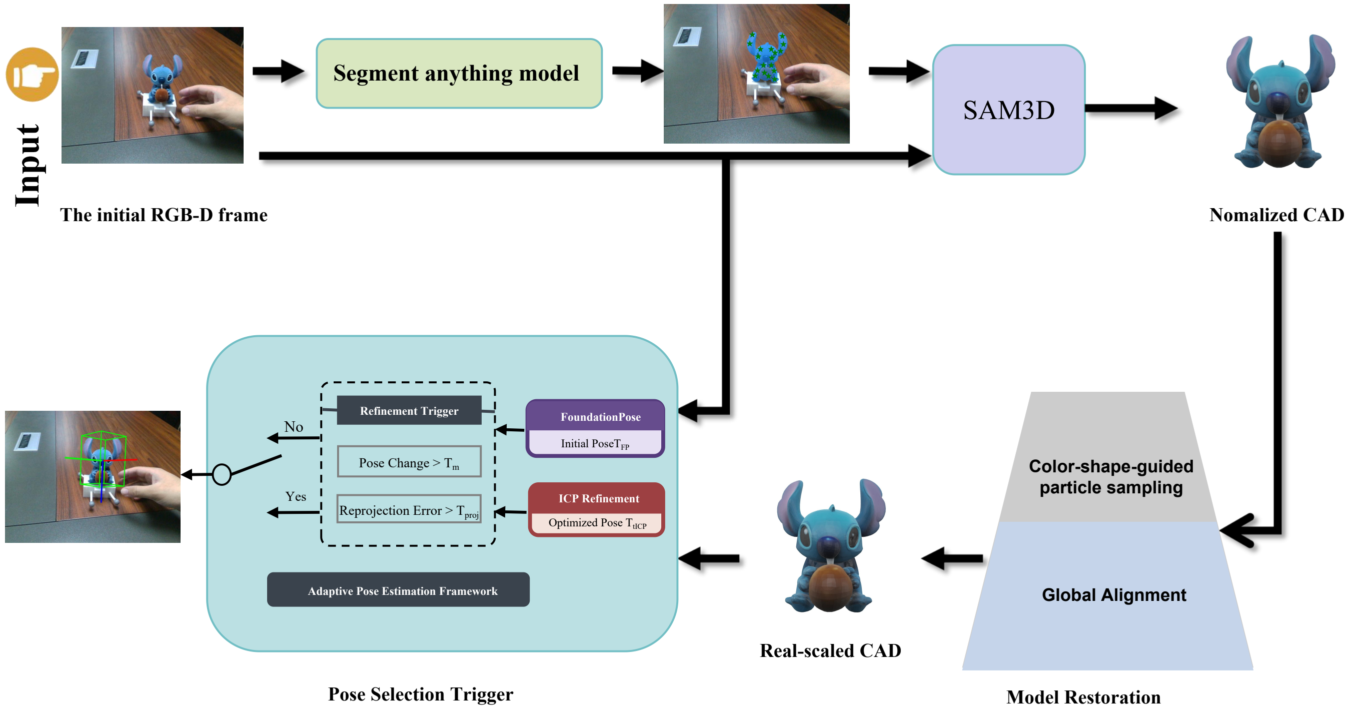

Fig. 2 The framework comprises three stages: preliminary segmentation and normalization, scale-consistent model restoration, and adaptive pose estimation. A feature-aware fitness function integrating geometric, color, and robustness terms guides particle-based global alignment to recover a metrically consistent object model. FoundationPose then provides high-accuracy pose estimation from RGB-D observations, serving as the primary localization module. ICP operates as a lightweight geometric tracker, refining and stabilizing the pose under motion through fast local alignment when triggered. This coarse-to-fine and learning-geometry hybrid design ensures metric correctness, robustness to occlusion, and real-time performance.

preserving model adaptability, offering a more feasible solution for robotic operations in open environments.

A key component of the proposed framework is a scale-aware registration strategy that improves model alignment during online reconstruction. We introduce a color-shape-guided probabilistic particle sampling mechanism to select representative points for scale alignment. Instead of relying solely on geometric consistency, the method jointly considers color distribution and structural cues to guide particle selection, enabling robust scale calibration even under partial observations. This strategy enhances reconstruction stability and reduces initialization drift.

For pose estimation and tracking, we design a hybrid decision-based localization scheme that combines coarse pose inference with geometric refinement. The system first performs pose initialization and temporal tracking using a learning-based estimator, and then selectively applies iterative geometric alignment to refine the pose when necessary. This decision-driven integration balances computational efficiency and localization accuracy. By adaptively invoking geometric refinement, the framework maintains real-time performance while preserving robustness in challenging scenarios.

The main contributions of this work are as follows:

1) **Complete interactive perception pipeline:** an end-to-end click-to-model framework unifying segmentation, 3D modeling, and pose tracking for real-time deployment.
2) **Scale-aware probabilistic registration:** a color-shape guided particle sampling method for robust online scale alignment.
3) **Decision-driven hybrid localization:** a combined learning-based and geometric alignment approach balancing real-time performance and accuracy.

