RETRIEVAL-AUGMENTED TRANSPARENT OBJECT PERCEPTION FOR AUTONOMOUS LABORATORIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Transparent objects are ubiquitous in autonomous laboratories, yet their geometry remains difficult to perceive due to refraction, reflection, and unreliable visual cues. Existing methods largely infer transparent geometry from observations alone, overlooking a distinctive property of laboratory environments: they are semi-open worlds, where many instruments are standardized products with known or retrievable 3D models. We introduce *R*etrieval-*A*ugmented *T*ransparent *D*epth *E*stimation (*RATDE*), a framework that grounds visual depth recovery with external object geometry. Given a transparent laboratory object, our approach retrieves candidate CAD models, encodes their geometry as point-cloud features, and establishes image-to-CAD correspondences to obtain pose-aware local geometric tokens. These geometric priors condition SeeClear's generative opacification. The generated image and consistency-filtered SGBM depth from the original transparent stereo pair are then passed to a fixed LingBot-Depth backend. The framework uses pose supervision during training while requiring no ground-truth pose at inference time. To study this problem, we construct a new laboratory transparent-object dataset, *LabSense-Glass*, comprising nine rendered laboratory scenes and 1,336 stereo samples with paired transparent–opaque images and geometric annotations. With the exact CAD assets of all transparent targets available in memory, we evaluate depth recovery and CAD retrieval, comparing the full pipeline against direct-depth and image-only opacification baselines, as well as a random-CAD control at fixed model weights. Our results demonstrate that exploiting retrievable geometry provides a strong prior for transparent perception, suggesting a broader principle for robotic laboratories: when geometry already exists in the world, perception should retrieve and ground it rather than infer it entirely from scratch.

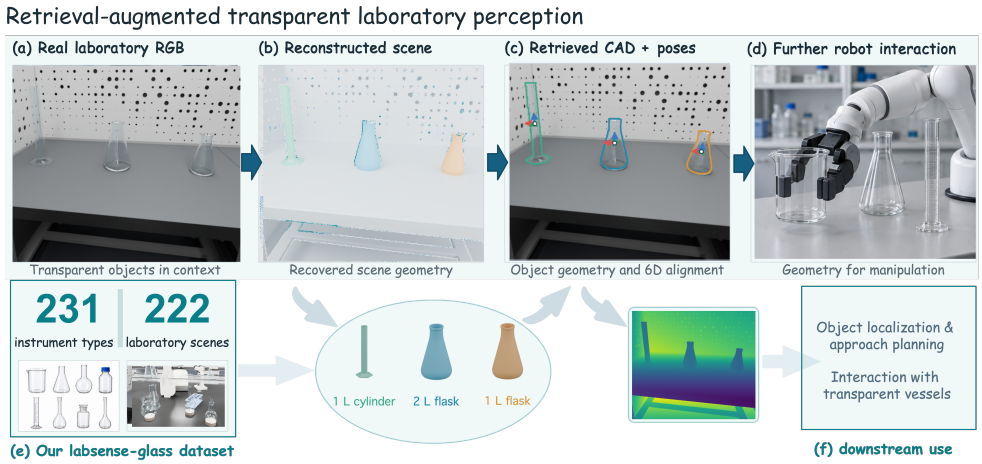

Figure 1: Retrieval-augmented transparent laboratory perception. Our LabSense-Glass dataset spans 231 instrument types and 222 lab scenes. Retrieved CAD geometry supports scene reconstruction and 6D pose estimation for further downstream robotic interaction.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Transparent objects are common in automated laboratories, where beakers, flasks, graduated cylinders, reagent bottles, and other instruments are frequently made of glass or transparent plastic (Ren et al., 2026; Li et al., 2026). Their reflection, refraction, and transmission make depth estimation particularly challenging: active depth sensors often produce missing or erroneous measurements, while RGB observations are strongly affected by background appearance visible through transparent surfaces. Reliable depth estimation of these objects is nevertheless important for downstream robotic tasks such as pose estimation, grasping, and manipulation (Sajjan et al., 2020; Fang et al., 2022; Chen et al., 2022).

Existing methods recover transparent-object geometry primarily from visual observations. ClearGrasp (Sajjan et al., 2020) predicts surface normals and occlusion boundaries to complete corrupted depth, while TransCG (Fang et al., 2022) learns depth completion from real RGB-D observations. More recent approaches exploit learned shape priors or generative models. AISPO (Chen et al., 2026) learns an affine-invariant shape prior to improve geometric consistency, whereas SeeClear (Wang et al., 2026) introduces generative opacification, converting transparent objects into opaque appearances from which off-the-shelf monocular depth estimators can recover depth. These methods have substantially improved transparent-object perception, but recovering geometry from transparent appearance remains ambiguous when visual cues are weak or distorted.

In laboratory environments, however, most transparent objects are standardized instruments with known 3D geometry. A beaker or reagent bottle may be difficult to reconstruct accurately from its transparent appearance, while its corresponding CAD model can already be available in a laboratory asset library. This provides a useful geometric prior that is largely complementary to visual evidence. We therefore ask a straightforward question: *Can known CAD geometry be used to improve depth estimation of transparent laboratory objects in the semi-open world?*

To address this problem, we propose ***R***etrieval-***A***ugmented ***T***ransparent ***D***epth ***E***stimation (***RATDE***), a framework that augments depth estimation with object-specific CAD geometry. Given the CAD model associated with an observed object, we sample its surface into a point cloud and encode it into geometric features. A key challenge is that the CAD model is represented in a canonical object coordinate system, while the object appears under an unknown pose in the image. We therefore establish correspondences between RGB features and CAD geometry to obtain local geometric features associated with the current observation. Ground-truth object poses are used to supervise this correspondence during training, but are not required during inference. The resulting CAD features are introduced as additional geometric conditions for generative opacification, allowing depth estimation to exploit both visual observations and object-specific 3D shape information.

To evaluate this approach in laboratory environments, we construct *labsense-glass*, a synthetic dataset comprising nine rendered laboratory scenes and 1,336 stereo samples. Each sample contains transparent left- and right-view images and their matched opaque counterparts, accompanied by geometric annotations. The source asset library contains 231 distinct instrument assets collected from 222 source scenes; The dataset covers diverse laboratory objects, poses, backgrounds, illumination conditions, and occlusions, and enables us to evaluate whether CAD conditioning improves depth estimation beyond visual conditions alone.

The main contributions of our work are summarized as follows:

- We investigate the use of object-specific CAD geometry as an additional condition for transparent depth estimation, motivated by the availability of standardized 3D assets for laboratory instruments.
- We develop a CAD-conditioned transparent depth estimation framework, *RATDE*, that encodes CAD point clouds and establishes RGB-CAD correspondences to provide local geometric conditions, without requiring ground-truth object poses at inference time.
- We construct a laboratory transparent-object dataset, *labsense-glass*, covering 231 instruments and 222 laboratory scenes, with observations associated with corresponding 3D assets, and use it to systematically evaluate CAD-conditioned transparent depth estimation.

2 Related Work

Transparent-object depth estimation. Transparent objects challenge conventional RGB-D perception because reflection, refraction, and transmission often produce unreliable depth. ClearGrasp recovers transparent geometry using predicted surface normals and boundaries (Sajjan et al., 2020), while TransCG introduces a real-world benchmark and RGB-D depth completion method (Fang et al., 2022). ClearPose extends transparent-object perception to segmentation, depth completion, and 6D pose estimation (Chen et al., 2022). More recently, AISPO incorporates an affine-invariant shape prior for transparent and non-Lambertian depth completion (Chen et al., 2026). In contrast, we exploit *object-specific CAD geometry* as an additional prior for standardized laboratory objects.

Generative opacification and depth reconstruction. SeeClear introduces *generative opacification*, converting transparent objects into geometrically consistent opaque appearances before depth estimation (Wang et al., 2026). General-purpose geometry models such as MoGe and MoGe-2 recover dense geometry from monocular images (Wang et al., 2025a;b), while LingBot-Depth learns RGB-guided depth completion and refinement from imperfect depth inputs (Tan et al., 2026). We build on these advances by conditioning generative opacification with object-specific CAD geometry, while using a fixed downstream depth estimator.

