DAVATaR: Depth-Aware Visually Augmented Tactile Learning Across the Air-Water Interface

Anonymous Author(s)
Affiliation
Address
email

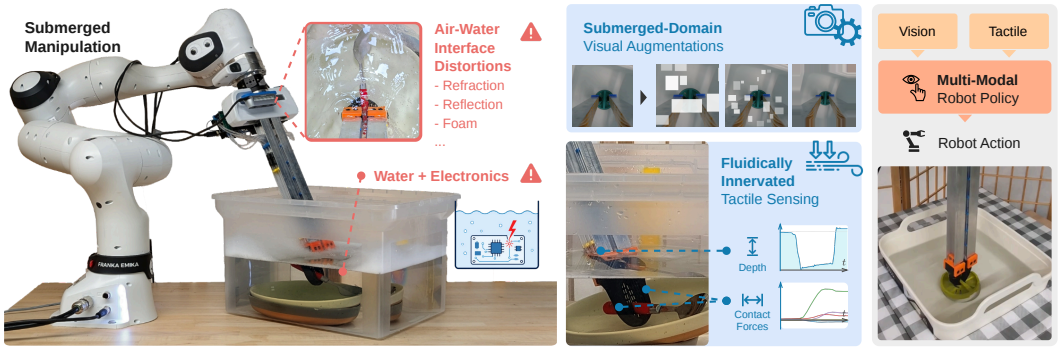


Figure 1: DAVATaR. Contact-rich submerged tasks are difficult as water introduces reflection and refraction artifacts that make visual observations unreliable, while water risks damaging electronics. We combine visual augmentation with waterproof tactile sensing to train visuo-tactile policies for submerged manipulation tasks.

Abstract: Manipulation in submerged settings (e.g., washing dishes, bathing, collecting water samples) is difficult because of the air-water interface. Water makes visual cues unreliable and damages electronics. We learn visuo-tactile policies for this regime by fusing visual observations with tactile and immersion-depth signals from a sensorized waterproof gripper. We train imitation learning transformer policies that use submerged-domain visual augmentations to reduce overfitting to turbidity, reflection- and refraction-related artifacts. Across the evaluated submerged manipulation tasks, our full policy achieves a mean success rate of 71%, compared to 41% for the best no-tactile variant. This improvement is especially pronounced under difficult visual conditions, such as floating foam and strong glare. Overall, our results demonstrate combining visual robustness with multimodal tactile sensing improves submerged manipulation. See website for details.

Keywords: Visuo-tactile Learning, Submerged Manipulation, Data Augmentation

1 Introduction

Submerged manipulation involves challenging contact-rich tasks, requiring robots in dry environments to reach into water. Unlike fully underwater systems, submerged tasks require repeated interaction across the air–water interface, introducing unique visual challenges [1, 2, 3]. In household settings, soap and foam films may cause occlusions, while outdoor settings may face sun reflections and suspended sediment. These conditions change dynamically, preventing standard visuomotor policies from robust generalization due to their assumption that visual observations remain stable and informative throughout the interaction [4, 5, 6, 7]. As visual reliability collapses, the robot must increasingly rely on tactile interaction signals. For example, scrubbing a plate or bathing a limb requires sustained force-aware physical contact. Meanwhile, water quality sampling must target specific depths to avoid floating surface contaminants and bottom sediment [8, 9]. These observations motivate a broader role for tactile sensing in submerged manipulation, to better support both contact-rich interaction and awareness of the gripper's relative position within the water.

Submitted to the 10th Conference on Robot Learning (CoRL 2026). Do not distribute.

To address these challenges, we introduce **DAVATaR**, a depth-aware visually augmented tactile robot learning framework for submerged manipulation (Fig. 1). DAVATaR trains a visuo-tactile imitation learning policy that combines tactile feedback with visual augmentations in training, enabling policies to remain effective in submerged environemnts. Contact force and depth sensing are provided by a multimodal sensorized gripper that operates robustly in both dry and wet environments. DAVATaR also introduces a physically motivated image augmentation strategy, targeting structured occlusion and scaling distortions caused by the submerged environment. We evaluate DAVATaR across four real-world tasks, demonstrating substantially improved robustness of visuo-tactile imitation learning policies in submerged environments. Our main contributions in DAVATaR are:

- A novel multimodal tactile gripper designed for submerged manipulation, enabling tactile interaction in both dry and submerged environments, together with immersion-depth sensing.
- A physically motivated visual augmentation strategy targeting structured submerged visual distortions, including occlusion, scaling, and geometric deformation.
- Training and real-world evaluation of multimodal (vision, tactile, immersion depth) imitation learning policies, demonstrating improved robustness for submerged tasks.

2 Related Work

DAVATaR is a novel method for submerged manipulation, integrating visuo-tactile signals into an imitation learning policy. DAVATaR achieves this through novel end-effector design and domain-specific visual augmentation strategies. In the following, we review prior work in those areas.

Visuo-tactile manipulation. Prior work in contact-rich manipulation leverages the complementary nature of vision and tactile sensing for robust performance [10, 11, 12]. Vision provides global scene context, while tactile provides local physical feedback, making them a good pairing so long as visual signals remain geometrically consistent throughout the interaction [13, 14, 15]. Submerged manipulation breaks this assumption as the air-water interface creates systematic visual distortions, fundamentally shifting the balance between visual and tactile feedback in favor of local interaction signals. DAVATaR advances visuo-tactile manipulation by demonstrating how to integrate force and depth sensing with vision for more robust submerged manipulation.

Underwater and submerged manipulation. Underwater manipulation has been extensively studied for submersibles [16, 17, 18], incorporating robot learning advances made for dry manipulation [19, 20, 21]. This prior work differs from our focus on submerged manipulation, as we are more interested in interaction across the air–water interface. Submersibles operate in a single consistent environment, and thus do not experience the visual artifacts of reflection and refraction or the abrupt interfacial changes in buoyancy and splashing. These dynamic changes require DAVATaR to depart from traditional underwater approaches to succeed in the submerged domain. For example, for tactile sensing design, traditional methods focus on waterproofing technologies originally developed for dry manipulation [22, 23, 24]. These approaches treat hydrostatic pressure variation as a nuisance variable, as it is does not occur in dry environments [23, 25, 26]. However, by focusing on submerged environments, we realize that hydrostatic pressure variation can serve as a useful sensing signal, especially as our gripper moves between air and water. This openness to the submerged task domain empowers DAVATaR to use unique signals for manipulation, enabling a single sensing architecture to provide both contact-force and immersion-depth feedback.

