Phys-Liquid: A Physics-Informed Dataset for Estimating 3D Geometry and Volume of Transparent Deformable Liquids

Ke Ma[1], Yizhou Fang[2], Jean-Baptiste Weibel[3], Shuai Tan[4], Xinggang Wang[5], Yang Xiao[6], Yi Fang[7,8], Tian Xia[2*]

[1]School of Artificial Intelligence and Automation, Huazhong University of Science and Technology
[2]School of Software and Engineering, Huazhong University of Science and Technology
[3]Human-Centered AI Lab, Institute of Forest Engineering, BOKU University
[4]Shanghai Jiao Tong University
[5]School of Electronic Information and Communications, Huazhong University of Science and Technology
[6]National Key Laboratory of Multispectral Information Intelligent Processing Technology, School of Artificial Intelligence and Automation, Huazhong University of Science and Technology
[7]Center for Artificial Intelligence and Robotics, New York University Abu Dhabi
[8]Embodied AI and Robotics (AIR) Lab, New York University
make@hust.edu.cn, {fangyizhou, xgwang, Yang_Xiao}@hust.edu.cn, jb.weibel@gmail.com, tanshuai0219@stju.edu.cn, yfang@nyu.edu, tianxia@hust.edu.cn

Abstract

Estimating the geometric and volumetric properties of transparent deformable liquids is challenging due to optical complexities and dynamic surface deformations induced by container movements. Autonomous robots performing precise liquid manipulation tasks—such as dispensing, aspiration, and mixing—must handle containers in ways that inevitably induce these deformations, complicating accurate liquid state assessment. Current datasets lack comprehensive physics-informed simulation data representing realistic liquid behaviors under diverse dynamic scenarios. To bridge this gap, we introduce Phys-Liquid, a physics-informed dataset comprising 97,200 simulation images and corresponding 3D meshes, capturing liquid dynamics across multiple laboratory scenes, lighting conditions, liquid colors, and container rotations. To validate the realism and effectiveness of Phys-Liquid, we propose a four-stage reconstruction and estimation pipeline involving liquid segmentation, multi-view mask generation, 3D mesh reconstruction, and real-world scaling. Experimental results demonstrate improved accuracy and consistency in reconstructing liquid geometry and volume, outperforming existing benchmarks. The dataset and associated validation methods facilitate future advancements in transparent liquid perception tasks.

Code and Datasets — https://dualtransparency.github.io/Phys-Liquid/

Extended version — https://arxiv.org/abs/2511.11077

Introduction

Understanding the geometric and volumetric properties of transparent deformable liquids is essential for embodied robots (Liu et al. 2024b) operating in autonomous laboratory environments. As robotics and large multimodal models (Radford et al. 2021; Li et al. 2023; Driess et al. 2023) advance, autonomous robots are increasingly tasked with handling complex procedures in biomedical and biochemical experiments (Dai et al. 2024; Szymanski et al. 2023; Triantafyllidis et al. 2023; Xie et al. 2023; Boiko et al. 2023; Burger et al. 2020). Precise liquid manipulation tasks-such as dispensing, aspiration, and mixing-require these systems to accurately perceive dynamic liquid deformation caused inevitably by container movements. Without robust and realistic simulation datasets that represent the complexities of deformable liquids, accurately assessing these liquids remains a significant challenge, potentially causing experimental failures and resource losses.

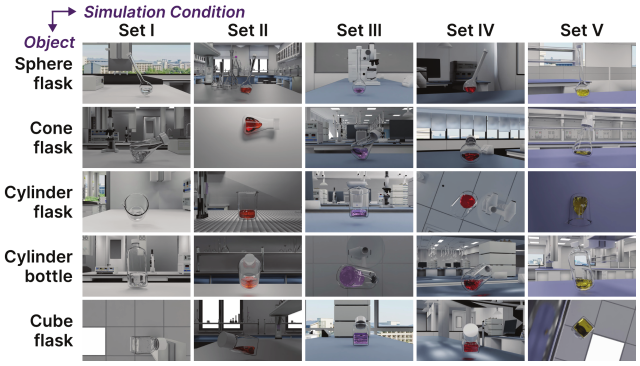


Figure 1: Simulation samples of five transparent containers under five distinct simulation sets from Phys-Liquid, showing variations in laboratory scenes, lighting conditions, container rotations, liquid colors, volumes, and deformations.

Current datasets are insufficient for addressing the complex challenges posed by transparent deformable liquids in dynamic laboratory environments. Existing large-scale 3D datasets like Objaverse (Deitke et al. 2023) primarily include rigid and opaque objects, with limited representation of liquid-filled transparent containers. Datasets such as Clear-

*Corresponding author.

Grasp (Sajjan et al. 2020), ClearPose (Chen et al. 2022), and TODD (Fang et al. 2022) focus on transparent object perception but neglect the contained liquids. Although datasets like DTLD (Wang et al. 2024) and the dataset by Narasimhan et al. (Narasimhan et al. 2022) include transparent liquids, they either provide static liquid states or limited views without simulating realistic deformation dynamics induced by container rotations. This lack of dynamic realism restricts the development of robust algorithms for precise robotic manipulation in real-world scenarios.

To address these limitations, we introduce Phys-Liquid, a physics-informed simulation dataset specifically designed to capture the dynamic behaviors of transparent deformable liquids under realistic laboratory conditions. Unlike previous datasets, Phys-Liquid systematically simulates liquid deformations induced by container rotations based on precise physical modeling governed by Navier-Stokes equations (Chorin 1968). We proposed a dataset generation workflow using Blender (Blender 2018) that includes defining diverse laboratory scenes, lighting conditions, liquid properties, and systematically rotating containers to induce temporal liquid deformations, subsequently rendered from multiple orthographic viewpoints. By comparing simulated liquid behaviors against corresponding real-world captures, we further validate the realism and fidelity of our simulation data. Our dataset comprises 97,200 simulation images generated from 20 common laboratory containers, encompassing diverse laboratory scenes, liquid colors, lighting conditions, and rotation modes. The dataset includes annotated 3D liquid meshes, providing precise geometric and volumetric information for rigorous validation and benchmarking.