CAD priors and RGB–CAD correspondence. CAD models have been widely used for geometric matching and pose estimation. SurfEmb learns image-to-surface correspondences (Haugaard & Buch, 2022), 2D3D-MATR performs image–point-cloud matching (Li et al., 2023), and FoundationPose and FoundPose exploit CAD models for novel-object pose estimation (Wen et al., 2024; Örnek et al., 2024). Unlike these methods, we use RGB–CAD correspondence to obtain observation-relevant geometric features that condition transparent-object reconstruction and subsequent depth estimation.

3 A paired laboratory dataset and object memory

3.1 Scene and asset construction

We construct *labsense-glass*, a paired laboratory dataset for transparent-object depth estimation. Our source library contains 231 distinct instrument assets collected from 222 source scenes. Vessel families include beakers, flasks, bottles, and graduated cylinders. We compose laboratory scenes in Universal Scene Description (USD) from referenced assets, separating vessel geometry, material assignments, and scene placement. Glass materials are assigned to vessel meshes, while liquids and non-glass accessories retain separate materials. Layouts, distractors, and camera viewpoints provide variation in apparent object scale, clutter, partial occlusion, and background appearance.

We render nine laboratory scenes in Isaac Sim (Gao et al., 2026) with RTX Path Tracing, producing 1,336 stereo samples. Each sample contains transparent left- and right-view RGB images and their matched opaque counterparts, yielding 2,672 input RGB images and opaque supervision images. Figure 2 illustrates representative laboratory scenes, paired transparent and opaque renderings, and the corresponding geometric annotations. We will release the USD files for the nine composed scenes and the asset dependencies required to reconstruct them, together with the paired images and ground-truth annotations. The complete source library is not included in the release. More details are provided in Appendix A.

3.2 Paired rendering and geometric annotations

The dataset provides stereo pairs and transparent–opaque counterparts for both views of every sample. Within a stereo sample, the left and right views share scene geometry, layout, materials, and illumination, with intrinsics and world poses recorded for each camera. For each transparent–opaque correspondence, we keep geometry, object placement, camera parameters, and light settings fixed while replacing the target vessel material with an opaque material. The transparent image provides the observation, and the corresponding opaque image supplies the opacification target. Material replacement can alter shadows, reflections, and indirect illumination, so paired images need not be identical outside the vessel mask.

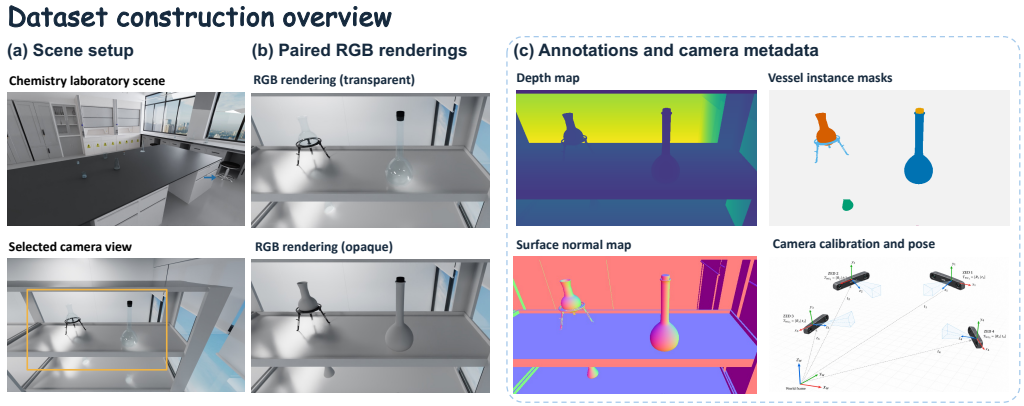

Figure 2: Overview of dataset construction and paired supervision. (a) Representative laboratory scene configurations and selected viewpoints. (b) Paired transparent and opaque RGB renderings with identical scene geometry and camera parameters. Opaque supervision targets are generated by replacing the transparent materials of the target vessels with opaque counterparts. (c) Corresponding axial depth maps, surface normals expressed in world coordinates, visible vessel instance masks, and a schematic illustration of camera poses.

RGB images and geometric annotations use the same camera poses and scene geometry. Isaac-native annotators provide camera-axis depth in metres, world-coordinate surface normals, and visible instance labels. Depth values are valid when finite and positive; zero denotes invalid depth. Instance masks describe visible regions without amodal completion. A shared instance mapping assigns consistent identifiers across the two stereo views. Category masks distinguish transparent vessels, liquids, and vessels with unverified material labels using these identifiers and designated scene nodes. Annotation conventions and the scope of independent geometric validation are described in Appendix A.

3.3 Memory construction and data organization

Each CAD memory entry stores a mesh, sampled surface points, reference renders, and source and coordinate metadata. Source identity and reversible source-to-canonical transforms connect the model to its representations. Canonical shape normalization and nominal asset dimensions are recorded separately from query-scene placement. Instance labels retain scene paths for grouping mesh components and linking known objects to their source assets. The memory contains the exact CAD asset of every transparent target object used in the experiments. Query-specific ground-truth identity, pose, and scale are excluded from inference.

Images and annotations are indexed by scene and stereo-sample identity, linking the two camera views and their transparent–opaque counterparts, which preserves correspondence across image inputs, geometric supervision, and CAD source records where an asset association is available.

The release will preserve these correspondences alongside the editable USD assets and scene descriptions. Researchers can use them to reconstruct the provided laboratory environments, rearrange instruments, and vary materials, illumination, and camera configurations. The same assets can also support new laboratory layouts and additional image–annotation pairs, allowing researchers to construct their own environments and evaluation protocols.

4 Retrieval-augmented laboratory scene perception

We propose *R*etrieval-*A*ugmented *T*ransparent *D*epth *E*stimation (*RATDE*), which combines retrieved point-cloud priors and RGB–CAD alignment to condition SeeClear's generative opacification. The generated scene image is passed to a frozen LingBot-Depth (Tan et al., 2026) backend together with consistency-filtered SGBM depth from the original transparent stereo pair (Hirschmuller, 2008). Figure 3 illustrates the pipeline.

4.1 Inputs, outputs, and assumptions

Given a calibrated, rectified stereo pair (I_L, I_R) and a CAD memory $\mathcal{M}$, we predict left-camera depth $\hat{D}$ and a point map $\hat{P}$. The exact CAD asset of each transparent target is available in $\mathcal{M}$,

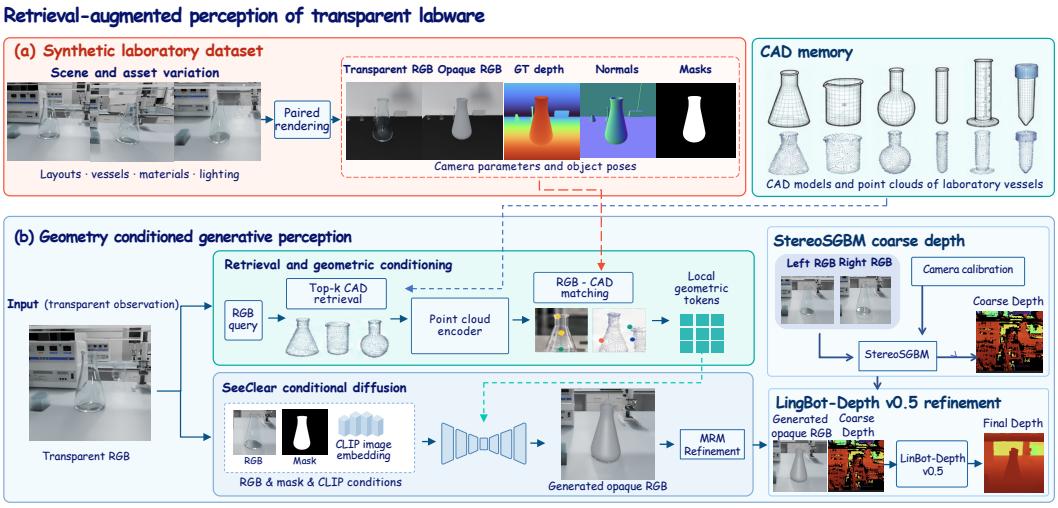

Figure 3: **Overview of RATDE for retrieval-augmented perception of transparent labware.** (a) Diverse synthetic laboratory scenes provide paired transparent–opaque renderings, with ground-truth depth, normals, masks, camera parameters, and object poses. (b) Given transparent RGB and an object mask, RATDE retrieves top-k CAD candidates and encodes their point clouds. These features and local geometric tokens from RGB–CAD matching augment SeeClear's RGB, mask, and CLIP conditions for opaque-image generation, followed by Mask Refinement Module (MRM) refinement. LingBot-Depth estimates scene depth from the composited opaque RGB and StereoSGBM depth computed from the original stereo pair. CAD models provide geometric priors; ground-truth object poses are not required at inference.