Domain-specific visual augmentation. Data augmentation is widely used to improve robustness in robot learning [27, 28]. Notably, research in autonomous driving has demonstrated improved robustness to adverse environmental conditions by augmenting training data with synthetic fog, rain, and darkness. [29, 30, 31, 32, 33]. Inspired by this approach, we introduce submerged-domain visual augmentations in DAVATaR, targeting structured distortions caused by the air–water interface, including reflection- and turbidity-caused occlusions and refraction-induced scale changes. Rather than modeling the full physics of underwater image formation, DAVATaR's augmentations expose the policy to the dominant visual perturbations encountered during submerged manipulation.

3 Design of DAVATaR

Our goal is to achieve successful manipulation in a submerged workspace, where only the robot end effector enters the water. We do so by designing **(1)** a waterproof end-effector with contact-force and immersion-depth sensing, **(2)** submerged-domain visual augmentation strategies and **(3)** a multimodal behavior-cloning transformer policy for submerged tasks.

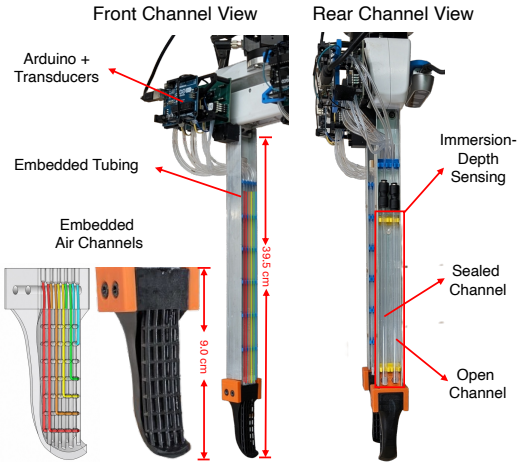


Figure 2: DAVATaR gripper design. The long-finger gripper uses fluidic tactile sensing to keep sensing electronics away from the water-exposed region. Embedded fingertip air channels provide contact-force sensing, while a rear open/sealed tube pair enables immersion-depth sensing through hydrostatic pressure.

3.1 Submerged Manipulation Gripper Design

To create hardware that can operate in wet and dry environments, we separate the submerged interaction region from the electronics through long gripper fingers with fluidic signal transmission (Fig. 2). Each finger consists of a $305\,\mathrm{mm}$ aluminum H-beam connected to a $90\,\mathrm{mm}$ compliant tactile fingertip with fingernail. We sense contact forces using fluidic innervation [34], a barometric tactile sensing method that allows us to separate readout electronics from the fingertip. Each fingertip contains 5 pairs of embedded L-shaped air channels. When the fingertip deforms under contact, the channel pressure changes and is measured by pressure transducers (All Sensors DLHR-L05D) as a tactile signal. Similar to [35], each transducer measures a pair of L-channels. This differential configuration reduces common-mode disturbances like temperature variation and hydrostatic-pressure-induced offsets. The same pressure-sensing principle is extended to immersion-depth sensing (i.e., estimating how deep each finger is in the water). Two fluidic tubes are mounted along the back side of each finger and connected to the two ports of one differential pressure transducer. One tube is open to the surrounding water, while the other is sealed as a reference channel. As the finger moves deeper below the water surface, the pressure in the open tube increases, producing a differential pressure relative to the sealed tube that correlates with depth. In summary, the gripper provides 11 sensing signals at $20\,\mathrm{Hz}$: 10 contact-force sensing signals and 1 immersion-depth sensing signal.

3.2 Submerged-Domain Visual Augmentations

Submerged manipulation tasks face visual distortions created by reflection and refraction (Fig. 3). Reflection causes glare that occludes the scene, while refraction causes objects to appear at different scales and positions. We address this through training-time perturbations. Let $\mathcal{A}$ denote the stochastic visual augmentation operator applied to a wrist-camera image I, producing an augmented image $\tilde{I} = \mathcal{A}(I)$. The operator first applies standard image augmentations, including random translation and color jitter, to cover generic camera variation. We then add two submerged-domain augmentations, $\mathcal{A}_{\mathrm{refl}}$ and $\mathcal{A}_{\mathrm{refr}}$ that mimic the distortions created by reflection and refraction, respectively. $\mathcal{A}_{\mathrm{refl}}$ overlays translucent bright rectangular patches on the image. We parameterize these patches by brightness and scattering. Brightness controls how strongly the patches are blended toward white, while scattering controls how coherent the patch spatial structure is. $\mathcal{A}_{\mathrm{refr}}$ applies a zoom perturbation parameterized by a zoom factor. The image is resized and then cropped or padded back to the original resolution, changing the apparent image scale while preserving the image size.

3.3 Policy Learning with multimodal Tactile Sensing

To achieve submerged manipulation across the air–water interface, DAVATaR combines wrist-camera observations with local gripper sensing in a single policy (Fig. 4). At timestep t, the robot receives an RGB image I_t from a wrist-mounted camera and a local sensing vector $s_t =$

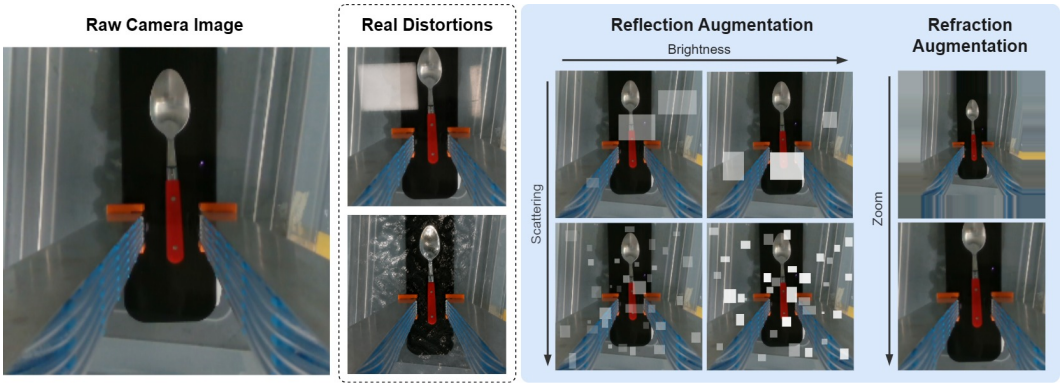


Figure 3: Visualization of submerged-domain visual augmentations. Distortions at the air–water interface affect raw wrist-camera images, adding occlusions and zoom effects. We augment images used in training by mimicking those distortions, such as overlaying light patches on the image and changing perceived image scale.

$[\tau_t, d_t]$, where $\tau_t \in \mathbb{R}^{10}$ denotes the contact-force sensing signals and $d_t \in \mathbb{R}$ denotes the immersion-depth sensing signal. The policy uses a temporal history of length H, written as $o_t = ((I_{t-H+1}, s_{t-H+1}), \ldots, (I_t, s_t))$, to predict the next action a_t, consisting of the desired operational-space end-effector motion and gripper command. Given the observation history o_t, the image sequence $I_{t-H+1:t}$ is encoded into visual tokens, while the local sensing sequence $s_{t-H+1:t}$ is encoded into local-sensing tokens. The visual and local-sensing tokens are passed through a transformer sequence model together with a learned readout token. The output at the readout token is used as the fused representation of the observation history for predicting the action distribution.