A key aspect of validating the realism and effectiveness of a physics-informed dataset is to demonstrate its utility in enabling accurate predictions of liquid geometry and volume from visual inputs. To this end, we developed a four-stage pipeline that leverages Phys-Liquid for reconstructing and estimating liquid geometry from single images. This pipeline includes segmentation of liquid regions, generation of multi-view liquid masks via diffusion models (Wang et al. 2025), 3D mesh reconstruction using triplane methods (Wang et al. 2025), and scaling reconstructed meshes to match real-world dimensions. By comparing the reconstructed liquid states against ground truth simulation meshes, we quantitatively assess the dataset's realism and applicability to real-world manipulation tasks.

Experimental validation demonstrates that Phys-Liquid significantly improves the accuracy and consistency of liquid geometry reconstruction compared to existing benchmarks. Moreover, evaluation on real-world datasets confirms that models trained on Phys-Liquid exhibit strong generalization capabilities. These results underscore the dataset's realistic representation of dynamic liquid behaviors and highlight its potential to substantially enhance future robotic manipulation and liquid perception research.

Our contributions are summarized as follows:

- We introduce Phys-Liquid, a physics-informed dataset addressing gaps in current liquid datasets by capturing realistic dynamic deformations, providing a foundational resource for future research in transparent liquid perception and precise liquid handling applications.
- We develop a four-stage reconstruction pipeline to validate the dataset's effectiveness, demonstrating superior performance over existing benchmarks on both simulation and real-world liquid datasets.

Related Work

Liquid Datasets in Laboratory Scenes

Existing large-scale 3D datasets like Objaverse (Deitke et al. 2023) include various chemistry lab glassware; however, most of these objects are typically empty and do not involve liquids. Transparent object datasets like TOD (Liu et al. 2020), TODD (Xu et al. 2021), TransCG (Fang et al. 2022), ClearGrasp (Sajjan et al. 2020), StereOBJ-1M (Liu, Iwase, and Kitani 2021), and ClearPose (Chen et al. 2022) mainly focus on 6D pose estimation for laboratory objects but generally do not represent liquids inside these containers. Gautham et al. (Narasimhan et al. 2022) introduced a dataset capturing transparent chemistry flasks with transparent liquids from a single camera viewpoint, lacking representation of dynamic liquid deformations. The DTLD dataset proposed by Wang et al. (Wang et al. 2024) comprises 27,458 images depicting liquids in four biomedical flasks captured under multi-view laboratory settings; however, the dataset only includes static liquid states without deformation dynamics induced by container movements. Thus, these datasets do not sufficiently address the dynamic characteristics crucial for precise liquid perception in realistic robotic manipulation scenarios.

Physics-Informed Liquid Simulation Datasets

Physics-informed methods have been increasingly employed in simulating realistic deformable or fluid behaviors for various computer vision tasks (Banerjee et al. 2024; Lin et al. 2025; Tang et al. 2025; Yan et al. 2023; Li et al. 2025; Wu et al. 2025), leveraging fundamental physical principles to enhance dataset authenticity. Recent liquid simulation studies have been explored across diverse applications, including robotic manipulation (Lin, Fu, and Xue 2023; Eppel et al. 2022; Moya et al. 2023; Qian et al. 2024), river flow velocimetry evaluation (Bodart et al. 2022), liquid volume estimation (Liu et al. 2023a), surface detection (Richter, Orosco, and Yip 2022), and temporal prediction (Wiewel, Becher, and Thuerey 2019). Various liquid datasets have been generated using different simulation pipelines: for instance, TransProteus (Eppel et al. 2022) simulates liquid shapes inside transparent vessels using Blender (Blender 2018) Mantaflow module (Thuerey and Pfaff 2016). Richter et al. (Richter, Orosco, and Yip 2022) simulated a fountain liquid dataset with Blender (Blender 2018) to reconstruct liquids from surface detections. However, existing physics-informed datasets generally do not simulate dynamic liquid deformations within transparent containers induced by physical manipulations, thus leaving a gap in realistic representation of laboratory liquid behaviors.

Liquid Reconstruction Methods from Images

Several methods have been proposed for reconstructing liquids from visual data to enable downstream estimation tasks. Eppel et al. (Eppel et al. 2022) reconstructed 3D liquid shapes from RGB images by predicting XYZ maps, though facing limitations in scenes lacking precise depth data for transparent liquids. Recent studies in single-view 3D reconstruction (Chan et al. 2022; Cheng et al. 2023; Gupta et al. 2023; Vahdat et al. 2022; Zheng et al. 2023; Liu et al. 2024a; Wu et al. 2024; Xu et al. 2024; Wang et al. 2025) have demonstrated high-quality mesh generation from single images, showing potential for liquid reconstruction tasks.

Physics-Informed Dataset

We present Phys-Liquid, a physics-informed dataset capturing the dynamic deformation of transparent liquids within transparent containers under systematic rotational motions in laboratory scenes. Figure 1 presents examples of five containers under comprehensive simulation conditions.

Physics-Informed Liquid Simulation Method

To simulate realistic liquid behaviors, we leverage the Navier-Stokes equations (Chorin 1968), which accurately describe the fundamental physics of fluid dynamics by accounting for velocity, pressure, viscosity, and external forces. Specifically, the fluid flow in our simulation is governed by the momentum equation:

$$\frac{D\mathbf{u}}{Dt} = -\frac{1}{\rho}\nabla p + \nu\nabla^2\mathbf{u} + \mathbf{g}, \tag{1}$$

where $\mathbf{u}$ represents the fluid's velocity field, indicating the direction and speed of each particle, ρ denotes fluid density, p is the pressure field, dictating flow based on high and low-pressure regions, ν is the kinematic viscosity that models internal friction within the fluid, affecting how smoothly it flows and $\mathbf{g}$ represents external forces that impact fluid behavior. This equation describes the fluid acceleration $\frac{D\mathbf{u}}{Dt}$ as a result of internal pressure forces $\left(-\frac{1}{\rho}\nabla p\right)$, viscous forces $(\nu\nabla^2\mathbf{u})$, and external forces $\mathbf{g}$. Additionally, we enforce the incompressibility condition to ensure constant fluid density and realistic liquid behavior:

$$\nabla \cdot \mathbf{u} = 0 \tag{2}$$

This constraint ensures that fluid density remains constant, maintaining the divergence-free nature of the velocity field. It enforces an accurate fluid behavior by preventing artificial compressions or expansions. We solve these equations using Mantaflow (Thuerey and Pfaff 2016), integrated within Blender (Blender 2018), enabling accurate simulation of liquid dynamics under various container rotations.