while its identity and pose in the query scene must be inferred from the observation. Retrieval and generation operate on I_L, while I_R supplies stereo evidence. A Mask R-CNN detector (He et al., 2017) provides instance masks $\{M_i\}$; each instance has its own crop, CAD candidates, pose hypotheses, and generated image. For each instance, A_i^{match} and A_i^{gen} map original left-image coordinates to the 224×224 matching crop and 512×512 generation crop, respectively. Each transform includes cropping, padding, resizing, and the corresponding pixel-center convention. The crop intrinsics are

$$K_i^{\text{match}} = A_i^{\text{match}} K_L, \qquad K_i^{\text{gen}} = A_i^{\text{gen}} K_L. \tag{1}$$

Each CAD entry stores a mesh, sampled surface points, reference views, and coordinate and scale metadata. Normalized shape coordinates and independently sourced asset dimensions are recorded separately; nominal proxies are not assumed to reproduce every physical surface detail. Inference uses predicted masks and public calibration, without query ground-truth identity, pose, depth, XYZ, or opaque RGB.

4.2 CAD RETRIEVAL AND RGB–CAD MATCHING

We construct a retrieved point-feature bank and an explicitly posed geometric condition using separate candidate rankings. For the point-feature bank, normalized silhouette overlap ranks each CAD across its reference views:

$$r_{ij} = \max_v \text{IoU}(\mathcal{N}(M_i), \mathcal{N}(S_{jv})), \qquad \mathcal{R}_i^{\text{pc}} = \text{TopK}_{j \in \mathcal{M}}\, r_{ij}. \tag{2}$$

Here $\mathcal{N}$ preserves aspect ratio on a common canvas; we use 24 views per asset and retain $K = 3$ distinct assets. For each asset, a frozen Utonia encoder (Zhang et al., 2026) supplies features at 256 point anchors:

$$g_{jn} = \phi_{\text{pc}}([\text{LN}(f_{jn}^{\text{U}}); \bar{X}_{jn}; n_{jn}; \log \ell_j]), \tag{3}$$

where $\bar{X}_{jn}$, n_{jn}, and ℓ_j denote normalized object coordinates, normals, and asset extents in the declared units.

For explicit alignment, we match the observed crop against CAD templates across the memory, following the image-to-model correspondence approach of FoundPose (Örnek et al., 2024). We extract spatial descriptors from the DINOv2-based encoder (Oquab et al., 2024) in a frozen MoGe-2 checkpoint (Wang et al., 2025b), without invoking its prediction heads. A learned shared projection

produces $f = \mathrm{normalize}(f^0 + Wf^0)$ from frozen descriptors f^0. Images of size 224×224 yield a 16×16 grid of 1024-dimensional descriptors; valid template patches retain their CAD surface coordinates. With mask-supported query patches $\mathcal{Q}_i$ and mutual nearest-neighbor matches $\mathcal{C}_{ijv}$, the identity score is

$$s_{ij} = \max_v \frac{1}{|\mathcal{Q}_i|} \sum_{(u,p)\in\mathcal{C}_{ijv}} \max(0, f_{iu}^\top f_{jvp}). \tag{4}$$

We retain the top three identities and up to four views per identity for pose fitting; an empty query support yields no alignment. This ranking is independent of $\mathcal{R}_i^{\mathrm{pc}}$, so the point-feature and pose branches may select different CADs.

4.3 Local RGB–CAD geometric conditioning

Matched query pixels u and template CAD points X define an object-to-camera pose $\hat{T}_{ij} = (\hat{R}_{ij}, \hat{t}_{ij})$ through

$$u^{\mathrm{match}} \simeq \pi(K_i^{\mathrm{match}}(\hat{R}_{ij}X + \hat{t}_{ij})). \tag{5}$$

We use EPnP–RANSAC followed by Levenberg–Marquardt refinement, checking correspondence spread, positive depth, inlier support, and reprojection error. Within each identity, valid poses are ranked by inlier count and then reprojection error. We examine the three candidate identities in descending local-matching score s_{ij}. If no pose for the top-ranked identity passes validation, we try the second-ranked identity and then the third. We select the best valid pose from the first identity that passes these checks. If all three identities fail, the local condition is disabled while point-feature conditioning remains active.

Let c_j be the stored object-frame center and $L_j > 0$ the scalar normalization length, both expressed in nominal meters. For $X_n = (X_m - c_j)/L_j$, a pose (R_n, t_n) fitted in normalized coordinates is converted to dimensional coordinates by

$$R_m = R_n, \qquad t_m = L_j t_n - R_n c_j. \tag{6}$$

Here X_m and X_n denote dimensional and normalized CAD coordinates, respectively; this conversion uses stored asset metadata.

Rendering the posed CAD with calibrated intrinsics gives first-hit axial depth z_i, camera-frame normals n_i^c, and silhouette S_i. Valid support V_i requires a unique, finite, positive first hit inside the predicted mask; overlapping hypotheses are assigned to the nearest predicted CAD surface with deterministic ties. We resample these channels into the 512×512 generation crop using its recorded crop transform, with nearest-neighbor sampling for the geometric channels and support masks. The crop is partitioned into a 16×16 grid. Each nonempty cell selects the valid pixel nearest its center and forms a token from an 11-dimensional descriptor:

$$c_{ip} = \phi_{\mathrm{loc}}([X_{ip}^c; n_{ip}^c; \log(1 + z_{ip}/z_{\mathrm{ref}}); \rho_{ip}^S; \tilde{u}_{ip}; \tilde{v}_{ip}; \rho_{ip}^V]). \tag{7}$$

Let u_{ip}^{src} be the original-image pixel associated with the selected crop pixel under the same nearest-neighbor sampling map. We compute $X_{ip}^c = z_{ip} K_L^{-1} \bar{u}_{ip}^{\mathrm{src}}$, where the bar denotes homogeneous coordinates. The positional channels $(\tilde{u}_{ip}, \tilde{v}_{ip})$ give the normalized location of this source pixel within the generation crop. The quantities ρ_{ip}^S and ρ_{ip}^V are the silhouette and valid-support fractions within the cell, respectively; $z_{\mathrm{ref}} = 1\,\mathrm{m}$ for nominal-meter assets. Empty cells are masked out; these camera-space tokens complement the separate Utonia point-feature bank.