II. RELATED WORK

A. Object Modeling from Segmentation and RGB-D Reconstruction

Recent advances in foundation models have significantly improved category-agnostic object segmentation, with the Segment Anything Model (SAM) [7] demonstrating strong generalization and prompt-based mask extraction capabilities. Based on segmentation output, SAM3D-style pipelines [1] perform object-centric reconstruction by isolating masked regions and fusing multi-view RGB or RGB-D observations to generate 3D models. SAM-6D [8] further extends this paradigm to zero-shot 6D pose estimation. While such approaches alleviate dependence on pre-scanned CAD models, RGB-driven reconstruction typically produces normalized object geometry without reliable metric scale, limiting its direct applicability to precise 6D pose estimation in robotic manipulation. Depth-aware fusion improves geometric fidelity but does not fully resolve scale inconsistency under partial observation. In contrast to prior work that focuses primarily on visual reconstruction quality, our approach explicitly addresses metric scale recovery during online reconstruction through a color-shape-guided probabilistic point selection

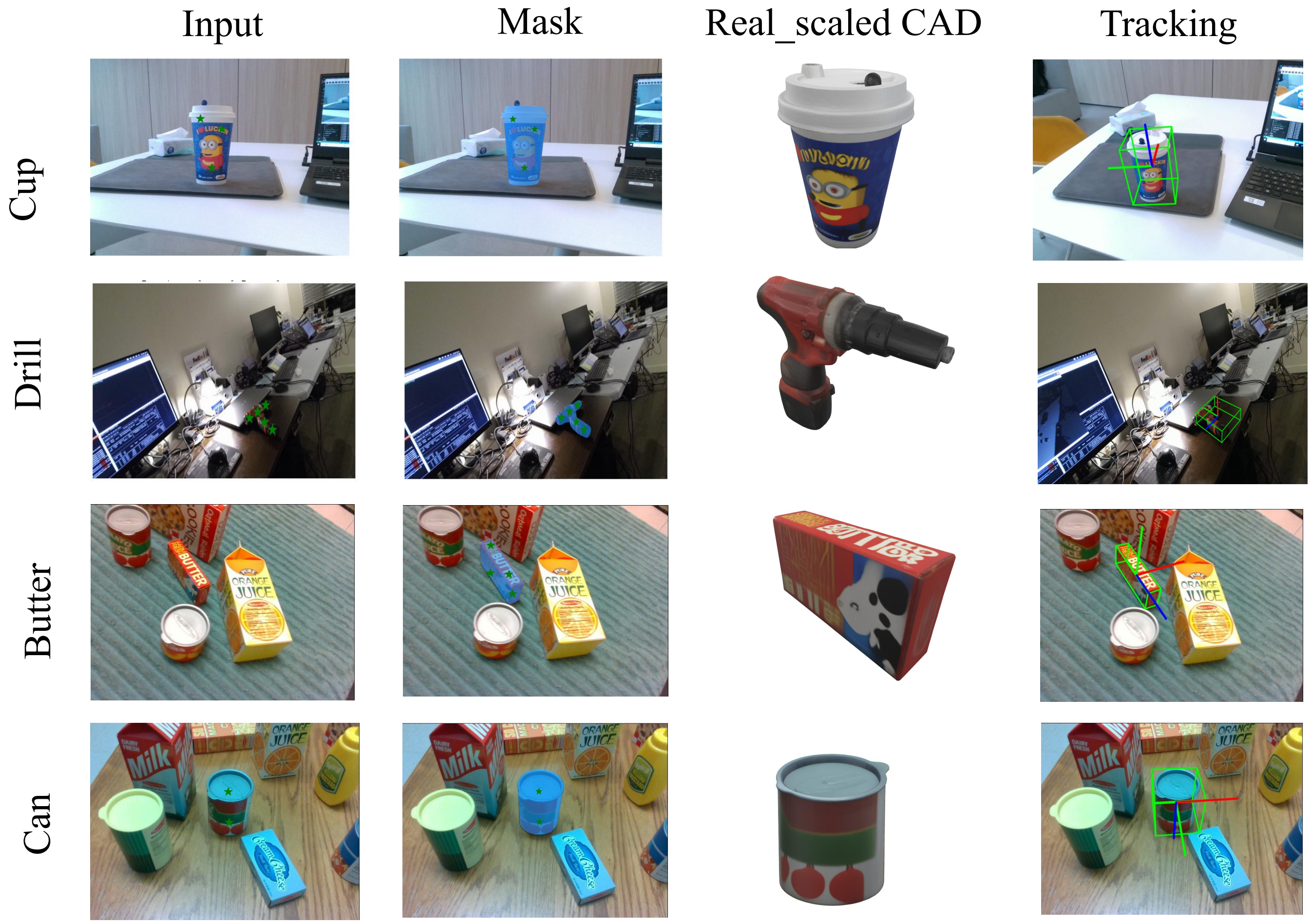

Fig. 3 Qualitative results on multiple datasets, including our self-collected dataset, the HOPE dataset, and the NVIDIA dataset. The figure illustrates the pipeline from RGB input and object mask extraction to scale-aligned CAD reconstruction and subsequent 6D pose tracking across different object categories.

strategy, enabling scale-consistent object modeling suitable for downstream pose tracking.