Data Generation Pipeline

Scene Creation Our data generation workflow begins with constructing diverse laboratory scenes, including realistic lab environments, lighting variations, commonly-used laboratory containers, liquid colors, and controlled camera setups. Specifically, we design five realistic laboratory scenes containing experimental setups, lab equipment, and workbenches. We simulate eight indoor lighting conditions varying in intensity and orientation. We select twenty common transparent laboratory containers covering a variety of shapes (cubes, cylinders, cones, spheres, composite forms) and sizes, each with distinct materials and textures. Transparent liquids are simulated in five typical colors and various initial volumes. Six orthographic virtual cameras (top, bottom, front, back, left, right) are positioned to capture the liquid deformations and container rotations from multiple viewpoints, illustrated in Figure 2. Each simulation set randomly selects configurations, assigning five distinct conditions per container, resulting in 100 unique combinations.

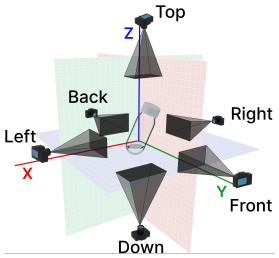


Figure 2: The setting of six orthographic camera views for the representation of triplane in the scene setting.

Object Rotation Over Time Containers undergo rotational motions to induce liquid deformation. Rotation patterns are defined along the X, Y, and Z axes, creating six distinct combinations (excluding rotations solely around the Z-axis), with rotations ranging from 0° to 80°. We record 81 time frames for continuous motion, with each frame corresponding to a specific object pose and liquid deformation.

Liquid Simulation The liquid is represented by a particle system enclosed within a mesh defining the liquid's surface illustrated in Figure 3. Key parameters controlling the simulation include mesh resolution, particle radius, particle limits, and the FLIP (Brackbill, Kothe, and Ruppel 1988) ratio. As the container rotates, liquid particles interact with the inner surfaces, undergoing collisions governed by the Navier-Stokes equations (Chorin 1968). This results in realistic liquid deformations and varying surface meshes influenced by container geometry and motion.

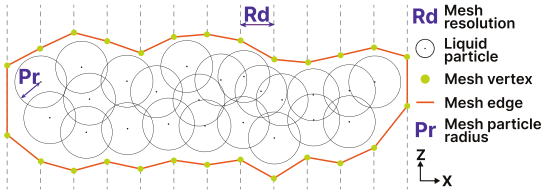
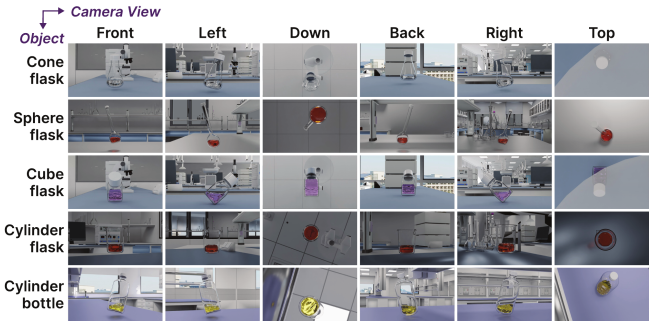
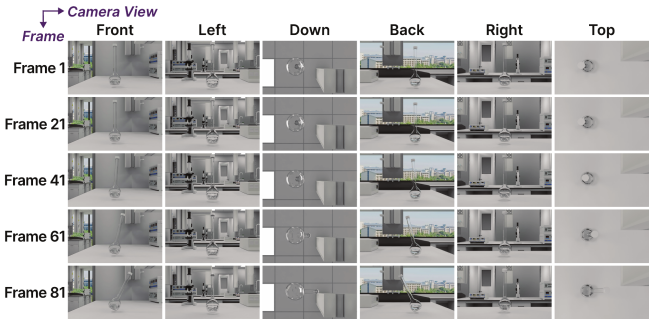


Figure 3: The liquid mesh formed by particles.

Multi-View Rendering We render six orthographic views using virtual cameras to capture perspectives that are challenging to achieve in real-world settings. Images are rendered every frame over the sequence of 81 frames capturing



(a) Image samples of different objects captured simultaneously from six orthographic camera views at the same time frame.



(b) Image samples of the same object captured from six orthographic camera views across multiple time frames.

Figure 4: Multi-view and temporal representations.

continuous motion. For each time frame, we generate scene images (container, liquid, background), liquid-only masks, and liquid meshes saved in OBJ format with calibrated real-world dimensions. Figure 4a depicts multi-view images for different containers at one time frame. Figure 4b shows temporal variation at 20-frame intervals across multiple views for a single container.

Comparison with Real-World Captures

To further validate the dataset realism, we performed comparative experiments between simulated and real-world liquid behaviors. For ten representative containers from the Phys-Liquid dataset, we recorded liquid deformations in physical experiments under comparable rotational conditions. We analyzed visual and physical properties, such as liquid deformation, flow patterns, and interactions with container surfaces. In Figure 5, the angles between the liquid's top surface and container side walls are annotated in red for simulation data and green for real-world data regarding two objects. These highlighted angle pairs at each rotational pose closely align, demonstrating that our simulations approximate real-world liquid deformation behaviors.