4.4 Geometry-conditioned generative opacification

We retain SeeClear's RGB, mask heatmap, and pooled CLIP conditions (Wang et al., 2026; Radford et al., 2021), adding both CAD residuals at the UNet middle block. At timestep t, pooled hidden queries $\bar{q}_i^t$ and mean point tokens $\bar{g}_j$ define candidate weights:

$$w_{ij}^t = \mathrm{softmax}_{j\in\mathcal{R}_i^{\mathrm{pc}}}(r_{ij}/\tau + \psi(\bar{q}_i^t, \bar{g}_j)), \qquad \tau = 0.1, \tag{8}$$

where ψ is a learned projected dot product. Spatial hidden queries attend to point tokens with these weights as priors:

$$a_{iujn}^t = \mathrm{softmax}_{j,n}\left(\frac{(q_{iu}^t)^\top W_K^{\mathrm{pc}} g_{jn}}{\sqrt{d_h}} + \log w_{ij}^t\right), \qquad C_{iu}^{\mathrm{pc,t}} = \sum_{j,n} a_{iujn}^t W_V^{\mathrm{pc}} g_{jn}. \tag{9}$$

Each candidate contributes the same number of points; a learned scalar gate γ_i^t scales the residual. The local branch uses four attention heads with a total width of 128 (32 dimensions per head). Queries attend to valid tokens in a 3×3 neighborhood of the corresponding grid cell, with a support mask m_i derived from V_i:

$$h_i^{t,\mathrm{pc}} = h_i^t + \gamma_i^t W_O^{\mathrm{pc}} C_i^{\mathrm{pc,t}}, \qquad h_i^{t,+} = h_i^{t,\mathrm{pc}} + m_i \odot W_O^{\mathrm{loc}} C_i^{\mathrm{loc,t}}. \tag{10}$$

Local queries use $h_i^{t,\mathrm{pc}}$, and their output projection is zero-initialized. Missing local geometry yields a zero local residual while retaining point-feature conditioning.

Writing the point and local conditions as $\mathcal{G}_i$ and $\mathcal{C}_i$, the opaque crop and insertion mask are

$$\hat{I}_i^{\mathrm{op}} = G_\theta(I_i, H(M_i), e_{\mathrm{img}}(I_i); \mathcal{G}_i, \mathcal{C}_i), \qquad \tilde{M}_i = R_{\mathrm{MRM}}(\hat{I}_i^{\mathrm{op}}, I_i, H(M_i)). \tag{11}$$

Mask Refinement Module (MRM), inherited from SeeClear, predicts the insertion support from the generated crop, original RGB crop, and mask prompt. After restoring crops and refined binary masks to scene coordinates, we composite them in fixed proposal order:

$$I^{(i)} = (1 - \bar{M}_i) \odot I^{(i-1)} + \bar{M}_i \odot \bar{I}_i^{\mathrm{op}}, \qquad I^{(0)} = I_L. \tag{12}$$

Later insertions overwrite overlaps; pixels outside refined supports retain the original RGB, giving $\tilde{I}_L^{\mathrm{op}} = I^{(N)}$.

4.5 Scene geometry reconstruction

We compute metric SGBM depth from the original transparent stereo pair and apply left–right consistency filtering with a one-pixel threshold. Rejected measurements are set to zero, while accepted depths retain their values, forming D_{in}. A LingBot-Depth v0.5 RGB-D completion backend (Tan et al., 2026) predicts

$$\hat{D} = H_{\mathrm{LingBot}}(\tilde{I}_L^{\mathrm{op}}, D_{\mathrm{in}}, K_L), \qquad \hat{P}(u) = \hat{D}(u) K_L^{-1} \bar{u}. \tag{13}$$

Depth is camera-axis z; per-instance depth is extracted from the scene prediction. Original-RGB and generated-RGB comparisons share the same backend, calibration, and filtered SGBM input, isolating the RGB representation change.

4.6 Training and inference protocol

We first adapt the detector and establish an image-only SeeClear model using low-rank updates (Hu et al., 2021). With the visual and Utonia encoders frozen, we train only the shared residual projection W for local descriptor matching. For a query descriptor f_q, a geometrically verified positive template descriptor f_{p^+}, and negatives $\mathcal{N}(q, p^+)$, the pairwise contrastive loss is

$$\ell_{\mathrm{match}}(q, p^+) = -\log \frac{\exp(f_q^\top f_{p^+}/\tau_{\mathrm{match}})}{\exp(f_q^\top f_{p^+}/\tau_{\mathrm{match}}) + \sum_{p^- \in \mathcal{N}(q,p^+)} \exp(f_q^\top f_{p^-}/\tau_{\mathrm{match}})}, \ \tau_{\mathrm{match}} = 0.10. \tag{14}$$

Positive pairs are mutually visible correspondences verified using training geometry. Negatives are spatially separated template patches from the same CAD and view. We average the loss across positive views for each query patch, then across supported query patches and instances. The projection is zero-initialized and fitted using the training data; the descriptor encoder remains frozen. Opaque supervision is rendered from the same scene and camera by replacing transparent vessel materials; predicted masks determine training crops and prompts. We adapt the point-feature conditioning module with the image-only generator frozen, then freeze this model and train the local residual; MRM and LingBot remain frozen.

For target opaque latent z_0, noisy latent z_t, and sampled noise ϵ, both CAD-conditioning adaptation stages minimize full-latent noise prediction error:

$$\mathcal{L}_{\mathrm{denoise}} = \mathbb{E}_{i,t,\epsilon}\left[\frac{1}{|z_0|}\,\|\epsilon - \epsilon_\theta(z_t, t; I_i, H(M_i), e_{\mathrm{img}}(I_i), \mathcal{G}_i, \mathcal{C}_i)\|_2^2\right]. \tag{15}$$

Only the conditioning module selected for each stage receives updates; its CAD and pose inputs remain predictions rather than ground-truth replacements. At inference, model parameters are fixed

Table 1: Comparison of depth estimation pipelines. The full RATDE model combines retrieved CAD point-feature conditioning and RGB–CAD geometric alignment for SeeClear-based opacification, followed by LingBot-Depth with SGBM depth from the original stereo pair. MAE and RMSE are in millimeters; coverage and S_{50} are percentages. Errors use valid final predictions. Coverage and S_{50} retain the full valid reference support.

	Depth Error ↓		Robustness ↑	
Method	MAE (mm)	RMSE (mm)	Coverage (%)	S_{50} (%)
SGBM	391.6	2611.1	60.6	21.6
MoGe-2	316.7	357.1	67.2	17.8
LingBot + SGBM	64.8	184.3	61.7	49.4
SeeClear + MoGe-2	256.7	282.2	60.6	2.2
SeeClear + T4T-tiny + MoGe-2	193.6	256.1	63.5	9.4
RATDE (Ours)	**48.6**	**97.5**	**75.4**	**64.4**

and poses are estimated per observation; query ground truth is not used by the inference pipeline. Our reported comparisons include the original-RGB depth baseline, SeeClear, the random-CAD control, and the full model. F0 uses the SeeClear released checkpoint without our image-only LoRA adaptation or CAD-conditioning modules. The random-CAD control and the full model share the same trained weights.

5 Experiments

5.1 Evaluation questions and setup

Questions. We evaluate three questions: *Q1.* whether RATDE improves transparent-surface depth accuracy and coverage over direct depth-estimation baselines; *Q2.* whether the full CAD-conditioned pipeline improves reconstruction over image-only opacification with the same depth backend; and *Q3.* whether retrieved CAD candidates outperform random candidates at fixed model weights.

Models. We compare Stereo SGBM, MoGe-2 on the original transparent RGB image, LingBot-Depth v0.5 with SGBM depth, SeeClear with shared predicted vessel masks followed by MoGe-2, SeeClear with Trans4Trans-tiny segmentation followed by MoGe-2, and the full RATDE model with point-feature and local geometric conditioning followed by LingBot-Depth v0.5. All pipelines start from the original observations. For LingBot-Depth configurations, SGBM depth is computed from the original stereo pair; the backend receives the original left RGB in the direct baseline or the composited opacified left image in generation-based variants. More details are in Appendix. B.2.

Metrics. We evaluate depth estimation using mean absolute error (MAE), root mean squared error (RMSE), coverage, and S_{50}. For each evaluation frame, all methods use the same transparent-object reference region, excluding liquid pixels and restricting evaluation to finite, positive reference depths. A final depth prediction is valid if and only if it is finite and strictly positive; evaluation imposes no additional confidence, depth-range, predicted-instance-mask, or SGBM-support filter. MAE and RMSE are computed on pixels with valid reference and predicted depth and are reported in millimeters. Coverage is the fraction of valid reference pixels with a valid final depth prediction. S_{50} is the fraction of valid reference pixels with absolute depth error at most 50 mm; missing or invalid predictions count as failures. Coverage and S_{50} are reported as percentages. All four metrics are computed independently for each frame and then averaged with equal weight across evaluation frames. Depth is evaluated at its predicted metric scale without ground-truth scale or affine alignment. The reported metrics describe the transparent-object reference region. Metric definitions appear in Appendix B.1.