During training, the image stream can be augmented using the visual augmentation operator $\mathcal{A}$ defined in Sec. 3.2. For an input history o_t, only the wrist images are augmented, while the paired local sensing vectors remain unchanged: $\tilde{o}_t = ((\mathcal{A}(I_{t-H+1}), s_{t-H+1}), \ldots, (\mathcal{A}(I_t), s_t))$. During validation and real-world rollout, $\mathcal{A}$ is disabled and the policy receives the unmodified o_t.

We train the policy from demonstrations by minimizing the negative log likelihood of the demonstrated action under the predicted distribution: $\mathcal{L}_{\mathrm{BC}}(\theta) = -\mathbb{E}_{(o_t, a_t)\sim\mathcal{D}}\left[\log p_\theta(a_t \mid \tilde{o}_t)\right]$. During rollout, the policy maintains the most recent H observations and predicts OSC actions from the fused visual and local-sensing inputs. See Appendix for further architecture and training details.

4 Experimental Evaluation

Our experiments are organized around the following research questions:

- **RQ1:** Does DAVATaR accurately estimate contact-force?
- **RQ2:** Does DAVATaR accurately estimate immersion depth?
- **RQ3:** Does DAVATaR improve overall submerged manipulation performance?
- **RQ4:** How much does multimodal tactile sensing contribute to DAVATaR's performance?
- **RQ5:** How much do visual augmentations contribute to DAVATaR's performance?

These questions evaluate both the sensing validity of DAVATaR's hardware and the extent to which tactile sensing and visual augmentation improve policy robustness for submerged manipulation.

RQ1: Does DAVATaR accurately estimate contact-force?

We evaluate contact-force estimation by conducting ten $1\,\mathrm{min}$ pressing trials at four immersion depths: $0\,\mathrm{cm}$ (dry), $5\,\mathrm{cm}$, $10\,\mathrm{cm}$, and $20\,\mathrm{cm}$. We simulate grip force by pushing the finger laterally against the tank wall, and simulate vertical push force by pushing the finger against the tank bottom. Ground-truth force is measured using a wrist force-torque sensor (BOTA SenseOne). We use the vertical force component for push-force evaluation and the lateral force component for grip-force evaluation. To estimate force from pressure readings, we use the tactile feature representation in [35] to train an SVR with an RBF kernel, using 5-fold cross-validation. For both push-force and grip-force, we demonstrate sub-newton force estimation accuracy across different immersion depths over

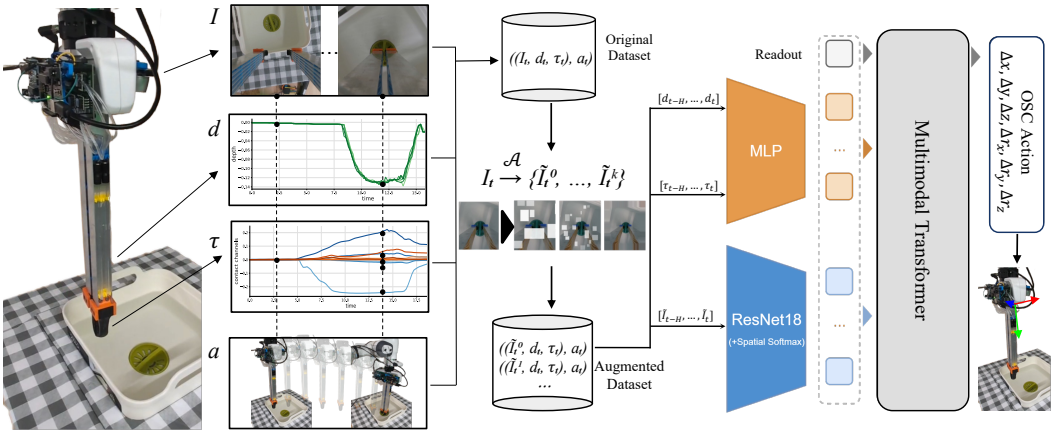


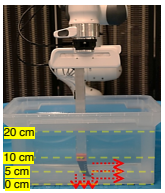
Figure 4: Policy training pipeline and architecture. At timestep t, each demonstration sample contains a wrist RGB image I_t, contact-force readings τ_t, immersion-depth reading d_t, and action label a_t. During training, the visual augmentation operator $\mathcal{A}$ produces augmented images $\tilde{I}_t^j = \mathcal{A}_j(I_t)$ while leaving d_t, τ_t, and a_t unchanged. The policy consumes a synchronized history of H observations ending at t: wrist images, contact-force readings, and immersion-depth readings. Image histories are encoded with a ResNet-18 encoder with spatial softmax, local sensing histories are encoded with an MLP, and the resulting tokens are fused by a multimodal transformer with a learned readout token to predict the operational-space action.

a 0–20 N force range (Fig. 5). Force-estimation accuracy decreases with depth, indicating that force measurements are affected by increased hydrostatic pressure. We further note that grip-force estimation is consistently more accurate than push-force estimation. This is a result of our differential channel layout, as lateral finger bending produces larger relative volume changes between paired channels than vertical compression. Overall, these results demonstrate that our sensorized gripper accurately measures contact forces, with RMSE on the order of $0.5 - 4\%$ of the signal value.

RQ2: Does DAVATaR accurately estimate immersion depth?

We evaluate immersion-depth sensing by moving the gripper inside a water tank filled to a depth of $25\,\mathrm{cm}$. Five $2\,\mathrm{min}$ trials are collected using randomized translational motions at $1\,\mathrm{cm\,s^{-1}}$ to generate varying immersion depths. Ground-truth immersion depth is defined as the vertical distance between the water surface and the base of the compliant tactile fingertip, and is computed from robot forward kinematics. The sensing signal exhibits a strong near-linear ($R^2 = 0.9996$) relationship with immersion depth across all trials (Fig. 5). A linear regression model achieves an RMSE of $1.6\,\mathrm{mm}$ for immersion-depth estimation. Notably, the sensing response exhibits a distinct transition at the air–water interface, emphasizing that we are measuring hydrostatic pressure. The sensing response is also highly repeatable across trials, with a standard deviation of only 0.30% of the full-scale range (FSR). Together with the contact-force characterization results, these experiments demonstrate that the same fluidic sensing architecture can simultaneously support both contact-force sensing and immersion-depth sensing.