Dataset Analysis and Visualization

Phys-Liquid comprises 97,200 simulation images (scene images and liquid masks) across 100 simulation configurations, 8,100 annotated OBJ-format liquid meshes with real-world size, and CAD models of 20 laboratory containers.

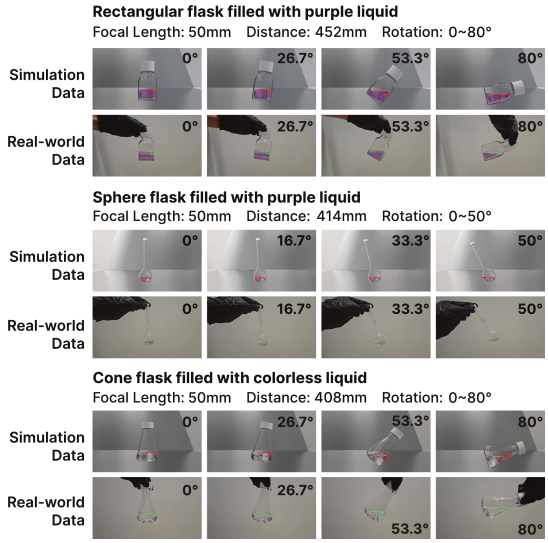


Figure 5: Validation of simulation realism by comparing liquid deformations with real-world experiments.

A detailed analysis of the simulation factors is provided in Table 1, where we also compare Phys-Liquid with existing datasets collected in laboratory scenes. The distribution of image samples across conditional settings are visualized in Figure 6. In detail, the five liquid colors are colorless, purple, red, orange, and yellow. The eight lighting conditions represent typical indoor illumination settings, encompassing variations in light intensity, direction, color, and mixing methods, labeled as L1 to L8. Three laboratory scenes are designated as Lab1 to Lab5 and six rotation modes are labeled from R1 to R6 (details in the appendix).

Each dataset image includes extensive metadata: container (CAD model, material, transparency), camera viewpoint (orthographic view, distance, focus), liquid properties (color, volume, mask, mesh), environmental settings (lighting, tabletop textures), physical rotation information (rotation angles, rotation mode), and image resolution.

Liquid Reconstruction Pipeline

Task Formulation

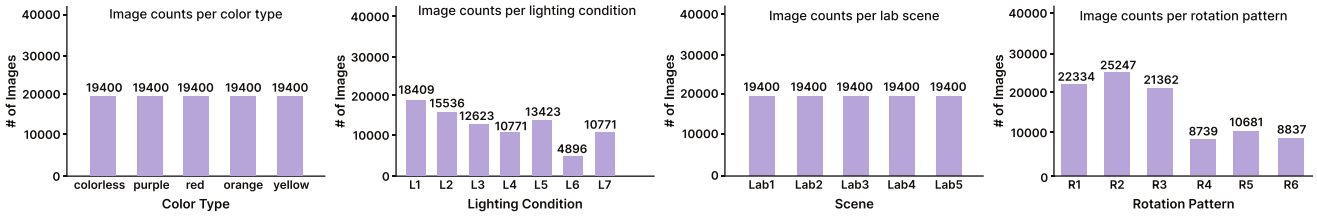
Our goal is to reconstruct the 3D mesh S of deformable, transparent liquids in real-world dimensions from a single input image I, formulated as:

$$S = F(I) = T(R(G(S(I))), s) \tag{3}$$

We first extract a high-confidence liquid mask $M = S(I)$ from the image I using segmentation; then, we generate six orthographic liquid masks $\{M_i\}_{i=1,2,\dots,6} = G(M)$ via a diffusion model to simulate multiple views. Next, we reconstruct the 3D mesh $V = R(\{M_i\}_{i=1,2,\dots,6})$ by integrating the multi-view masks using a triplane method; finally, we scale the reconstructed mesh to real-world dimensions $S = T(V, s)$, where s is a scaling factor. The pipeline framework is illustrated in Figure 7.

Dataset	# objects	# scenes	# images	# liquid color	# lighting condition	liquid deformation	multiple viewpoints	temporal changes	continuous volume
Gautham et al.	2	1	4,601	2	1	×	×	×	✓
DTLD	4	3	27,458	5	7	×	✓	×	✓
Phys-Liquid(Ours)	20	5	97,200	5	8	✓	✓	✓	✓

Table 1: The analysis of Phys-Liquid dataset compared with related liquid datasets.



(a) Liquid color distribution. (b) Lighting distribution. (c) Scene distribution. (d) Rotation mode distribution.

Figure 6: Data visualization of Phys-Liquid.

Liquid Segmentation

We use SAM2 (Ravi et al. 2024) to segment liquid regions from the input RGB image. To enhance segmentation accuracy, we integrate YOLO-world (Cheng et al. 2024), a real-time detector, by specifying "liquid" and "colored liquid" as positive classes and explicitly excluding the "bottle" class. A high-confidence bounding box from YOLO-world guides SAM2, resulting in precise liquid masks. As shown in the input module of Figure 7, the pipeline can also take multi-view images as input where segmented masks from these images are fed to the multi-view generation module.

Multi-View Liquid Mask Generation

Generating accurate multi-view liquid masks is challenging for pre-trained multi-view diffusion models due to complex physics-driven deformations. We utilize the multi-view diffusion model CRM (Wang et al. 2025), fine-tuned on our Phys-Liquid dataset. Our approach leverages the dataset's six orthographic views per timestep during fine-tuning. Canonical coordinate maps (CCMs) generated alongside these masks capture spatial consistency and deformation details. During generation, the fine-tuned model uses single or multi-view masks from the segmentation module to predict and complete masks from other views.

3D Mesh Reconstruction

We reconstruct the 3D liquid mesh from multi-view masks and CCMs using the convolutional reconstruction model from CRM (Wang et al. 2025). This model employs a triplane representation, aggregating spatial features from three orthogonal planes (xy, xz, yz). A convolutional U-Net encodes multi-view masks into triplane features, which multilayer perceptrons (MLPs) (Rumelhart, Hinton, and Williams 1986) subsequently decode into a textured 3D mesh. Due to the high fidelity of the input multi-view images, no additional fine-tuning on Phys-Liquid is necessary.