B. 6D Pose Estimation and Tracking

6D pose estimation has been extensively studied using both model-based geometric alignment and learning-based render-and-compare frameworks [9, 10]. Classical approaches rely on feature correspondence and iterative closest point (ICP) [11, 12] refinement given accurate CAD models [2, 13], achieving strong geometric consistency but requiring reliable initialization and precise object geometry. Instance-level methods such as DenseFusion [14], PVN3D [15], and PVNet [16] leverage dense correspondences for pose regression. Recent learning-based methods, such as FoundationPose [6], significantly improve robustness and tracking stability by leveraging learned feature representations and synthetic-to-real generalization. However, these approaches generally assume access to accurate CAD models and pre-defined masks, limiting their use in scenarios involving previously unseen objects. Methods such as DPOD [17], FFB6D [18], SurfEmb [19], and GNC-Pose [20] improve dense correspondence learning. Hybrid pipelines combining learned initialization with geometric refinement have been explored [21], yet reconstruction and tracking are typically treated as independent stages. Our method integrates online model generation with a decision-based hybrid localization scheme that adaptively combines FoundationPose and ICP refinement, achieving a balance between computational efficiency and pose accuracy under real-time robotic constraints.

C. Model-Free and Unseen Object Pose Estimation

Traditional 6D pose estimation methods typically rely on pre-scanned CAD models and category-specific training data [22, 23], limiting their applicability to unseen objects [24]. To address this limitation, recent works have explored model-free and generalizable pose estimation paradigms. OnePose [25] and OnePose++ [26] employ template-based matching without CAD models. Gen6D [27] and BundleTrack [28] enable model-free tracking of novel objects. Category-level methods such as NOCS [29], FS-Net [30], and GPV-Pose [31] leverage learned shape priors. MegaPose [32] and LatentFusion [33] adopt render-and-compare strategies. Some approaches leverage large-scale synthetic training and feature matching to align RGB observations with pre-learned shape priors, while others employ template retrieval or image-to-image matching strategies to infer pose without explicit CAD supervision. Another line of research focuses on prompt-based or segmentation-driven modeling, where object masks or user interaction serve as the initialization cue for pose estimation. These methods reduce dependency on pre-built object models and improve flexibility in interactive settings. However, they often face challenges in scale recovery, robustness under large viewpoint changes, or maintaining stable tracking over long sequences. Despite notable progress, including recent advances in Gaussian splatting for pose estimation [34, 35] and robust ICP variants, achieving accurate, real-time, and robust pose tracking for previously unseen objects in robotic manipulation scenarios remains an open problem, particularly when interactive specification and scale-consistent reconstruction are required.

III. METHODOLOGY

A. Overview

The Click-to-Model (CLM) framework addresses two central challenges in interactive robotic perception of unknown objects: precise metric scale recovery and stable real-time 6D pose tracking. Unlike conventional pipelines that depend on pre-scanned CAD models with known dimensions, CLM directly reconstructs scale-consistent object representations from RGB-D observations and maintains reliable pose estimation through an adaptive hybrid tracking strategy. To resolve scale ambiguity introduced by RGB-driven reconstruction, CLM leverages aligned depth information and formulates scale recovery as a probabilistic particle-based optimization problem. Randomly sampled surface particles are evaluated using a unified objective that integrates photometric consistency and geometric saliency, enabling metric alignment between reconstructed geometry and physical depth measurements. For pose tracking, CLM combines FoundationPose [6] for robust global inference under moderate motion with ICP [11, 12] for efficient local alignment during rapid changes. A motion-aware mechanism monitors inter-frame pose increments to adaptively switch between estimators, ensuring both stability and responsiveness. Recent works on robust registration and deep point cloud alignment [36] offer complementary strategies. This unified framework achieves a balanced trade-off between accuracy and real-time performance in unseen object manipulation scenarios.

B. Color–Shape Consistent Particle-Based Scale Recovery

In interactive object modeling based on SAM3D, the reconstructed object model is typically produced at a normalized scale and lacks metric consistency with the physical world. This scale ambiguity often leads to multiple feasible solutions or degenerate configurations due to symmetry, repetitive structures, and partial visibility. To address this issue, we introduce a global isotropic scale parameter $s > 0$ and formulate scale recovery as a depth-anchored optimization guided by color–shape consistent particle selection.