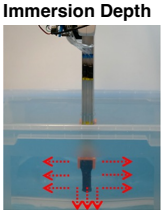
Contact Forces



RMSE (N) by Water Depth, mean ± std

Depth (cm)	Grip force (N)	Push force (N)
0	0.11 ± 0.02	0.46 ± 0.37
5	0.22 ± 0.23	0.56 ± 0.12
10	0.30 ± 0.16	0.82 ± 0.86
20	0.85 ± 0.27	0.80 ± 0.33
Overall	$\mathbf{0.37 \pm 0.28}$	$\mathbf{0.66 \pm 0.16}$

Immersion Depth



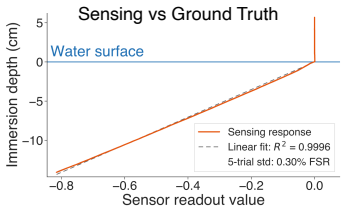


Figure 5: DAVATaR gripper characterization. Characterization results show stable force estimation across water depths and near-linear immersion-depth sensing and Red arrows represent gripper motion directions.

RQ3: Does DAVATaR improve overall performance of submerged manipulation?

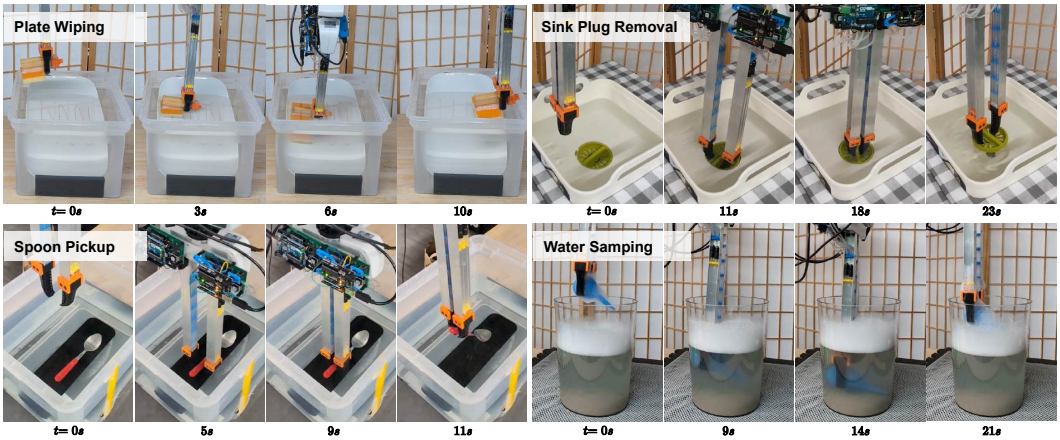


Figure 6: Submerged manipulation task suite. We evaluate DAVATaR across four contact-rich tasks, focusing on tasks that involve sustained surface contact, fixture interaction, grasping, and sampling in water.

Table 1: Task-level success rates. Entries report success rates averaged over normal and adversarial visual settings for each task. DAVATaR achieves the highest success across all evaluated tasks, showing that multimodal tactile sensing and submerged-domain visual augmentation improve submerged manipulation performance.

Method	Water sampling	Spoon pickup	Sink plug removal	Plate wiping	Mean
`VisionOnly`	0.11	0.40	0.24	0.37	0.28
`DAVATaR-T`	0.22	0.52	0.41	0.48	0.41
`DAVATaR-VA`	0.56	0.61	0.37	0.64	0.55
`DAVATaR`	**0.67**	**0.75**	**0.61**	**0.81**	**0.71**

Having established that DAVATaR provides accurate contact-force and immersion-depth measurements, we evaluate whether these sensing capabilities, combined with physically-grounded visual augmentations improve policy learning for submerged manipulation.

Experimental Setup: We evaluate RQ3-5 using four tasks: *water sampling*, *spoon pickup*, *sink plug removal*, and *plate cleaning* (Fig. 6). Together, these tasks cover immersion-depth control, submerged object retrieval, fixture manipulation, and sustained contact-rich interaction under water-induced visual distortion. We compare DAVATaR against a vanilla vision-only baseline (`VisionOnly`), which corresponds to a standard visuomotor policy [4], and two ablations: DAVATaR without tactile (`DAVATaR-T`) and DAVATaR without visual augmentations (`DAVATaR-VA`). All policies are trained from $N = 40$ demonstrations per task on a Franka Emika Panda robot and trained using the same behavior-cloning negative log-likelihood loss. For evaluation, we run $k = 18$ real-world robot rollouts for each method and task setting, corresponding to 6 trials for each of 3 random seeds. Each task includes a normal setting and one or more adversarial visual conditions with stronger water-induced distortions (see Appendix for details).

Experimental Results for RQ3: Table 1 provides the main task-level comparison. DAVATaR achieves high success in all the tasks with the highest mean success rate (71%). Notably, neither tactile sensing (41% for `DAVATaR-T`) nor submerged-domain visual augmentation (55% for `DAVATaR-VA`) alone is sufficient to achieve the performance of DAVATaR, suggesting that the two components address distinct sources of failure in submerged manipulation.

RQ4: How much does multimodal tactile sensing contribute to DAVATaR's performance?

To isolate the contribution of tactile sensing, we compare `DAVATaR` against `DAVATaR-T`, which uses the same visual augmentation strategy but does not receive contact-force or immersion-depth observations. Incorporating tactile sensing leads to a 30% increase in success across tasks, showing that multimodal tactile sensing consistently improves performance. These gains arise because submerged tasks often depend on physical information that is difficult to infer reliably from vision alone. For instance, in *water sampling*, foam occludes the water surface, preventing a vision-only

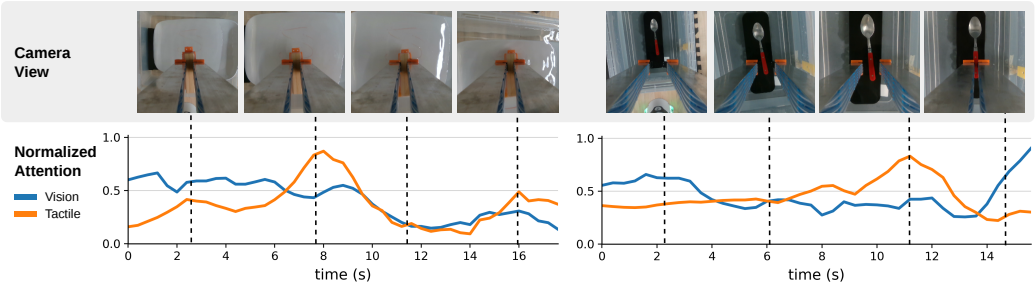


Figure 7: Modality attention during representative rollouts. We visualize normalized readout-token attention to visual and tactile tokens for plate wiping and spoon pickup. The marked frames correspond to key task phases, including approach, contact/alignment, wiping or grasping, and completion. The attention patterns show increased tactile attention during contact-sensitive phases.

policy to infer how far the syringe has entered the water. Similarly, in *spoon pickup*, refraction causes the gripper to appear closer to the spoon than it physically is, causing vision-only policies to close before reaching the spoon. Tactile feedback provides exact grasp state information for both tasks, informing the policy of the gripper's current depth and when it has hit the bottom of the tank.