Scaling to Real-World Dimensions

To align the reconstructed 3D mesh with real-world dimensions, we develop a mesh scaling model that enables the pipeline to finally estimate the geometric and volumetric properties of transparent deformable liquids in laboratory settings. The scaling model utilizes a multi-view Vision Transformer (ViT) (Dosovitskiy 2020) architecture depicted in Figure 7. Six orthographic views are encoded separately via the ViT backbone, then integrated using positional encoding and transformer encoder layers. The combined features are processed by an MLP(Rumelhart, Hinton, and Williams 1986) to regress a scaling factor s. The model is supervised using an L2 loss, with ground-truth scaling factors provided by Phys-Liquid. s is calculated by comparing reconstructed dimensions with the real-world dimensions from Phys-Liquid S_{PI} along the x, y, and z axes:

$$s = \sqrt[3]{\frac{S_{\text{PI},x}}{V_x} \cdot \frac{S_{\text{PI},y}}{V_y} \cdot \frac{S_{\text{PI},z}}{V_z}} \tag{4}$$

Experiment

We performed one experiment comparing our method with baselines, four experiments evaluating the pipeline on real-world generalization, multi-view consistency, fine-tuning impact, and temporal consistency, and one ablation study on pipeline modules. The multi-view diffusion model was fine-tuned on two RTX 6000 Ada 48GB GPUs for 16 hours over 10k iterations and the scaling model trained on one RTX 6000 Ada 48GB GPU for 12 hours across 500 iterations.

Dataset

We used the Phys-Liquid dataset consisting of 100 separate temporal sequences, splitting it into training and testing sets (9:1) by entire sequences. Each sequence contains 81 timesteps with multi-view images. Entire sequences (all 81 timesteps within one sequence) are strictly assigned to either the training or testing set, ensuring no overlap between consecutive frames from the same sequence across sets.

Evaluation Metrics

For 2D mask accuracy, we report Intersection over Union (IoU). To assess 3D mesh similarity, we use Chamfer Distance (CD), Volumn IoU, and F-Score (following One-2-3-45 (Liu et al. 2023b)). Additionally, we compute the Root

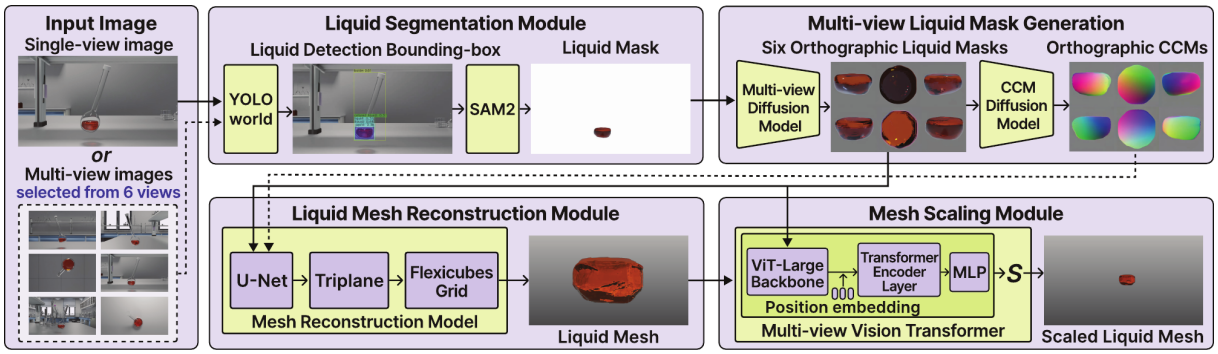


Figure 7: Overview of our four-step pipeline for reconstructing and scaling 3D meshes of deformable liquids. The input module accepts single or multi-view images, conditioned on the laboratory setup. Output examples for each step are shown.

Mean Squared Error (RMSE) for the reconstructed mesh's length, width, and height, and the Mean Absolute Percentage Error (MAPE) for the estimated scaling factor. Precise metric definitions are provided in the appendix.

Comparisons with Reconstruction Baselines

Qualitative Results We compared the reconstruction network in our pipeline with baseline methods, including InstantMesh (Xu et al. 2024) and TripoSR (Tochilkin et al. 2024). We show two selected example cases in the test set in Figure 8. We included six orthographic simulation masks from Phys-Liquid for reference, comparing these to outputs from baseline methods and our pipeline with and without fine-tuning multi-view diffusion model. The baseline methods show visual discrepancies from simulation results and fail to capture precise morphological features of liquid deformation, demonstrating the higher fidelity of our method.

Quantitative Results We quantitatively compared our method with InstantMesh (Xu et al. 2024) and TripoSR (Tochilkin et al. 2024) using 50 randomly selected test images. Meshes were scaled to a unified bounding box for fair evaluation. Results using Chamfer Distance, Volume IoU, and F-Score are shown in Table 2. Our method outperformed two baselines, with higher reconstruction accuracy.

Method	Chamfer Distance	Volume IoU	F-Score (%)
InstantMesh (Xu et al. 2024)	0.0189	0.2794	46.18
TripoSR (Tochilkin et al. 2024)	0.0275	0.2275	38.06
Our method without fine-tuning	0.0128	0.3246	58.19
Our method with fine-tuning	**0.0085**	**0.6236**	**78.57**

Table 2: Quantitative reconstruction comparison with baseline methods (threshold = 0.005) using **50** test images.

Generalization to Real-World Data

To evaluate the applicability and scalability of our method beyond simulated laboratory environments, we tested our model trained exclusively on the Phys-Liquid dataset directly on the DTLD dataset (Wang et al. 2024), a real-world dataset featuring static liquid states without deformation.