5.2 Experimental Results

On the evaluated laboratory scenes, RATDE achieves the lowest MAE and RMSE and the highest coverage and S_{50} among the compared methods (Table 1), with an MAE of 48.6 mm, an RMSE of 97.5 mm, 75.4% coverage, and 64.4% S_{50}. LingBot-Depth is employed as an RGB-D depth completion and refinement backend, with SGBM predictions computed from the original transparent stereo images serving as its initial depth input. Keeping this depth input and the LingBot backend

Table 2: Opacification and CAD-input comparisons (Q2). MAE and RMSE are in millimeters; coverage and S_{50} are percentages. Local denotes RGB–CAD matching. SeeClear uses its released checkpoint. The random-CAD control and full model share trained weights and matching architecture; random CAD replaces candidates in both conditioning branches.

	CAD Configuration		Depth Performance			
Variant	CAD Input	Local	MAE (mm) ↓	RMSE (mm) ↓	Coverage (%) ↑	S_{50} (%) ↑
LingBot + SGBM	–	No	64.8	184.3	61.7	49.4
SeeClear + LingBot + SGBM	–	No	58.2	177.3	63.5	52.1
Random-CAD	Random	Yes	56.5	155.6	62.7	53.0
Full RATDE	**Retrieved**	**Yes**	**48.6**	**97.5**	**75.4**	**64.4**

unchanged, our CAD-guided opacification pipeline reduces MAE by 25.0% and RMSE by 47.1% relative to the original-RGB baseline, while increasing coverage and S_{50} by 13.7 and 15.0 percentage points, respectively. This comparison measures the effect of replacing the original RGB with the output of the complete adapted opacification pipeline, with the depth-completion backend held fixed. It does not isolate the effects of laboratory adaptation, point-feature conditioning, or RGB–CAD alignment. Standalone SGBM exhibits a disproportionately large RMSE of 2611.1 mm, highlighting severe large-error outliers. Potential failure sources include incorrect CAD retrieval, pose misalignment, and incomplete instance masks, which can compromise the geometric consistency of the generated opaque images.

5.3 Ablations: CAD retrieval and geometric conditioning

We conduct ablation studies to identify where the improvement of our method comes from. The results are shown in Table. 2 and we have the following findings (More details are in Appendix. B.3):

Contribution of retrieved CAD. At fixed model weights, replacing random CADs with retrieved candidates reduces MAE by 14.0% and RMSE by 37.3%, while increasing coverage and S_{50} by 12.7 and 11.4 percentage points, respectively. This comparison supports the value of retrieval-selected CAD information across the combined conditioning pathway. Relative to SeeClear, the full model reduces MAE by 16.5% and RMSE by 45.0%; these improvements reflect the combined effects of model adaptation and CAD conditioning. Error metrics use each method's valid prediction support and are interpreted jointly with coverage and S_{50}.

Retrieval accuracy and interpretation. For the silhouette-based ranking r_{ij} in Eq. 2, exact-asset accuracy is 62.0% at Top-3 and 86.7% at Top-5, evaluated on the same retrieval queries. A query is counted as correct when its exact CAD asset appears among the top-k candidates. These rates do not measure the independent local-matching ranking s_{ij} or the correctness of the CAD selected after pose validation. We hypothesize that some incorrect candidates still provide useful shape information: beakers or conical flasks may have similar shapes while differing in physical size. However, this explanation remains unverified. The aggregate gains could be concentrated among queries with correct retrieval; establishing benefits from shape-similar incorrect candidates requires query-level analysis.

6 Conclusion

We present the LabSense-Glass dataset and RATDE, a retrieval-augmented framework for transparent-object perception in laboratories. Our key idea is to exploit a distinctive property of laboratories: many transparent vessels are standardized objects whose geometry can be retrieved from a known CAD library. By retrieving the corresponding CAD asset and using it to condition transparent-to-opaque reconstruction, our framework converts ambiguous transparent observations into a representation that is more amenable to conventional geometric perception. We further introduced a laboratory dataset with paired transparent and opaque observations to support systematic evaluation of this setting. Experiments demonstrate that incorporating retrieved geometric priors consistently improves transparent-object reconstruction and downstream depth estimation over approaches relying on visual observations alone. These results suggest a practical path toward robust geometric perception of transparent laboratory objects, providing a perceptual foundation for autonomous laboratory systems.

AI USE STATEMENT

In this work, we used generative AI tools for polishing our paper and helping us with coding. We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Xiaotong Chen, Huijie Zhang, Zeren Yu, Anthony Opipari, and Odest Chadwicke Jenkins. Clearpose: Large-scale transparent object dataset and benchmark. In *European conference on computer vision*, pp. 381–396. Springer, 2022.

Zhiming Chen, Linfang Zheng, Kun Zhang, Hyung Jin Chang, Wei Zhang, Hongyu Yu, and Hua Chen. Aispo: Enhancing depth reliability for robotic manipulation of non-lambertian objects via affine-invariant shape prior. *IEEE Robotics and Automation Letters*, 2026.

Hongjie Fang, Hao-Shu Fang, Sheng Xu, and Cewu Lu. Transcg: A large-scale real-world dataset for transparent object depth completion and a grasping baseline. *IEEE Robotics and Automation Letters*, 7(3):7383–7390, 2022.

Sicong Gao, Maurice Pagnucco, Tomasz Bednarz, and Yang Song. Nvidia isaac sim: Enabling scalable, gpu-accelerated simulation for robotics. *arXiv preprint arXiv:2606.03551*, 2026.

Rasmus Laurvig Haugaard and Anders Glent Buch. Surfemb: Dense and continuous correspondence distributions for object pose estimation with learnt surface embeddings. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6739–6748. IEEE, 2022.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

Heiko Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on pattern analysis and machine intelligence*, 30(2):328–341, 2008.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Minhao Li, Zheng Qin, Zhirui Gao, Renjiao Yi, Chenyang Zhu, Yulan Guo, and Kai Xu. 2d3d-matr: 2d-3d matching transformer for detection-free registration between images and point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14128–14138, 2023.

Rui Li, Zixuan Hu, Wenxi Qu, Jinouwen Zhang, Zhenfei Yin, Sha Zhang, Xuantuo Huang, Hanqing Wang, Tai Wang, Jiangmiao Pang, et al. Labutopia: High-fidelity simulation and hierarchical benchmark for scientific embodied agents. *Advances in Neural Information Processing Systems*, 38, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research* , pp. 1–31, January 2024. doi: 10.48550/arxiv.2304.07193. URL `https://hal.science/hal-04376640`.

Evin Pınar Örnek, Yann Labbé, Bugra Tekin, Lingni Ma, Cem Keskin, Christian Forster, and Tomáš Hodaň. Foundpose: Unseen object pose estimation with foundation features. *European Conference on Computer Vision (ECCV)*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Baochang Ren, Xinjie Liu, Xi Chen, Yanshuo Liu, Chenxi Li, Daqi Gao, Zeqin Su, Jintao Xing, Zirui Xue, Rui Li, et al. Labvla: Grounding vision-language-action models in scientific laboratories. *arXiv preprint arXiv:2606.13578*, 2026.

Shreeyak Sajjan, Matthew Moore, Mike Pan, Ganesh Nagaraja, Johnny Lee, Andy Zeng, and Shuran Song. Clear grasp: 3d shape estimation of transparent objects for manipulation. In *2020 IEEE international conference on robotics and automation (ICRA)*, pp. 3634–3642. IEEE, 2020.

Bin Tan, Changjiang Sun, Xiage Qin, Hanat Adai, Zelin Fu, Tianxiang Zhou, Han Zhang, Yinghao Xu, Xing Zhu, Yujun Shen, et al. Masked depth modeling for spatial perception. In *European Conference on Computer Vision*, pp. 453–470. Springer, 2026.

Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5261–5271. IEEE, 2025a.

Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. MoGe-2: Accurate monocular geometry with metric scale and sharp details. *arXiv preprint arXiv:2507.02546*, 2025b. URL https://arxiv.org/abs/2507.02546.

Xiaoying Wang, Yumeng He, Jingkai Shi, Jiayin Lu, Yin Yang, Ying Jiang, and Chenfanfu Jiang. SeeClear: Reliable transparent object depth estimation via generative opacification. *arXiv preprint arXiv:2603.19547*, 2026. URL https://arxiv.org/abs/2603.19547.

Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. Foundationpose: Unified 6d pose estimation and tracking of novel objects. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17868–17879. IEEE, 2024.

Yujia Zhang, Xiaoyang Wu, Yunhan Yang, Xianzhe Fan, Han Li, Yuechen Zhang, Zehao Huang, Naiyan Wang, and Hengshuang Zhao. Utonia: Toward one encoder for all point clouds. *arXiv preprint arXiv:2603.03283*, 2026.

A Dataset specification and additional examples

A.1 Asset records and rendering configuration

Asset records associate CAD meshes and sampled surface points with source identifiers, canonical coordinate frames, and nominal dimensions. Canonical shape normalization is recorded separately from object placement in the rendered scenes. The composed USD scenes retain object transforms, material assignments, lighting, and camera configurations. Images and annotations are indexed by scene and stereo-sample identity, linking the two camera views and their transparent–opaque counterparts. Table 3 summarizes the rendering configuration.

Table 3: Dataset rendering configuration.

Field	Configuration
Renderer	NVIDIA Isaac Sim with RTX Path Tracing; 256 samples per pixel.
Resolution and cameras	1280×720 pixels; parallel virtual stereo cameras with a 0.12 m baseline and a nominal focal length of 2.1 mm. Intrinsics and camera poses are recorded for each view.
Materials and illumination	OmniGlass material overrides use a refractive index of 1.49 and roughness of 0.02. Source-scene lighting is retained with scene-specific intensity multipliers ranging from 0.10 to 0.30. Liquids and non-glass accessories retain separate material assignments.
Scene sampling	1,336 stereo samples from nine laboratory scenes, using existing USD layouts and recorded camera viewpoints. Layouts and viewpoints provide variation in background appearance, clutter, and partial occlusion.
Depth / normal convention	Camera-axis depth in metres and surface normals in world coordinates, exported using Isaac Sim geometry annotators. Finite positive depth values are valid; zero denotes invalid depth.

A.2 Annotation conventions and reference support

RGB images and geometric annotations use the same scene geometry and camera calibration for each view. The depth annotations represent renderer-derived geometry, rather than simulated or measured sensor depth. Surface normals are expressed in world coordinates. Instance labels describe visible regions without amodal completion, and a shared instance mapping preserves object identities across the stereo views.

For transparent–opaque pairs, material replacement preserves object geometry, placement, camera parameters, and light settings. Changes in shadows, reflections, and indirect illumination may nevertheless alter pixels outside the target vessel masks.

Category labels distinguish transparent vessels, liquids, and vessels with unverified material labels. Transparent-object depth evaluation excludes liquid regions and uses finite, positive reference depth, following Appendix B.1.

B Implementation Details

B.1 Evaluation Metrics

Depth metrics. For each evaluation frame, let Ω_{GT} denote transparent-object reference pixels, excluding liquid regions, with finite, positive reference depth. Let $\Omega_{\mathrm{pred}} \subseteq \Omega_{\mathrm{GT}}$ contain those pixels with finite, positive final predicted depth. No additional confidence, depth-range, predicted-instance-mask, or SGBM-support filter is applied to the final predictions for evaluation. Reference

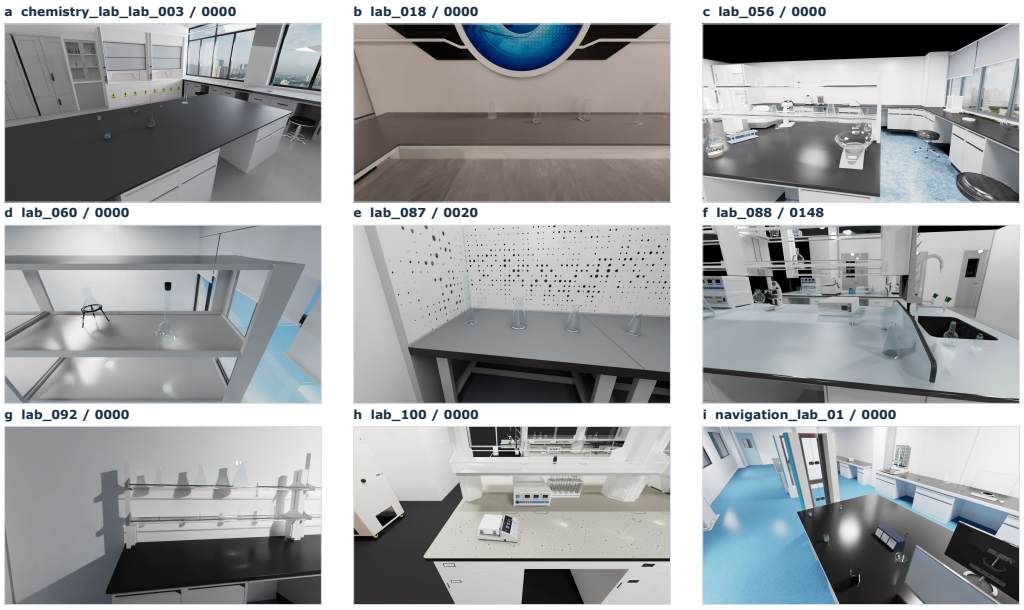

Figure 4: Additional dataset examples from the nine rendered laboratory scenes, illustrating variation in layout, background appearance, illumination, apparent vessel scale, and occlusion. Each panel identifies the source scene and frame.

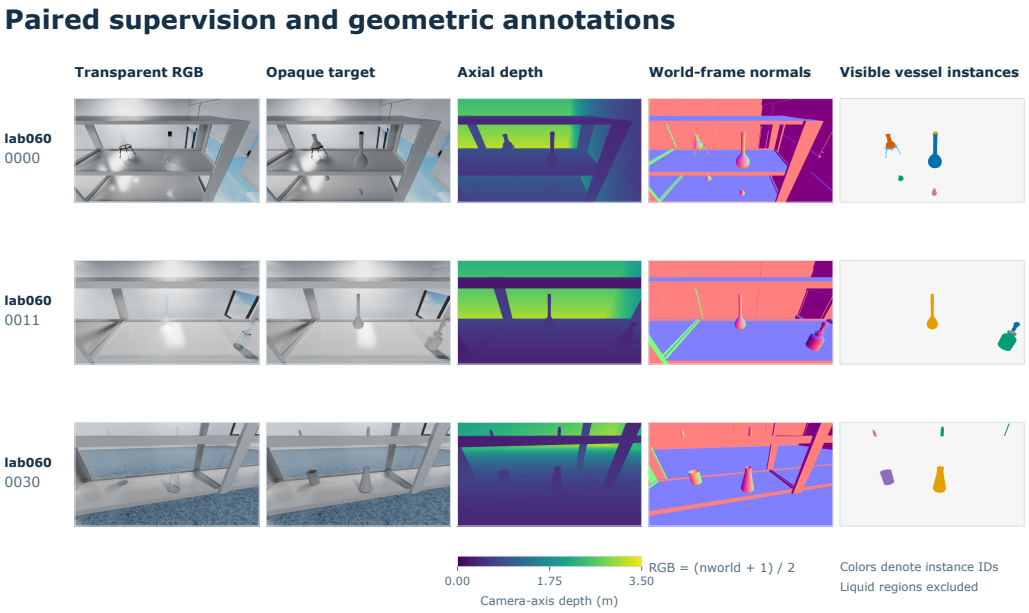

Figure 5: Paired supervision and geometric annotations. Each row shows a transparent RGB observation, its opaque target, camera-axis depth, world-coordinate surface normals, and visible vessel instance labels from the same view. Paired renderings share scene geometry and camera parameters. Depth maps use a common color scale, and instance colors distinguish vessel labels within each frame.