Fig. 7 provides additional evidence that DAVATaR uses tactile sensing during contact-sensitive phases. The Appendix contains details on how modality attention is computed. In *plate wiping*, tactile attention increases when the sponge first contacts and aligns to the tilted plate, and again near the container boundary where the gripper should stop or adjust. In *spoon pickup*, tactile attention rises when the fingers reach the bottom and prepare to grasp. Together, the quantitative gains and attention visualization suggest that multimodal tactile sensing helps DAVATaR use immersion depth, contact/grasp state, and visually hidden physical state during submerged manipulation.

RQ5: How much do visual augmentations contribute to DAVATaR's performance?

Our submerged-domain augmentation improves policies both with and without multi-modal tactile sensing, indicating that it strengthens visual robustness rather than compensating for missing tactile feedback. As shown in Table 1, augmentation improves policies without tactile sensing from 28% (`VisionOnly`) to 41% (`DAVATaR-T`), and policies with tactile sensing from 55% (`DAVATaR-VA`) to 71% (`DAVATaR`).

The benefit of augmentation is most pronounced under adversarial visual conditions (Fig. 8, left), where water-induced artifacts more strongly corrupt the visual stream. In these settings, augmentation improves policies without tactile sensing from 9% (`VisionOnly`) to 25% (`DAVATaR-T`), and policies with tactile sensing from 37% (`DAVATaR-VA`) to 61% (`DAVATaR`).

Among individual tasks (reported in full in the Appendix), *sink plug removal* demonstrates the clearest task-level example of how augmentations help in adversarial visual conditions. Successfully completing this task requires a centered grasp on the plug before rotation. Under strong submerged visual distortion, policies trained without augmentation frequently miss the grasp due to refraction, reflection, and partial occlusion, preventing the robot from unlocking and removing the plug. We observe that in the adversarial settings, augmentation improves `VisionOnly` from 0% to 25% and `DAVATaR-VA` from 11% to `DAVATaR` at 50%. This suggests that augmentation helps the policy maintain more accurate visual guidance under water-induced distortion.

We further investigate how each augmentation helps by conducting an ablation study on the *spoon pickup* task. We compare policies with no submerged-domain augmentation against policies using only reflection-like augmentation or only refraction-like augmentation (Fig. 8). In the Reflective setting, using only $\mathcal{A}_{\text{refl}}$ improves the policy without tactile sensing from 0% (`VisionOnly`) to 54% (`DAVATaR-T-RefrA`), and improves the policy with tactile sensing from 46% (`DAVATaR-VA`) to 71% (`DAVATaR-RefrA`). This indicates that reflection-like augmentation improves robustness to glare-like occlusions that hide or corrupt local visual evidence. In the Deep-water setting, using only $\mathcal{A}_{\text{refr}}$ changes the policy without tactile sensing only marginally, from 38% (`VisionOnly`) to 39% (`DAVATaR-T-ReflA`), but improves the policy with tactile sensing from 37% (`DAVATaR-VA`) to

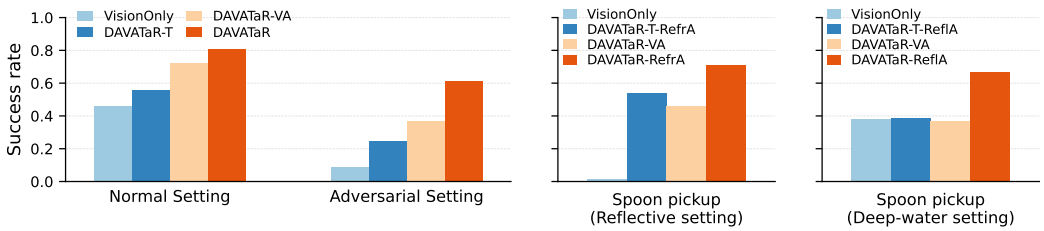


Figure 8: Success rates under adversarial visual conditions. We evaluate all policies under adversarial versions of the tasks (see Appendix). **Left:** Success rates averaged across all evaluated tasks under Normal and Adversarial settings; DAVATaR achieves the highest success in both settings. **Middle and right:** Single-augmentation ablations on spoon pickup. The middle panel evaluates the Reflective setting, where `DAVATaR-T-RefrA` and `DAVATaR-RefrA` use only reflection-like augmentation, i.e., the refraction-like augmentation is removed from the full augmentation set. The right panel evaluates the Deep-water setting, where `DAVATaR-T-ReflA` and `DAVATaR-ReflA` use only refraction-like augmentation, i.e., the reflection-like augmentation is removed from the full augmentation set. `VisionOnly` and `DAVATaR-VA` are the corresponding no-augmentation baselines without and with tactile sensing, respectively. The ablations show that targeted visual augmentation improves performance in the corresponding visual setting, with refraction-like augmentation helping most when paired with tactile sensing.

67% (`DAVATaR-ReflA`). Unlike reflections, refraction-like distortion primarily alters the apparent geometry of the scene by changing perceived object scale and relative position rather than visibility itself. As a result, vision-only policies may still observe the object but act on incorrect geometric estimates leading to missed grasps or unsafe collisions with sink bottom. In contrast, `DAVATaR-VA` policies compensate with tactile feedback to resolve this ambiguity, making the refraction-like augmentation substantially more effective. These results suggest that different submerged visual distortions induce different policy failure modes: reflections primarily degrade visibility, while refraction more strongly distorts geometric interpretation.

5 Conclusion, Discussion and Limitations

We introduced DAVATaR, a framework for submerged manipulation that combines multimodal tactile sensing with submerged-domain visual augmentation. We achieved this by integrating a compliant gripper that provides force and depth sensing with a policy-learning framework designed to improve robustness to water-induced visual artifacts. Experimental results demonstrate that our sensing architecture provides accurate contact-force and immersion-depth estimates across submerged conditions, and that DAVATaR has substantially higher performance than baseline across diverse real-world tasks. Our analysis further shows that tactile sensing and visual augmentation address complementary failure modes, yielding the strongest performance when used together.

Despite the positive results, DAVATaR has several limitations. First, our gripper's compliance can mean significant deformation during tasks. Repeated large contact forces can cause the sealant used to close each channel to peel, reducing long-term durability. Future work will investigate stiffer fingertip geometries and more robust channel sealing techniques to improve reliability under sustained contact. Second, the long fluidic transmission pathways required by the elongated finger design are sensitive to temperature variation.Temperature changes will alter the pressure of the embedded air channels due to the Ideal Gas Law, introducing sensor drift over time. Future work will investigate improved thermal isolation and temperature compensation methods. Finally, our visual augmentations model only zoom- and patch-based distortions, and do not capture time-varying water motion, dynamic lensing, or refractive defocus. More realistic camera or optics models could produce more diverse water-induced visual distortions.