Let the SAM3D-generated object model be represented as a colored point cloud

$$\mathcal{S} = \{(\mathbf{s}_i, \mathbf{c}_i)\}_{i=1}^{N_s}, \tag{1}$$

and the observed RGB-D scene be

$$\mathcal{T} = \{(\mathbf{t}_j, \mathbf{d}_j)\}_{j=1}^{N_t}. \tag{2}$$

We denote the rigid transformation as $\mathbf{T} = (\mathbf{R}, \mathbf{t}) \in SE(3)$. Under scale and pose transformation, each model point becomes

$$\hat{\mathbf{s}}_i = \mathbf{R}(s\mathbf{s}_i) + \mathbf{t}. \tag{3}$$

a) Particle-Based Anchor Selection.: To avoid unstable correspondences in geometrically repetitive or textureless regions [13, 37], we randomly sample n particles on the model surface as candidate anchors. For each particle $\mathbf{p}_k$, local geometric and color descriptors, denoted as $\phi_g(\mathbf{p}_k)$ and $\phi_c(\mathbf{p}_k)$ respectively, are computed within its neighborhood.

We estimate the density of each particle in geometric and color feature spaces using kernel functions:

$$\rho_g(\mathbf{p}_k) = \sum_{l \neq k} \exp\Big(- \frac{\|\phi_g(\mathbf{p}_k) - \phi_g(\mathbf{p}_l)\|^2}{2\sigma_g^2} \Big), \tag{4}$$

$$\rho_c(\mathbf{p}_k) = \sum_{l \neq k} \exp\Big(- \frac{\|\phi_c(\mathbf{p}_k) - \phi_c(\mathbf{p}_l)\|^2}{2\sigma_c^2} \Big). \tag{5}$$

Particles located in repetitive regions exhibit high density and are therefore less reliable. We define a joint distinctiveness score

$$U(\mathbf{p}_k) = \alpha\big(- \log \rho_g(\mathbf{p}_k)\big) + (1 - \alpha)\big(- \log \rho_c(\mathbf{p}_k)\big), \tag{6}$$

where $\alpha \in [0, 1]$ balances geometric and color contributions. The score is normalized into a probabilistic weight

$$\omega_k = \frac{\exp(\beta U(\mathbf{p}_k))}{\sum_l \exp(\beta U(\mathbf{p}_l))}, \tag{7}$$

with temperature parameter $\beta > 0$. Through iterative filtering, we obtain a stable anchor set $\mathcal{A} = \{\mathbf{a}_m\}_{m=1}^{M}$ composed of highly distinctive particles.

b) Depth-Anchored Scale Estimation.: Given anchor correspondences between model anchors $\mathbf{a}_m$ and scene points $\mathbf{b}_m \in \mathcal{T}$, scale recovery is formulated as a weighted least-squares problem:

$$\min_s \sum_{m=1}^{M} \omega_m \left\|\mathbf{R}(s\mathbf{a}_m) + \mathbf{t} - \mathbf{b}_m\right\|_2^2 . \tag{8}$$

With $\mathbf{R}$ and $\mathbf{t}$ fixed, this becomes a one-dimensional quadratic optimization in s. Taking the derivative with respect to s and setting it to zero yields the closed-form solution:

$$s^* = \frac{\sum_{m=1}^{M} \omega_m \mathbf{a}_m^\top \mathbf{R}^\top (\mathbf{b}_m - \mathbf{t})}{\sum_{m=1}^{M} \omega_m \|\mathbf{a}_m\|_2^2}. \tag{9}$$

This expression directly ties the scale parameter to metric depth observations in the RGB-D scene. Since the weights ω_m encode color–shape distinctiveness, unreliable or symmetric regions exert limited influence on the solution.

Overall, the proposed method decomposes scale recovery into two coupled stages: distinctive anchor selection and depth-consistent metric estimation. By integrating color and geometric cues during particle evaluation and leveraging RGB-D measurements for metric anchoring, the approach significantly reduces ambiguity under symmetry and occlusion, providing a stable and physically consistent scale foundation for subsequent 6D pose estimation and tracking.

C. Hybrid FoundationPose–ICP Object Tracking

After obtaining a scale-consistent object model $s^*\mathcal{S}$, object pose estimation and temporal tracking are formulated as a joint learning–geometry optimization problem. Let the RGB-D observation at time t be denoted as $I_t = \{I^t_{rgb}, I^t_{depth}\}$. The objective is to estimate the object pose $\mathbf{T}_t \in SE(3)$ in the camera coordinate frame at each frame.

a) Primary Global Localization via FoundationPose.: FoundationPose serves as the primary global localization module. Given the current observation and the metric-consistent model, FP predicts the object pose as

$$\mathbf{T}_t^{FP} = f_\theta(I_t, s^*\mathcal{S}), \tag{10}$$

where $f_\theta(\cdot)$ denotes the trained network. By jointly leveraging RGB semantic representations and depth geometry, FP performs direct pose regression in $SE(3)$ and maintains robustness under occlusion and background clutter.