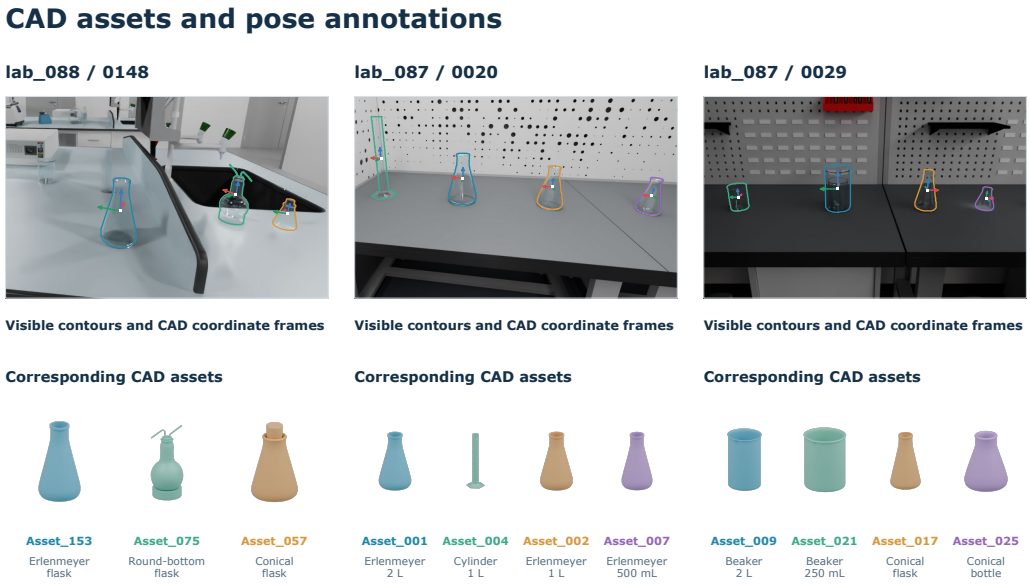

Figure 6: CAD assets and pose annotation examples. The upper panels overlay visible instance contours and CAD coordinate frames on the original RGB observations. Coordinate frames use ground-truth scene transforms, with origins at the CAD bounding-box centers. The lower panels show the corresponding CAD assets, scaled independently for display.

depth D and predicted depth $\hat{D}$ are expressed in millimeters. We compute

$$\mathrm{MAE} = \frac{1}{|\Omega_{\mathrm{pred}}|} \sum_{u \in \Omega_{\mathrm{pred}}} \left| \hat{D}(u) - D(u) \right|, \tag{16}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{|\Omega_{\mathrm{pred}}|} \sum_{u \in \Omega_{\mathrm{pred}}} \left(\hat{D}(u) - D(u) \right)^2}, \tag{17}$$

$$\mathrm{Coverage} = 100 \, \frac{|\Omega_{\mathrm{pred}}|}{|\Omega_{\mathrm{GT}}|}, \tag{18}$$

$$S_{50} = \frac{100}{|\Omega_{\mathrm{GT}}|} \sum_{u \in \Omega_{\mathrm{pred}}} \mathbf{1}\left[\left| \hat{D}(u) - D(u) \right| \leq 50\,\mathrm{mm} \right]. \tag{19}$$

MAE and RMSE are reported in millimeters, while coverage and S_{50} are percentages. Coverage measures valid-depth availability within the reference region. Missing or invalid predictions count as failures in S_{50}.

Metrics are computed independently for each frame. Frames without valid reference support are excluded from all four metrics. For MAE and RMSE, frames with no valid predictions are also excluded, and each metric is averaged with equal weight over the remaining frames for that method. Coverage and S_{50} are averaged with equal weight over all frames with valid reference support, including frames with no valid predictions, for which both metrics are zero.

All methods use the same transparent-object reference region, although their valid prediction supports and the frames included in MAE and RMSE may differ. These error metrics are therefore interpreted jointly with coverage and S_{50}. Depth is evaluated at the predicted metric scale without ground-truth scale or affine alignment.

CAD retrieval accuracy. Let $\mathcal{Q}$ denote the evaluated retrieval queries, $j_q^\star$ the exact reference CAD identity for query q, and $\mathcal{R}_q^{(k)}$ the top-k distinct CAD identities under the silhouette-based ranking in Eq. (2). Exact-asset Top-k accuracy is

$$\mathrm{Acc}\,@k = \frac{100}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \mathbf{1}\left[j_q^\star \in \mathcal{R}_q^{(k)} \right]. \tag{20}$$

Top-3 and Top-5 accuracy are evaluated on the same queries. A hit requires the exact asset identity; a category match or a geometrically similar asset does not count as a correct retrieval. These metrics evaluate the silhouette-based candidate ranking, rather than the independent RGB–CAD local-matching ranking or the identity selected after pose validation. Reference identities are used only for scoring and are not provided to candidate selection at inference time.

B.2 Model Details

We compare Stereo SGBM, MoGe-2 on the original transparent RGB image, LingBot-Depth v0.5 with SGBM depth, SeeClear with shared predicted vessel masks followed by MoGe-2, SeeClear with Trans4Trans-tiny segmentation followed by MoGe-2, and the full RATDE model with point-feature and local geometric conditioning followed by LingBot-Depth v0.5. All pipelines start from the original observations. For LingBot-Depth configurations, SGBM depth is computed from the original stereo pair; the backend receives the original left RGB in the direct baseline or the composited opacified left image in generation-based variants. We use matched preprocessing, masks, calibration, and backend settings where applicable, and report segmentation and input-modality differences. SeeClear uses its released checkpoint, whereas RATDE includes the adaptation stages described in Section 4. Experiments use the rendered laboratory dataset described in Section 3. The CAD memory includes the exact assets of all transparent targets, and retrieval and alignment are performed from the input observations.

B.3 Ablation details

We compare the original-RGB baseline, SeeClear, a random-CAD control, and the full RATDE model (Table 2). All variants use the same LingBot-Depth v0.5 backend, calibration, and consistency-filtered SGBM depth from the original stereo pair. SeeClear uses its released checkpoint without our image-only LoRA adaptation or CAD-conditioning modules. The random-CAD control shares the full model's trained weights and replaces candidate CAD identities in both the point-feature and local-geometry branches with random CADs from the memory. Point-feature inputs are constructed from the replacement CADs, and RGB–CAD matching and pose estimation are rerun for the local-geometry branch. Thus, random CAD versus retrieved CAD tests the effect of candidate selection at fixed weights, whereas SeeClear versus the full model measures the combined effects of adaptation and CAD conditioning.

C LabSense-Glass Dataset

We provide additional details on the construction, asset composition, rendering configuration, annotations, and organization of LabSense-Glass. The dataset is designed around a property of laboratory environments that is central to our method: many laboratory objects are standardized products whose 3D geometry can be stored in an external object memory. We therefore construct the dataset such that visual observations, geometric annotations, and source CAD assets remain explicitly associated throughout the data pipeline.

C.1 Dataset Scope and Composition

Our data resources consist of three complementary components: an existing transparent-object benchmark, laboratory scene and asset resources, and our paired rendered data. The broader data collection contains 222 scene entries across these components. Importantly, these 222 entries should not be interpreted as 222 distinct laboratory environments: 204 correspond to scenes from the existing ClearDepth benchmark, while the laboratory-specific portion is based on nine LabUtopia environments from which our paired data are rendered.

The dataset used to train and evaluate RATDE is LABSENSE-GLASS, comprising nine rendered laboratory environments and 1,336 stereo samples. Each stereo sample contains synchronized left and right observations. For each view, we additionally render an opaque counterpart while keeping scene geometry, object placement, camera parameters, and lighting configuration fixed. This results in 2,672 transparent RGB images and 2,672 corresponding opaque supervision images in the current experimental release.