More broadly, our results suggest that submerged manipulation constitutes a distinct robot learning domain in which visual observations are frequently unreliable and local physical feedback becomes critical for successful interaction. We hope DAVATaR motivates future research on sensing, learning, and manipulation across the air-water interface and in other environments where perception is degraded by structured environmental distortions.

References

[1] Y.-H. Kwon and J. B. Casebolt. Effects of light refraction on the accuracy of camera calibration and reconstruction in underwater motion analysis. *Sports Biomechanics*, 5(2):315–340, 2006.

[2] T. Treibitz and Y. Y. Schechner. Turbid scene enhancement using multi-directional illumination fusion. *IEEE Transactions on Image Processing*, 21(11):4662–4667, 2012.

[3] A. Jordt-Sedlazeck and R. Koch. Refractive structure-from-motion on underwater images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 57–64, 2013.

[4] T. Zhao, V. Kumar, S. Levine, and C. Finn. Learning fine-grained bimanual manipulation with low-cost hardware. *Robotics: Science and Systems XIX*, 2023.

[5] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 44(10-11):1684–1704, 2025.

[6] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter, et al. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

[7] A. Mandlekar, D. Xu, J. Wong, S. Nasiriany, C. Wang, R. Kulkarni, L. Fei-Fei, S. Savarese, Y. Zhu, and R. Martín-Martín. What matters in learning from offline human demonstrations for robot manipulation. In *Conference on Robot Learning*, pages 1678–1690. PMLR, 2022.

[8] L. R. Shelton. *Field guide for collecting and processing stream-water samples for the National Water-Quality Assessment Program*, volume 94. US Geological Survey, 1994.

[9] M. Liess and R. Schulz. Sampling methods in surface waters. *Handbook of water analysis*, pages 1–24, 2000.

[10] R. Calandra, A. Owens, D. Jayaraman, J. Lin, W. Yuan, J. Malik, E. H. Adelson, and S. Levine. More than a feeling: Learning to grasp and regrasp using vision and touch. *IEEE Robotics and Automation Letters*, 3(4):3300–3307, 2018. doi:10.1109/LRA.2018.2852779.

[11] M. A. Lee, Y. Zhu, K. Srinivasan, P. Shah, S. Savarese, L. Fei-Fei, A. Garg, and J. Bohg. Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8943–8950. IEEE, 2019.

[12] Y. She, S. Wang, S. Dong, N. Sunil, A. Rodriguez, and E. Adelson. Cable manipulation with a tactile-reactive gripper. *The International Journal of Robotics Research*, 40(12-14):1385–1401, 2021.

[13] Y. Chen, A. Sipos, M. Van der Merwe, and N. Fazeli. Visuo-tactile transformers for manipulation. In *Conference on Robot Learning*, 2022.

[14] B. Huang, Y. Wang, X. Yang, Y. Luo, and Y. Li. 3d-vitac: Learning fine-grained manipulation with visuo-tactile sensing. In *Conference on Robot Learning*, 2024.

[15] X. Zhu, B. Huang, and Y. Li. Touch in the wild: Learning fine-grained manipulation with a portable visuo-tactile gripper. In *Advances in Neural Information Processing Systems*, 2025.

[16] G. Marani, S. K. Choi, and J. Yuh. Underwater autonomous manipulation for intervention missions auvs. *Ocean Engineering*, 36(1):15–23, 2009. doi:10.1016/j.oceaneng.2008.08.007.

[17] G. Antonelli. *Underwater Robots: Motion, Control and Coordination*, volume 123 of *Springer Tracts in Advanced Robotics*. Springer, 4 edition, 2018. doi:10.1007/978-3-319-77899-0.

[18] E. Morgan, I. Carlucho, W. Ard, and C. Barbalata. Autonomous underwater manipulation: Current trends in dynamics, control, planning, perception, and future directions. *Current Robotics Reports*, 3:187–198, 2022. doi:10.1007/s43154-022-00089-2.

[19] R. Liu, H. Ha, M. Hou, S. Song, and C. Vondrick. Self-improving autonomous underwater manipulation. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 16915–16922. IEEE, 2025.

[20] O. Bagoren, M. Micatka, K. A. Skinner, and A. Marburg. Pugs: Perceptual uncertainty for grasp selection in underwater environments. In *IEEE International Conference on Robotics and Automation*, 2025.

[21] H. Li, L. Y. Chung, J. Goler, R. Zhang, X. Xie, H. Ha, S. Song, and M. Cutkosky. Umi-underwater: Learning underwater manipulation without underwater teleoperation. In *Robotics: Science and Systems*, 2026.

[22] R. A. S. I. Subad, L. B. Cross, and K. Park. Soft robotic hands and tactile sensors for underwater robotics. *Applied Mechanics*, 2(2):356–382, 2021. doi:10.3390/applmech2020021.

[23] J. Sun, Q. Zhang, Y. Lu, B. Huang, and Q. Li. A review of touching-based underwater robotic perception and manipulation. *Machines*, 13(1):41, 2025. doi:10.3390/machines13010041.

[24] N. Guo, X. Han, S. Zhong, Z. Zhou, J. Lin, J. S. Dai, F. Wan, and C. Song. Proprioceptive state estimation for amphibious tactile sensing. *IEEE Transactions on Robotics*, 40:4662–4683, 2024. doi:10.1109/TRO.2024.3463509.

[25] P. Kampmann and F. Kirchner. A tactile sensing system for underwater manipulation. In *Proceedings of the 7th ACM/IEEE International Conference on Human-Robot Interaction*, 2012.

[26] J. Zhang, P. Han, Q. Liu, S. Li, and B. Li. The design of underwater tactile force sensor with differential pressure structure and backpropagation neural network calibration. *Measurement and Control*, 57(2):124–138, 2024. doi:10.1177/00202940231194116.

[27] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2017.

[28] M. Laskin, K. Lee, A. Stooke, L. Pinto, P. Abbeel, and A. Srinivas. Reinforcement learning with augmented data. In *Advances in Neural Information Processing Systems*, 2020.

[29] C. Sakaridis, D. Dai, and L. Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126:973–992, 2018.

[30] V. Kilic, D. Hegde, A. B. Cooper, V. M. Patel, and M. Foster. Lidar light scattering augmentation (lisa): Physics-based simulation of adverse weather conditions for 3d object detection. In *ICASSP 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

[31] A. Anoosheh, T. Sattler, R. Timofte, M. Pollefeys, and L. Van Gool. Night-to-day image translation for retrieval-based localization. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 5958–5964. IEEE, 2019.