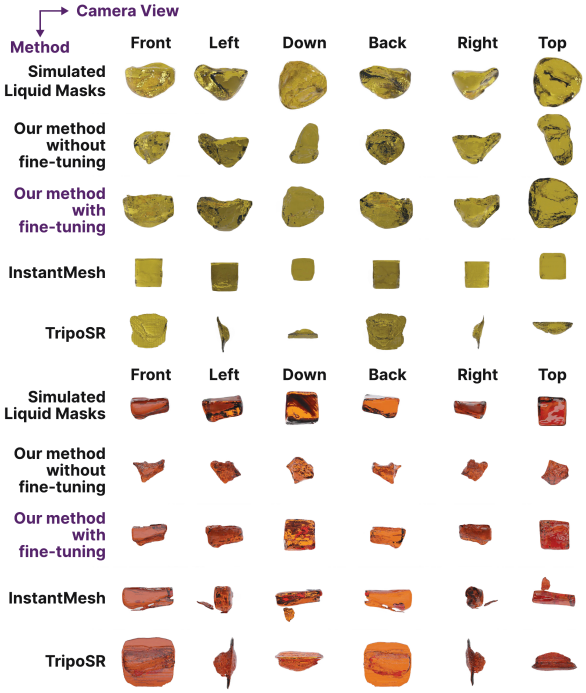


Figure 8: Qualitative comparison of reconstructed meshes by baseline methods and our pipeline with and without fine-tuning against physics-informed simulation meshes.

Performance was measured and presented in Table 3. Although DTLD (Wang et al. 2024) lacks liquid deformation, the results indicate that our method maintains reasonable accuracy in real-world scenarios, indicating that the physical priors and visual characteristics in Phys-Liquid are closely aligned with real-world conditions to support transferability.

Multi-View Consistency of Mask Generation

We evaluated fine-tuning the diffusion model on Phys-Liquid dataset for multi-view mask generation. Performance was assessed on the test set by comparing generated and simulated masks using IoU and visual inspection. Fine-tuning improved average IoU from 74.38% to 90.05%,

Dataset	RMSE	Chamfer Distance	Volume IoU	F-Score (%)
DTLD dataset	0.0266	0.0172	0.3861	62.43
Phys-Liquid Test	0.0192	0.0079	0.4748	75.38

Table 3: Quantitative comparison of reconstruction performance on DTLD and Phys-Liquid datasets on **entire** test set.

yielding masks with more realistic boundaries. Figure 9 presents a visual comparison, validating the dataset's effectiveness in improving generation precision.

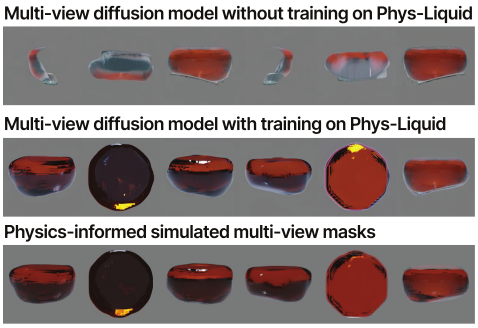


Figure 9: Multi-view liquid masks before and after diffusion model fine-tuning, compared with physics-informed masks.

We further evaluated multi-view consistency using different single-view inputs. We computed the average IoU between predicted multi-view masks and simulated masks across the test set for each camera view (Table 4). Results confirm high consistency across multi-view predictions.

Camera View	Front	Left	Down	Back	Right	Top
Average IoU (%)	91.76	90.16	89.32	91.03	90.82	89.21

Table 4: Average IoU between generated and physics-informed simulation masks across camera views.

Diffusion Model Fine-Tuning Impact

We evaluated the impact of diffusion model fine-tuning on the overall 3D reconstruction quality. We generated multi-view masks before and after fine-tuning and reconstructed meshes using our pipeline. Reconstructions were quantitatively assessed. As shown in Table 5, the fine-tuned model achieved improvement across all metrics, reducing the RMSE from 2.54% to 1.92% on the test set. The results confirm that fine-tuning enhances reconstruction accuracy.

Temporal Consistency of Reconstruction

We evaluated the temporal consistency of our pipeline in reconstructing liquid meshes across sequential frames from the test set of the Phys-Liquid dataset. We computed the RMSE of reconstructed meshes at each timestep within individual temporal sequences, and then calculated the variance and standard deviation of RMSE values for each sequence to quantify reconstruction stability. Over 100 temporal sequences, the average variance and standard deviation of RMSE were 0.00038038 and 0.00643858. These low values indicate that our pipeline maintains high temporal consistency, capturing stable and accurate liquid deformation states across consecutive frames.

Method	RMSE	Chamfer Distance	Volume IoU	F-Score (%)
without fine-tuning (test)	0.0254	0.0139	0.2850	46.19
fine-tuning (training)	0.0170	0.0083	0.4514	72.33
fine-tuning (test)	0.0192	0.0079	0.4748	75.38

Table 5: Reconstruction quality comparison before and after fine-tuning the diffusion model (threshold = 0.005).

Ablation Study on Pipeline Modules

We evaluated the contribution of each module in our pipeline by sequentially replacing their outputs with simulation results on the Phys-Liquid test set. Results of reconstruction quality are presented in Table 6, highlighting that mesh reconstruction and mesh scaling modules influence more on reconstruction performance.

Module Replaced	RMSE	Chamfer Distance	Volume IoU	F-Score (%)
Segmentation	0.0130	0.0075	0.5504	78.42
Multi-view Mask Generation	0.0105	0.0067	0.6532	81.36
Mesh Reconstruction	0.0085	0.0058	0.7687	85.64
Mesh Scaling	0.0071	0.0042	0.7511	88.47
Entire pipeline (test)	0.0192	0.0079	0.4748	75.38

Table 6: Ablation study evaluating contributions of pipeline modules by substituting their outputs with simulation data.

Conclusions and Limitations

We introduced Phys-Liquid, a physics-informed dataset designed specifically to support research on transparent deformable liquid perception. By accurately modeling liquid dynamics under realistic laboratory conditions, our dataset addresses the critical gap left by existing static or simplified datasets. In particular, it captures 3D deformations of liquids under temporally evolving container rotations, expanding the problem space from 3D to spatiotemporal (4D) domain and enabling the study of dynamic fluid perception.