Table 4 summarizes the role of each component.

Table 4: Composition of the data resources used in this work. ClearDepth is retained as an external transparent-object benchmark, while LabUtopia provides laboratory environments and source assets. LABSENSE-GLASS denotes our paired rendered dataset used for the proposed laboratory experiments.

Component	Scenes	Samples / Frames	Role
ClearDepth	204	15,634 frames	External benchmark
LabUtopia source environments	9	–	Scene and asset source
LABSENSE-GLASS	9	1,336 stereo samples	Paired laboratory data

C.2 LABORATORY ASSET LIBRARY

We construct a laboratory asset library containing **231 distinct 3D assets**. The library is broader than the set of transparent targets appearing in the rendered experiments and is intended to represent the semi-open asset space encountered in laboratory environments. It contains three groups: 130 vessel and labware assets, 63 laboratory equipment assets, and 38 additional glassware assets.

The 130 vessel and labware assets cover both conventional chemistry glassware and standardized experimental consumables. They include beakers, Erlenmeyer flasks, round-bottom flasks, volumetric flasks, graduated cylinders, test tubes, vials, condensers, reservoirs, tube racks, well plates, and related laboratory containers. These assets are collected from multiple existing asset sources, including 40 LabUtopia assets, 35 Opentrons labware models, 30 models from a laboratory glassware bundle, 23 existing laboratory vessel models, and two additional 3D resources.

The 63 equipment assets provide surrounding laboratory context and include centrifuges, incubators, heating devices, balances, biological safety cabinets, spectrometers, chromatographic instruments, sterilization equipment, pipettes, and other common laboratory devices. The additional glassware collection contains 38 geometrically distinct glass objects spanning vessel families such as carafes, drinking glasses, and other transparent containers.

Table 5: Composition of the 3D laboratory asset library. The full library contains 231 assets; only a subset appears as transparent targets in the rendered experiments.

Asset group	Number	Representative contents
Vessels and labware	130	Beakers, flasks, cylinders, vials, well plates
Laboratory equipment	63	Centrifuges, heaters, balances, analytical devices
Additional glassware	38	Carafes and other transparent vessel geometries
Total	**231**	

For each asset used by RATDE, we retain its source identity, mesh geometry, nominal dimensions when available, and coordinate metadata. A CAD-memory entry additionally stores sampled surface points and multi-view reference renderings. Source-to-canonical transformations are recorded so that normalized representations used by the network remain traceable to the original asset geometry. Query-scene placement is stored separately from canonical object geometry.

C.3 LABORATORY SCENE CONSTRUCTION

The paired dataset is rendered from nine laboratory environments covering five relatively simple laboratory layouts, two more cluttered layouts, and two navigation-oriented environments. The

source environments include chemistry workbenches, general laboratory spaces, and larger navigation scenes. Across these environments, variation in layout and camera placement produces changes in background appearance, object scale, clutter, and partial occlusion.

We compose scenes in Universal Scene Description (USD), keeping vessel geometry, scene placement, and material assignments independently controllable. Transparent materials are assigned to target vessel meshes, while liquids, caps, racks, and other non-glass components retain their corresponding materials. This separation is important for paired rendering because it allows the optical appearance of the target vessel to be changed without modifying its geometry or pose.

Fig. 4 shows examples from all nine rendered environments. The scenes exhibit substantial differences in workspace layout, illumination, background structure, vessel density, and camera viewpoint rather than repeatedly rendering a single tabletop configuration.

C.4 Paired Transparent–Opaque Rendering

A central property of LabSense-Glass is the availability of paired transparent and opaque observations. For every pair, the scene geometry, object transforms, camera intrinsics and extrinsics, and lighting configuration are kept fixed. We then modify the material of the target transparent vessel to obtain its opaque counterpart. Consequently, corresponding pixels describe the same underlying scene geometry while exposing different material appearances.

The transparent rendering serves as the model observation, whereas the opaque rendering provides supervision for generative opacification. We emphasize that material replacement can also modify shadows, reflections, and indirect illumination. Therefore, transparent and opaque images are not assumed to be pixel-identical outside the vessel region, despite sharing the same scene geometry and camera configuration.

Rendering is performed in NVIDIA Isaac Sim using RTX Path Tracing. Images are rendered at 1280×720 resolution using a virtual ZED 2i stereo configuration. The stereo baseline is 0.12 m, with a nominal focal length of 2.1 mm. The current rendering configuration uses 256 samples per pixel. Camera viewpoints are positioned approximately 0.4–0.7 m from the target workspace with a downward viewing angle of approximately $27.5°$–$30°$.

Table 6: Rendering and camera configuration for LabSense-Glass.

Field	Configuration
Simulator	NVIDIA Isaac Sim
Renderer	RTX Path Tracing
Image resolution	1280×720
Camera model	Virtual ZED 2i stereo camera
Stereo baseline	0.12 m
Nominal focal length	2.1 mm
Camera intrinsics	$f_x = f_y = 529.8$ px; $(c_x, c_y) = (638.5, 358.5)$
Path-tracing samples	256 spp
Camera–target distance	approximately 0.4–0.7 m
Camera downward angle	approximately $27.5°$–$30°$

C.5 Geometric Annotations

In addition to paired RGB observations, we record geometric information required for depth estimation, correspondence learning, and CAD alignment. The annotation set includes stereo camera calibration, camera poses, depth, surface normals, visible instance labels, object identities, and scene metadata. Fig. 4 illustrates representative transparent–opaque pairs together with their geometric annotations.

Depth follows the camera-axis convention and is expressed in metres. Finite positive values are considered valid, while zero denotes invalid depth. Surface normals are represented in world coordinates. Instance masks correspond to visible object regions and do not perform amodal completion. Shared object identifiers are retained across the left and right views so that stereo observations, geometric annotations, and CAD records can be associated with the same physical instance.

We additionally retain object-to-world transformations for rendered CAD instances. These transformations are used to construct training correspondences between visible image regions and points on the corresponding CAD surface. Ground-truth poses are used only for supervision and dataset analysis; RATDE does not receive query-specific ground-truth object identity or pose during inference.

For transparent-object evaluation, liquid pixels are treated separately from vessel surfaces and are excluded from the transparent-vessel depth region used by our reported metrics. This avoids conflating reconstruction of the vessel wall with the geometry of its contents.

C.6 CAD Association and Object Memory

Unlike transparent-object datasets in which object meshes are used only as evaluation annotations, LabSense-Glass explicitly preserves the association between an observed object and an external CAD memory. This association reflects our intended laboratory setting, where standardized instruments can be registered in an inventory before deployment.

Each memory entry contains the source mesh, sampled point cloud, reference renderings, canonical-coordinate metadata, and nominal asset dimensions when available. For the transparent targets used in our experiments, the exact source CAD is present somewhere in the memory, but its identity and query-specific pose are not provided to the model at inference time. The model must instead retrieve candidate assets and establish RGB–CAD correspondences from the observation.

This organization also permits controlled retrieval experiments. Because the reference asset identity is known to the dataset but hidden from candidate selection, we can measure exact-asset Recall@K and directly replace retrieved candidates with random CAD assets while keeping model weights fixed. The resulting protocol separates the availability of an object memory from the correctness of the retrieval mechanism.

C.7 Data Organization and Reproducibility

Samples are indexed by scene and stereo-sample identity. Each index links the left and right RGB observations, their transparent–opaque counterparts, camera calibration, geometric annotations, visible instance identities, and source asset records where applicable. This organization preserves correspondence across visual observations, geometric supervision, and the CAD memory.

For reproducibility, we plan to release the nine composed USD laboratory scenes together with the dependencies required to reconstruct them, paired renderings, geometric annotations, camera metadata, and sample manifests. The editable USD representation allows researchers to modify object placement, materials, illumination, and camera configuration, or to construct additional layouts using the same laboratory assets. The complete upstream source library is not redistributed when its original licensing or distribution conditions do not permit inclusion; instead, provenance and asset associations are retained where applicable.