From an operational perspective, FP is responsible for (i) initial pose acquisition, (ii) recovery from large inter-frame motion, and (iii) re-localization after tracking failure. Since FP relies on statistical learning, strict geometric consistency with the observed depth map is not explicitly enforced, and minor pose drift may accumulate over long sequences.

b) Auxiliary Local Refinement via ICP.: To enhance temporal continuity and metric precision, we introduce ICP as a lightweight auxiliary tracking module. Let $\mathbf{T}^*_{t-1}$ denote the optimized pose from the previous frame. ICP assumes small inter-frame motion and performs local refinement by minimizing point-to-plane residuals:

$$\Delta\mathbf{T}_t = \arg\min_{\Delta\mathbf{T}} \sum_i \left(\mathbf{n}_{\pi(i)}^\top \big(\Delta\mathbf{T}\mathbf{T}^*_{t-1}\mathbf{s}_i - \mathbf{t}_{\pi(i)} \big) \right)^2, \tag{11}$$

where $\mathbf{s}_i$ denotes model points, $\mathbf{t}_{\pi(i)}$ denotes corresponding observed points, and $\mathbf{n}_{\pi(i)}$ represents surface normals of the observed cloud. The refined pose is updated as

$$\mathbf{T}_t^{ICP} = \Delta\mathbf{T}_t \mathbf{T}^*_{t-1}. \tag{12}$$

Since ICP searches only within a small local neighborhood in $SE(3)$, it converges rapidly and incurs low computational cost, making it suitable for high-frequency pose propagation. However, ICP cannot handle large pose jumps or global misalignment.

c) Adaptive Execution Mechanism.: To balance global robustness and local efficiency, we explicitly define a hierarchical execution strategy in which FP acts as the primary estimator and ICP serves as a secondary refinement module.

First, we quantify the inter-frame motion magnitude between the propagated pose and the FP prediction:

$$\delta_t = \left\| \log\left((\mathbf{T}^*_{t-1})^{-1} \mathbf{T}_t^{FP} \right) \right\|_2, \tag{13}$$

where $\log(\cdot)$ denotes the Lie algebra mapping from $SE(3)$ to $se(3)$. This measures the relative pose change in the tangent space.

Second, we evaluate geometric consistency using depth residual:

$$E_{depth}(\mathbf{T}) = \frac{1}{N} \sum_i \|\Pi(\mathbf{T}\mathbf{s}_i) - \mathbf{u}_i\|_2, \tag{14}$$

where $\Pi(\cdot)$ denotes the camera projection function and $\mathbf{u}_i$ are observed depth points.

We then define a unified trigger variable:

$$\alpha_t = I\big(\lambda_1\delta_t + \lambda_2 E_{depth}(\mathbf{T}_t^{FP}) + \lambda_3(1 - c_t) > \tau\big), \tag{15}$$

where c_t is the network confidence score, λ_i are balancing coefficients, and τ is a threshold.

The final pose selection follows a hierarchical rule:

$$\mathbf{T}_t^* = \begin{cases} \mathbf{T}_t^{FP}, & \alpha_t = 1, \\ \mathbf{T}_t^{ICP}, & \alpha_t = 0. \end{cases} \tag{16}$$

In practice, this mechanism operates as follows:

- When large motion, geometric inconsistency, or low network confidence is detected ($\alpha_t = 1$), FP is invoked to perform global re-localization.
- Otherwise, ICP propagates the pose locally to ensure temporal smoothness and computational efficiency.

d) Optimization Interpretation.: From an optimization viewpoint, the hybrid strategy constitutes a hierarchical minimization framework. FP provides a globally informed approximation in a semantic feature space, while ICP enforces local metric consistency in the geometric space. The trigger mechanism dynamically switches between global search and local refinement, thereby reducing drift accumulation, avoiding unnecessary repeated network inference, and preserving real-time performance.

Together with the scale recovery module, the proposed system forms a closed-loop optimization pipeline: scale recovery guarantees metric correctness, FP ensures global pose observability, and ICP maintains stable geometric tracking. This unified design enables robust and continuous object tracking under occlusion, abrupt motion, and complex scene conditions.

IV. Experiments

A. experimental Setup

Experiments are conducted using a workstation equipped with an NVIDIA RTX 4090 GPU and an Intel i9 CPU. All algorithms are implemented in Python and executed on Ubuntu 22.04. RGB-D observations are captured using an Intel RealSense D415i camera operating at a resolution of 640 × 480 with a frame rate of 30 FPS. The RGB stream is used for interactive object segmentation and model reconstruction, while the aligned depth stream provides metric geometric information for scale recovery and pose estimation.

To obtain accurate ground-truth object poses in real-world experiments, a calibrated motion capture system is employed. The motion capture cameras track reflective markers attached to the target object and provide high-precision pose measurements in the global coordinate frame. These measurements are synchronized with the RGB-D data and used as reference poses to evaluate the accuracy of the proposed method.

We evaluate the proposed CLM framework on both public benchmarks [2] and a self-collected dataset. Specifically, experiments are conducted on the HOPE and Nvidia's datasets, which contain challenging manipulation scenarios with varying viewpoints, occlusions, and object appearances. These datasets provide RGB-D observations together with accurate object models and ground-truth poses, allowing a comprehensive evaluation of pose estimation accuracy, as shown in Fig. 3.