[32] C. Qian, Y. Guo, Y. Mo, and W. Li. Weatherdg: Llm-assisted diffusion model for procedural weather generation in domain-generalized semantic segmentation. *arXiv preprint arXiv:2410.12075*, 2024.

[33] Y. Zhu, Z. Zhu, M. Hašan, J. Yang, J. Xie, and B. Wang. Weatherdiffusion: Weather-guided diffusion model for forward and inverse rendering. *arXiv e-prints*, pages arXiv–2508, 2025.

[34] R. L. Truby, L. Chin, A. Zhang, and D. Rus. Fluidic innervation sensorizes structures from a single build material. *Science Advances*, 8(31):eabq4385, 2022. doi:10.1126/sciadv.abq4385.

[35] S. Shang, M. Seo, Y. Zhu, and L. Chin. Forte: Tactile force and slip sensing on compliant fingers for delicate manipulation. *IEEE Robotics and Automation Letters*, 2026.

A Experimental Details

A.1 Task Setup Details

Each task is evaluated under one normal setting and one or more adversarial settings. Demonstrations are collected in the normal setting, while adversarial settings are used during evaluation to test robustness to stronger water-induced visual distortions. Fig. 9 shows representative normal and adversarial conditions, and Table 2 summarizes the task variables and adversarial visual conditions used during evaluation. Each condition is evaluated with 3 trained policy seeds and 6 rollouts per seed.



Figure 9: Normal and adversarial conditions. We evaluate policies under normal settings and adversarial settings that introduce stronger water-induced visual distortions. The adversarial conditions include increased foam thickness, bubbles and surface foam, deep-water refraction, and light reflection/glare. These variations distort or corrupt the wrist-camera observation while preserving the underlying task objective, allowing us to evaluate policy robustness to air–water interface artifacts. Scale bars indicate 10 cm for reference.

Table 2: Task-specific evaluation conditions. The top block lists task variables used in the normal setting. The bottom block lists additional adversarial visual conditions used during evaluation.

Condition	Water sampling	Spoon pickup	Sink plug removal	Plate wiping
Normal setting				
Water depth	12–20 cm	8–12 cm	7–10 cm	18–21 cm
Pose / orientation	Container pose: ± 7 cm in xy	Spoon pose: $\pm 30°$, ± 5 cm in xy	Sink pose: $\pm 30°$, ± 7 cm in xy	Plate angle: $20°$, $25°$, $30°$ $150°$, $155°$, $160°$
Adversarial setting				
Occlusion	Thick foam 3–15 cm	–	Floating foam (bubble)	Floating foam (bubble)
Reflection	–	✓	✓	–
Refraction	–	Deep water 17–20 cm	–	–

Water sampling. A container is filled with water, with foam floating on the surface and sand covering the bottom. The policy lowers a syringe through the foam-covered surface, fills it within a valid immersion-depth range, and retrieves it without contacting the sand. A trial is successful if the syringe collects water from the valid depth range and is retrieved without touching the sand-covered

bottom. A trial fails if the syringe does not collect water, contacts the sand, or times out. The foam visually occludes the water surface while the sand constrains the valid insertion depth, making depth estimation difficult from vision alone.

Spoon pickup. A spoon rests at the bottom of a water-filled container. The policy grasps and retrieves the spoon from the container. A trial is successful if the spoon is grasped and lifted out of the water/container. A trial fails if the gripper closes before reaching the spoon, misses the spoon, drops it during lifting, or times out. Refraction and varying water depth distort the apparent spoon position and relative geometry between the gripper and object, making grasp localization visually ambiguous.

Sink plug removal. A sink plug is submerged underwater inside a sink fixture. The plug remains mechanically locked until rotated into a valid unlocking configuration. The policy aligns with the plug, rotates it into the liftable state, and removes it from the sink. A trial is successful if the plug is unlocked and removed from the fixture. A trial fails if the policy lifts before unlocking, over-rotates or misaligns the plug, loses contact, or times out. Successful execution requires contact-rich interaction because the plug state cannot be inferred reliably from visual observations alone.

Plate wiping. A plate coated with residue is submerged underwater at a tilted angle. The policy uses a sponge to wipe away the residue while maintaining surface contact. A trial is successful if the sponge contacts the plate surface and removes the marked residue during the wiping motion. A trial fails if the sponge misses the plate, loses sustained contact, fails to remove the residue, or times out. The task requires sustained force-aware interaction and alignment between the sponge and submerged surface under visual distortion.

A.2 Full Result Breakdown

Robot evaluation runs at $5\,\mathrm{Hz}$, and each rollout is capped at 300 control steps. For each method and evaluation condition, we evaluate 3 trained policy seeds with 6 rollouts per seed, for 18 rollouts per condition. For tasks with multiple adversarial conditions, the adversarial entry averages over all adversarial conditions.

Table 3: Full per-task success rates. Entries report success rates for each method under normal and adversarial settings. For tasks with multiple adversarial settings, Adv. averages over those settings; Avg. is the mean of Normal and Adv.

	Water sampling			Spoon pickup			Sink plug removal			Plate wiping		
Method	Normal	Adv.	Avg.	Normal	Adv.	Avg.	Normal	Adv.	Avg.	Normal	Adv.	Avg.
`VisionOnly`	0.22	0.00	0.11	0.60	0.19	0.40	0.47	0.00	0.24	0.56	0.17	0.37
`DAVATaR-T`	0.33	0.11	0.22	0.57	0.47	0.52	0.57	0.25	0.41	0.78	0.17	0.48
`DAVATaR-VA`	0.56	0.56	0.56	0.81	0.42	0.61	0.63	0.11	0.37	0.89	0.39	0.64
`DAVATaR`	0.78	0.56	0.67	0.81	0.69	0.75	0.73	0.50	0.61	0.94	0.67	0.81

A.3 Modality-Attention Computation for Visualization

We compute the modality-attention curves offline from saved trajectories. For each saved frame, we construct the policy input history with length $H = 5$ and run the trained policy in evaluation mode. For each timestep in the history, the transformer input contains one readout token, one RGB token, and one tactile token. The tactile token encodes the full multimodal tactile sensing stream, including both contact-force and immersion-depth readings.