While modest in size, the dataset emphasizes physically grounded diversity, with variations in lighting, liquid types, container shapes, and motion profiles. A reconstruction pipeline demonstrates the dataset's validity, supported by evaluations on both synthetic and real-world benchmarks.

Built within Blender, Phys-Liquid also enables the rendering of additional optical modalities—such as surface normals and refractive flow—that are difficult to obtain in real settings. This flexibility makes Phys-Liquid not only a benchmark, but also a versatile tool for training and evaluating physically grounded perception systems. We believe it provides a foundation for further research into multi-modal fluid representation and physics-aware visual reasoning.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 62525604, "Cell Therapy Designed by Artificial Intelligence". This work was also supported by the National Natural Science Foundation of China under Grant 62271221.

References

Banerjee, C.; Nguyen, K.; Fookes, C.; and George, K. 2024. Physics-informed computer vision: A review and perspectives. *ACM Computing Surveys*, 57(1): 1–38.

Blender, O. 2018. Blender—A 3D modelling and rendering package. *Retrieved. represents the sequence of Constructs1 to*, 4.

Bodart, G.; Le Coz, J.; Jodeau, M.; and Hauet, A. 2022. Synthetic river flow videos for evaluating image-based velocimetry methods. *Water Resources Research*, 58(12): e2022WR032251.

Boiko, D. A.; MacKnight, R.; Kline, B.; and Gomes, G. 2023. Autonomous chemical research with large language models. *Nature*, 624(7992): 570–578.

Brackbill, J. U.; Kothe, D. B.; and Ruppel, H. M. 1988. FLIP: a low-dissipation, particle-in-cell method for fluid flow. *Computer Physics Communications*, 48(1): 25–38.

Burger, B.; Maffettone, P. M.; Gusev, V. V.; Aitchison, C. M.; Bai, Y.; Wang, X.; Li, X.; Alston, B. M.; Li, B.; Clowes, R.; et al. 2020. A mobile robotic chemist. *Nature*, 583(7815): 237–241.

Chan, E. R.; Lin, C. Z.; Chan, M. A.; Nagano, K.; Pan, B.; De Mello, S.; Gallo, O.; Guibas, L. J.; Tremblay, J.; Khamis, S.; et al. 2022. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16123–16133.

Chen, X.; Zhang, H.; Yu, Z.; Opipari, A.; and Chadwicke Jenkins, O. 2022. Clearpose: Large-scale transparent object dataset and benchmark. In *European conference on computer vision*, 381–396. Springer.

Cheng, T.; Song, L.; Ge, Y.; Liu, W.; Wang, X.; and Shan, Y. 2024. Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16901–16911.

Cheng, Y.-C.; Lee, H.-Y.; Tulyakov, S.; Schwing, A. G.; and Gui, L.-Y. 2023. Sdfusion: Multimodal 3d shape completion, reconstruction, and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4456–4465.

Chorin, A. J. 1968. Numerical solution of the Navier-Stokes equations. *Mathematics of computation*, 22(104): 745–762.

Dai, T.; Vijayakrishnan, S.; Szczypiński, F. T.; Ayme, J.-F.; Simaei, E.; Fellowes, T.; Clowes, R.; Kotopanov, L.; Shields, C. E.; Zhou, Z.; et al. 2024. Autonomous mobile robots for exploratory synthetic chemistry. *Nature*, 1–8.

Deitke, M.; Schwenk, D.; Salvador, J.; Weihs, L.; Michel, O.; VanderBilt, E.; Schmidt, L.; Ehsani, K.; Kembhavi, A.; and Farhadi, A. 2023. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13142–13153.

Dosovitskiy, A. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Driess, D.; Xia, F.; Sajjadi, M. S.; Lynch, C.; Chowdhery, A.; Ichter, B.; Wahid, A.; Tompson, J.; Vuong, Q.; Yu, T.; et al. 2023. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*.

Eppel, S.; Xu, H.; Wang, Y. R.; and Aspuru-Guzik, A. 2022. Predicting 3D shapes, masks, and properties of materials inside transparent containers, using the TransProteus CGI dataset. *Digital Discovery*, 1(1): 45–60.

Fang, H.; Fang, H.-S.; Xu, S.; and Lu, C. 2022. 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.

Gupta, A.; Xiong, W.; Nie, Y.; Jones, I.; and Oğuz, B. 2023. 3dgen: Triplane latent diffusion for textured mesh generation. *arXiv preprint arXiv:2303.05371*.

Li, B.; Zhang, D.; Zhao, Z.; Gao, J.; and Li, X. 2025. Stitchfusion: Weaving any visual modalities to enhance multimodal semantic segmentation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 1308–1317.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Lin, H.; Fu, Y.; and Xue, X. 2023. PourIt!: Weakly-supervised Liquid Perception from a Single Image for Visual Closed-Loop Robotic Pouring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 241–251.

Lin, J.; Wang, Z.; Xu, D.; Jiang, S.; Gong, Y.; and Jiang, M. 2025. Phys4DGen: Physics-Compliant 4D Generation with Multi-Material Composition Perception. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 10398–10407.

Liu, F.; Wang, H.; Chen, W.; Sun, H.; and Duan, Y. 2024a. Make-Your-3D: Fast and Consistent Subject-Driven 3D Content Generation. *arXiv preprint arXiv:2403.09625*.

Liu, J.; Chen, Y.; Ni, B.; Mao, J.; and Yu, Z. 2023a. Inferring fluid dynamics via inverse rendering. *arXiv preprint arXiv:2304.04446*.

Liu, M.; Xu, C.; Jin, H.; Chen, L.; Varma T, M.; Xu, Z.; and Su, H. 2023b. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *Advances in Neural Information Processing Systems*, 36: 22226–22246.

Liu, X.; Iwase, S.; and Kitani, K. M. 2021. Stereobj-1m: Large-scale stereo image dataset for 6d object pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10870–10879.

Liu, X.; Jonschkowski, R.; Angelova, A.; and Konolige, K. 2020. Keypose: Multi-view 3d labeling and keypoint estimation for transparent objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11602–11610.