In addition, we construct a custom RGB-D dataset to further evaluate the performance of CLM in interactive and real-world environments. During data acquisition, objects are observed under different viewing angles, motion speeds, and partial occlusion conditions. The RealSense D415i camera records synchronized RGB and depth streams, while the motion capture system provides precise pose annotations. This setup enables a controlled evaluation of both pose accuracy and tracking stability. All experiments are conducted using the same hardware and configuration to ensure fair comparisons across different methods.

B. Metrics

To closely follow standard evaluation protocols for real-time 6D pose tracking, we adopt the following metrics:

ADD(-S). We report the recall of ADD and ADD-S under the 0.1 object diameter threshold (ADD-0.1d) [2], which measures geometric pose accuracy and is widely used in pose estimation benchmarks.

Runtime (FPS). To evaluate the real-time performance of the proposed framework, we measure the average frames per second (FPS) during the online tracking stage. The runtime includes both the learned pose estimation of FoundationPose and the geometric refinement using ICP within the hybrid tracking strategy. Since object modeling and scale recovery are performed only once during initialization, these stages are excluded from the FPS measurement. For fair evaluation, the system is first warmed up for several frames to eliminate initialization overhead. The FPS is then computed over 300 consecutive frames. This measurement reflects the effective runtime of the proposed CLM tracking pipeline under continuous operation.

Relocalization Success Rate. Relocalization Success Rate (RSR) measures the ability of the tracker to recover the correct object pose after tracking failure or large object motion. A frame is considered successful if the estimated pose satisfies the ADD(-S) threshold (e.g., ADD(-S) $< 0.1d$, where d is the object diameter). The metric is defined as the percentage of frames in which the system successfully re-establishes accurate pose estimation.

C. Comparison with Prior Work

As shown in Table I, we compare CLM with representative instance-level, zero-shot, and category-level 6D pose methods [39, 40] to evaluate performance under different prior assumptions and deployment settings.

FoundationPose serves as a strong instance-level baseline. When accurate CAD models and segmentation masks are provided, it achieves high pose accuracy and stable temporal refinement. However, this performance depends on object-specific priors that are unavailable in unseen-object scenarios. CLM removes both CAD and mask dependencies by constructing an object model online from a single click and recovering metric scale through geometric alignment. Despite operating under weaker prior assumptions, CLM achieves competitive ADD(-S) and AR while maintaining real-time tracking.

We further compare with zero-shot approaches such as Any6D [38] and other foundation-model-based pose estimators. These methods directly regress 6D pose without instance CAD, but are typically designed for single-frame estimation and lack explicit metric-scale recovery or integrated temporal tracking. As a result, robustness degrades under occlusion, viewpoint changes, or tracking interruptions. In contrast, CLM integrates online modeling, metric-scale alignment, and learned refinement, enabling stable tracking and reliable re-localization.

ZeroPose proposes a CAD-prompted zero-shot 6D pose estimation framework that follows a Discovery–Orientation–Registration pipeline to detect object instances and estimate poses using feature matching between RGB-D observations and a given CAD model. It generalizes to unseen objects without retraining but still requires an accurate CAD model and relies on multi-stage matching. In contrast, CLM eliminates CAD dependence by interactively reconstructing scale-consistent object models from RGB-D input and enabling real-time pose tracking, making it more suitable for interactive robotic manipulation scenarios.

Quantitative results in Table II support these observations. While FoundationPose achieves the highest accuracy under full supervision, CLM substantially narrows the performance gap without requiring CAD models or masks. Compared to zero-shot and category-level methods, CLM demonstrates consistently higher re-localization success and occlusion robustness, together with competitive real-time performance. These results indicate that CLM achieves a favorable trade-off between prior dependency, tracking stability, and deployment practicality.

D. Ablation Study

We further analyze the contribution of color and shape cues in the proposed particle-based scale recovery module. Removing either cue leads to a noticeable decrease in pose accuracy and relocalization robustness, while the full model achieves the best performance, demonstrating the effectiveness of the proposed color-shape joint optimization.

TABLE I: System-level comparison of representative 6D pose estimation methods:

Method	No CAD	Model Generalizability	Interaction	Real-time Tracking	Metric Scale Recovery
NOCS [29]	✗	✗	✗	Limited	✓
FS-Net [30] / SPD	✗	✗	✗	Limited	✓
ZeroPose	✗	✗	✗	✗	Partial
Any6D [38]	✓	✓	✗		✓
FoundationPose [6]	✗	✓	✗	✓	✓
CLM (Ours)	✓	✓	✓	✓	✓

TABLE II: Comparison of representative 6D pose estimation pipelines in terms of pose accuracy, robustness, and runtime efficiency.

Method	ADD(-S)↑	RSR (%)↑	FPS↑
FoundationPose [6]	0.95	0	16
Any6D	0.96	0	
ZeroPose	0.83	0	
Category-level 6D Pose	0.76	0	
CLM (Ours)	**0.95**	**0.96 (Click)**	**19**

TABLE III: Ablation Study of a Particle Model Size Recovery Module Driven by Color and Shape.