We compute attention using attention rollout. Let $A^{\ell,h}$ denote the self-attention matrix from transformer layer ℓ and head h. We first average over attention heads, add the residual identity connection,

and row-normalize:

$$\hat{A}^{\ell} = \mathrm{row_norm}\left(\frac{1}{N_h}\sum_{h=1}^{N_h} A^{\ell,h} + I\right).$$

The rollout attention matrix is obtained by multiplying the normalized attention matrices across layers:

$$R = \hat{A}^{L}\hat{A}^{L-1}\cdots\hat{A}^{1}.$$

We use the readout token corresponding to the current policy-output timestep as the query token. For each frame, we group the rollout attention from this readout token into visual and tactile modalities:

$$\alpha_{\mathrm{vis}} = \sum_{j\in\mathcal{V}} R_{q,j}, \qquad \alpha_{\mathrm{tac}} = \sum_{j\in\mathcal{T}} R_{q,j},$$

where q is the readout-token index, $\mathcal{V}$ denotes RGB-token indices, and $\mathcal{T}$ denotes tactile-token indices. In the plotted checkpoints, each modality contributes one token per timestep, so these sums correspond to the rollout attention assigned to the RGB and tactile tokens. Because the tactile token contains both contact-force and immersion-depth readings, the visualization does not separate contact-force attention from immersion-depth attention.

For visualization, the raw visual and tactile scores are independently min–max normalized to $[0, 1]$ over the trajectory and smoothed with a 5-frame moving average. The curves should therefore be interpreted qualitatively as readout-token attention-rollout attribution proxies, not as causal evidence that a modality determined the action.

B Method Details

B.1 Tactile Sensing Details

The compliant tactile fingertips are fabricated from EPU40 elastomer (approximately Shore A 72) using a Carbon M2 printer. The distal ends of the embedded air channels are sealed with Sil-Poxy silicone adhesive. Each fingertip contains five differential tactile sensing channel pairs, resulting in ten contact-force sensing channels across the two-finger gripper. The embedded air channels have a diameter of $1\,\mathrm{mm}$, while the lattice structures within the fingertip have a diameter of $2\,\mathrm{mm}$. All pressure measurements are acquired using DLHR-L05D-E1BD-C-NAV8 differential pressure transducers mounted near the finger roots. The tactile sensing channels are connected to the pressure transducers using silicone tubing with an inner diameter of $1.59\,\mathrm{mm}$ and an outer diameter of $3.18\,\mathrm{mm}$. The immersion-depth sensing tubes have an inner diameter of $3.18\,\mathrm{mm}$ and an outer diameter of $6.35\,\mathrm{mm}$. The larger tubing diameter used for immersion-depth sensing reduces flow resistance and improves the stability of hydrostatic-pressure measurements along the longer sensing pathway.

B.2 Visual Augmentation Details

All visual augmentations are applied online during training and only to the wrist RGB image. Tactile observations, immersion-depth readings, and action labels are not augmented. During validation and real-world rollout, visual augmentations are disabled.

The full augmentation operator first applies standard RGB augmentations and then applies the two submerged-domain augmentations:

$$\mathcal{A} = \mathcal{A}_{\mathrm{refr}} \circ \mathcal{A}_{\mathrm{refl}} \circ \mathcal{A}_{\mathrm{color}} \circ \mathcal{A}_{\mathrm{trans}}.$$

Table 4 summarizes the augmentation parameters.

The translation augmentation randomly shifts the image and crops back to the original resolution. For temporal image inputs, the same random crop is applied across the history window. The color augmentation applies standard brightness, contrast, saturation, and hue jitter.

Table 4: Visual augmentation parameters.

Symbol / parameter	Meaning	Value
δ	random translation range	15 pixels
p_{color}	color jitter probability	0.9
$c_{\text{b}}, c_{\text{c}}, c_{\text{s}}, c_{\text{h}}$	brightness / contrast / saturation / hue jitter	0.3 each
p_{refl}	reflection-like augmentation probability	0.3
b	reflection brightness	$\mathcal{U}(0.3, 0.8)$
γ	reflection scattering	$\mathcal{U}(0.2, 0.8)$
ρ	target reflected area fraction	0.06
p_{refr}	refraction-like augmentation probability	0.3
z	refraction zoom factor	$\mathcal{U}(0.95, 1.05)$

The reflection-like augmentation $\mathcal{A}_{\text{refl}}$ overlays translucent bright patches to emulate glare and reflected regions on the water surface. For each forward pass, we sample a brightness value b and scattering value γ. The brightness controls how strongly the patches are blended toward white. The scattering controls the spatial structure of the patches: lower scattering produces smaller, more scattered patches, while higher scattering produces larger, more coherent glare regions. The total patch area is chosen to cover approximately a fraction ρ of the image.

The refraction-like augmentation $\mathcal{A}_{\text{refr}}$ samples a zoom factor z and applies a resize-based zoom perturbation while preserving the output image resolution. This exposes the policy to apparent scale changes caused by viewing objects through water.

The training implementation samples one stochastic augmented image sequence per forward pass. Thus, the notation $I_t \to \{\tilde{I}_t^0, \ldots, \tilde{I}_t^k\}$ in Fig. 4 is conceptual and denotes the distribution of possible augmentations rather than an explicit set generated at each iteration. For the single-augmentation ablations, reflection-only variants keep $\mathcal{A}_{\text{refl}}$ and remove $\mathcal{A}_{\text{refr}}$, while refraction-only variants keep $\mathcal{A}_{\text{refr}}$ and remove $\mathcal{A}_{\text{refl}}$.

B.3 Network Architecture and Training Details

All policies use the same behavior-cloning transformer architecture. The policy conditions on a history of $H = 5$ observations. At each timestep, the policy receives a 256×256 wrist RGB image. Policies with tactile sensing additionally receive a tactile input formed by concatenating 10 contact-force tactile channels with one immersion-depth tactile channel. Images, tactile inputs, and actions are normalized using dataset statistics.

The RGB image is encoded with an ImageNet-pretrained ResNet-18 followed by spatial softmax with 32 keypoints. The resulting visual feature is projected into a token representation. The tactile input is encoded into a tactile token through an MLP. For each timestep, the modality tokens are passed to the transformer together with a learned readout token. The readout token output parameterizes a mixture-density action head that predicts the demonstrated action distribution.

Each policy is trained from $N = 40$ demonstrations per task, collected in the normal setting using space mouse teleoperation. We use a sequence-level $90\%/10\%$ train/validation split. Policies are trained with behavior cloning by minimizing the negative log likelihood of the demonstrated action under the predicted mixture distribution.

Table 5: Policy architecture hyperparameters.

Hyperparameter	Value
History length H	5
Token dimension	64
Transformer layers	4
Attention heads	6
Transformer MLP hidden dimension	256
Transformer dropout	0.1
MDN components	5
Modality dropout, with tactile sensing	0.2
Modality dropout, without tactile sensing	0.0

Table 6: Training hyperparameters.

Hyperparameter	Value
Optimizer	AdamW
Learning rate	10^{-4}
Weight decay	0.01
Batch size	32
Epochs	101
Gradient steps per epoch	300
Learning-rate schedule	Cosine annealing
Minimum learning rate	10^{-5}
Validation frequency	Every 5 epochs
Checkpoint frequency	Every 20 epochs