Liu, Y.; Chen, W.; Bai, Y.; Liang, X.; Li, G.; Gao, W.; and Lin, L. 2024b. Aligning cyber space with physical world: A comprehensive survey on embodied ai. *arXiv preprint arXiv:2407.06886*.

Moya, B.; Badías, A.; González, D.; Chinesta, F.; and Cueto, E. 2023. A thermodynamics-informed active learning approach to perception and reasoning about fluids. *Computational Mechanics*, 72(3): 577–591.

Narasimhan, G.; Zhang, K.; Eisner, B.; Lin, X.; and Held, D. 2022. Self-supervised transparent liquid segmentation for robotic pouring. In *2022 International Conference on Robotics and Automation (ICRA)*, 4555–4561. IEEE.

Qian, X.; Xu, J.; Gao, Y.; Li, M.; Li, W.; and Yin, X.-C. 2024. Understanding Dynamic Auditory and Tactile Perception for Water Filling Level Estimation. *International Journal of Social Robotics*, 1–10.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ravi, N.; Gabeur, V.; Hu, Y.-T.; Hu, R.; Ryali, C.; Ma, T.; Khedr, H.; Rädle, R.; Rolland, C.; Gustafson, L.; et al. 2024. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*.

Richter, F.; Orosco, R. K.; and Yip, M. C. 2022. Image based reconstruction of liquids from 2d surface detections. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13811–13820.

Rumelhart, D. E.; Hinton, G. E.; and Williams, R. J. 1986. Learning representations by back-propagating errors. *nature*, 323(6088): 533–536.

Sajjan, S.; Moore, M.; Pan, M.; Nagaraja, G.; Lee, J.; Zeng, A.; and Song, S. 2020. Clear grasp: 3d shape estimation of transparent objects for manipulation. In *2020 IEEE international conference on robotics and automation (ICRA)*, 3634–3642. IEEE.

Szymanski, N. J.; Rendy, B.; Fei, Y.; Kumar, R. E.; He, T.; Milsted, D.; McDermott, M. J.; Gallant, M.; Cubuk, E. D.; Merchant, A.; et al. 2023. An autonomous laboratory for the accelerated synthesis of novel materials. *Nature*, 624(7990): 86–91.

Tang, H.; Li, Z.; Zhang, D.; He, S.; and Tang, J. 2025. Divide-and-Conquer: Confluent Triple-Flow Network for RGB-T Salient Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(3): 1958–1974.

Thuerey, N.; and Pfaff, T. 2016. MantaFlow, 2018. *URL http://mantaflow. com*.

Tochilkin, D.; Pankratz, D.; Liu, Z.; Huang, Z.; Letts, A.; Li, Y.; Liang, D.; Laforte, C.; Jampani, V.; and Cao, Y.-P. 2024. Triposr: Fast 3d object reconstruction from a single image. *arXiv preprint arXiv:2403.02151*.

Triantafyllidis, E.; Acero, F.; Liu, Z.; and Li, Z. 2023. Hybrid hierarchical learning for solving complex sequential tasks using the robotic manipulation network ROMAN. *Nature Machine Intelligence*, 5(9): 991–1005.

Vahdat, A.; Williams, F.; Gojcic, Z.; Litany, O.; Fidler, S.; Kreis, K.; et al. 2022. Lion: Latent point diffusion models for 3d shape generation. *Advances in Neural Information Processing Systems*, 35: 10021–10039.

Wang, X.; Ma, K.; Zhong, R.; Wang, X.; Fang, Y.; Xiao, Y.; and Xia, T. 2024. Towards dual transparent liquid level estimation in biomedical lab: Dataset, methods and practices. In *European Conference on Computer Vision*, 198–214. Springer.

Wang, Z.; Wang, Y.; Chen, Y.; Xiang, C.; Chen, S.; Yu, D.; Li, C.; Su, H.; and Zhu, J. 2025. Crm: Single image to 3d textured mesh with convolutional reconstruction model. In *European Conference on Computer Vision*, 57–74. Springer.

Wiewel, S.; Becher, M.; and Thuerey, N. 2019. Latent space physics: Towards learning the temporal evolution of fluid flow. In *Computer graphics forum*, volume 38, 71–82. Wiley Online Library.

Wu, K.; Liu, F.; Cai, Z.; Yan, R.; Wang, H.; Hu, Y.; Duan, Y.; and Ma, K. 2024. Unique3D: High-Quality and Efficient 3D Mesh Generation from a Single Image. *arXiv preprint arXiv:2405.20343*.

Wu, S.; Zhang, H.; Liu, Z.; Chen, H.; and Jiao, Y. 2025. Enhancing Human Pose Estimation in the Internet of Things via Diffusion Generative Models. *IEEE Internet of Things Journal*.

Xie, Y.; Feng, S.; Deng, L.; Cai, A.; Gan, L.; Jiang, Z.; Yang, P.; Ye, G.; Liu, Z.; Wen, L.; et al. 2023. Inverse design of chiral functional films by a robotic AI-guided system. *Nature Communications*, 14(1): 6177.

Xu, H.; Wang, Y. R.; Eppel, S.; Aspuru-Guzik, A.; Shkurti, F.; and Garg, A. 2021. Seeing glass: joint point cloud and depth completion for transparent objects. *arXiv preprint arXiv:2110.00087*.

Xu, J.; Cheng, W.; Gao, Y.; Wang, X.; Gao, S.; and Shan, Y. 2024. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*.

Yan, S.; Chen, G.; Gao, A.; Liu, C.; and Xiong, Z. 2023. BiSPD-YOLO: Surface defect detection method for small features and low-resolution images. In *2023 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM)*, 709–714. IEEE.

Zheng, X.-Y.; Pan, H.; Wang, P.-S.; Tong, X.; Liu, Y.; and Shum, H.-Y. 2023. Locally attentional sdf diffusion for controllable 3d shape generation. *ACM Transactions on Graphics (ToG)*, 42(4): 1–13.