Method	ADD(-S)↑	RSR (%)↑
w/o color + shape	0.64	0.46
w/o shape	0.72	0.61
w/o color	0.85	0.68
Ours	**0.96**	**0.97**

The ablation results in Table III demonstrate the effectiveness of incorporating both color and shape cues into the proposed particle-based optimization module. When both cues are removed (w/o color + shape), the system relies only on basic geometric alignment, resulting in relatively low performance with an ADD(-S) score of 0.64 and a relocalization success rate of 0.46. Introducing shape information alone (w / o color) significantly improves the results, indicating that geometric structure plays a critical role in stable pose estimation. Similarly, incorporating only color information (w/o shape) provides moderate improvement, suggesting that appearance cues also contribute to identifying reliable correspondences. The full model, which jointly leverages both color and shape features, achieves the best performance with an ADD(-S) of 0.96 and a relocalization success rate of 0.97. These results indicate that the complementary nature of color and geometric cues effectively enhances both pose accuracy and tracking robustness. Joint optimization enables more reliable particle selection on the reconstructed surface, leading to improved scale alignment and more stable pose estimation during tracking.

Importantly, the improvements from geo-only PSO to geo+color PSO are not marginal refinements but systematic gains across accuracy and robustness metrics. This trend indicates that robust initialization quality—rather than solely temporal refinement—plays a decisive role in stable unseen-object tracking. The results demonstrate that the proposed color-shape alignment is not merely an auxiliary enhancement, but a key component enabling reliable deployment under realistic visual conditions.

V. CONCLUSIONS

We propose an interactive framework, CLM, for real-time 6D pose estimation of unseen objects. This method resolves scale inconsistencies in RGB-D-driven reconstruction through a color-shape consistency optimization mechanism, enabling metric model recovery without relying on pre-scanned CAD models. Based on a motion-aware hybrid tracking strategy, it further balances stability and real-time performance by integrating global learning estimation with local geometric refinement. This framework maintains stability in dynamic scenes while reducing model dependency. Limitations include sensitivity to depth quality and the absence of an online shape update mechanism, which may affect performance under severe initial occlusions. Future research could explore the integration of online shape refinement techniques and collision-aware planning to enhance robustness and adaptability in complex operational environments.

REFERENCES

[1] X. Chen, F.-J. Chu, P. Gleize, K. J. Liang, A. Sax, H. Tang, W. Wang, M. Guo, T. Hardin, X. Li *et al.*, "Sam 3d: 3dfy anything in images," *arXiv preprint arXiv:2511.16624*, 2025.

[2] T. Hodan, F. Michel, E. Brachmann, W. Kehl, A. GlentBuch, D. Kraft, B. Drost, J. Vidal, S. Ihrke, X. Zabulis *et al.*, "Bop: Benchmark for 6d object pose estimation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 19–34.

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

[4] M. Sundermeyer, A. Mousavian, R. Triebel, and D. Fox, "Contact-graspnet: Efficient 6-dof grasp generation in cluttered scenes," in *2021 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2021, pp. 13 438–13 444.

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

[6] B. Wen, W. Yang, J. Kautz, and S. Birchfield, "Foundationpose: Unified 6d pose estimation and tracking of novel objects," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 17 868–17 879.

[7] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4015–4026.

[8] J. Lin, L. Liu, D. Lu, and K. Jia, "Sam-6d: Segment anything model meets zero-shot 6d object pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 27 906–27 916.

[9] S. Hoque, M. Y. Arafat, S. Xu, A. Maiti, and Y. Wei, "A comprehensive review on 3d object detection and 6d pose estimation with deep learning," *IEEE Access*, vol. 9, pp. 143 746–143 770, 2021.

[10] P. Knap, "Human modelling and pose estimation overview," *arXiv preprint arXiv:2406.19290*, 2024.

[11] P. J. Besl and N. D. McKay, "Method for registration of 3-d shapes," in *Sensor fusion IV: control paradigms and data structures*, vol. 1611. Spie, 1992, pp. 586–606.

[12] S. Rusinkiewicz and M. Levoy, "Efficient variants of the icp algorithm," in *Proceedings third international conference on 3-D digital imaging and modeling*. IEEE, 2001, pp. 145–152.

[13] S. Hinterstoisser, V. Lepetit, S. Ilic, S. Holzer, G. Bradski, K. Konolige, and N. Navab, "Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes," in *Asian conference on computer vision*. Springer, 2012, pp. 548–562.

[14] C. Wang, D. Xu, Y. Zhu, R. Martín-Martín, C. Lu, L. Fei-Fei, and S. Savarese, "Densefusion: 6d object pose estimation by iterative dense fusion," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3343–3352.

[15] Y. He, W. Sun, H. Huang, J. Liu, H. Fan, and J. Sun, "Pvn3d: A deep point-wise 3d keypoints voting network for 6dof pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 632–11 641.

[16] S. Peng, Y. Liu, Q. Huang, X. Zhou, and H. Bao, "Pvnet: Pixel-wise voting network for 6dof pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4561–4570.

[17] S. Zakharov, I. Shugurov, and S. Ilic, "Dpod: 6d pose object detector and refiner," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1941–1950.

[18] Y. He, H. Huang, H. Fan, Q. Chen, and J. Sun, "Ffb6d: A full flow bidirectional fusion network for 6d pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 3003–3013.

[19] R. L. Haugaard and A. G. Buch, "Surfemb: Dense and continuous correspondence distributions for object pose estimation with learnt surface embeddings," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 6749–6758.

[20] X. Liu, "Gnc-pose: Geometry-aware gnc-pnp for accurate 6d pose estimation," *arXiv preprint arXiv:2512.06565*, 2025.

[21] S. Iwase, X. Liu, R. Khirodkar, R. Yokota, and K. M. Kitani, "Repose: Fast 6d object pose refinement via deep texture rendering," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3303–3312.

[22] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes," *arXiv preprint arXiv:1711.00199*, 2017.

[23] W. Kehl, F. Manhardt, F. Tombari, S. Ilic, and N. Navab, "Ssd-6d: Making rgb-based 3d detection and 6d pose estimation great again," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1521–1529.

[24] W. Li, Y. Luo, P. Wang, Z. Qin, H. Zhou, and H. Qiao, "Recent advances on application of deep learning for recovering object pose," in *2016 IEEE International Conference on Robotics and Biomimetics (ROBIO)*. IEEE, 2016, pp. 1273–1280.

[25] J. Sun, Z. Wang, S. Zhang, X. He, H. Zhao, G. Zhang, and X. Zhou, "Onepose: One-shot object pose estimation without cad models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 6825–6834.

[26] X. He, J. Sun, Y. Wang, D. Huang, H. Bao, and X. Zhou, "Onepose++: Keypoint-free one-shot object pose estimation without cad models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 35 103–35 115, 2022.

[27] Y. Liu, Y. Wen, S. Peng, C. Lin, X. Long, T. Komura, and W. Wang, "Gen6d: Generalizable model-free 6-dof object pose estimation from rgb images," in *European Conference on Computer Vision*. Springer, 2022, pp. 298–315.

[28] B. Wen and K. Bekris, "Bundletrack: 6d pose tracking for novel objects without instance or category-level 3d models," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 8067–8074.

[29] H. Wang, S. Sridhar, J. Huang, J. Valentin, S. Song, and L. J. Guibas, "Normalized object coordinate space for category-level 6d object pose and size estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 2642–2651.

[30] W. Chen, X. Jia, H. J. Chang, J. Duan, L. Shen, and A. Leonardis, "Fs-net: Fast shape-based network for category-level 6d object pose estimation with decoupled rotation mechanism," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 1581–1590.

[31] Y. Di, R. Zhang, Z. Lou, F. Manhardt, X. Ji, N. Navab, and F. Tombari, "Gpv-pose: Category-level object pose estimation via geometry-guided point-wise voting," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 6781–6791.

[32] Y. Labbé, L. Manuelli, A. Mousavian, S. Tyree, S. Birchfield, J. Tremblay, J. Carpentier, M. Aubry, D. Fox, and J. Sivic, "Megapose: 6d pose estimation of novel objects via render & compare," *arXiv preprint arXiv:2212.06870*, 2022.

[33] K. Park, A. Mousavian, Y. Xiang, and D. Fox, "Latentfusion: End-to-end differentiable reconstruction and rendering for unseen object pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 710–10 719.

[34] T. Cao, F. Luo, J. Qin, Y. Jiang, Y. Wang, and C. Xiao, "ig-6dof: Model-free 6dof pose estimation for unseen object via iterative 3d gaussian splatting," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 6436–6446.

[35] Y. Jin, V. Prasad, S. Jauhri, M. Franzius, and G. Chalvatzaki, "6dope-gs: Online 6d object pose estimation using gaussian splatting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 8032–8043.

[36] W. Lu, G. Wan, Y. Zhou, X. Fu, P. Yuan, and S. Song, "Deepvcp: An end-to-end deep neural network for point cloud registration," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 12–21.

[37] H. Zhao, S. Wei, D. Shi, W. Tan, Z. Li, Y. Ren, X. Wei, Y. Yang, and S. Pu, "Learning symmetry-aware geometry correspondences for 6d object pose estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 14 045–14 054.

[38] T. Lee, B. Wen, M. Kang, G. Kang, I. S. Kweon, and K.-J. Yoon, "Any6d: Model-free 6d pose estimation of novel objects," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025, pp. 11 633–11 643.

[39] T. Hodan, D. Barath, and J. Matas, "Epos: Estimating 6d pose of objects with symmetries," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 703–11 712.

[40] Z. Li, G. Wang, and X. Ji, "Cdpn: Coordinates-based disentangled pose network for real-time rgb-based 6-dof object pose estimation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 7678–7687.