CLP: A Real-World Dataset of Contaminated Lens Protectors for Robust Semantic Segmentation

Supplementary Material

This supplementary material provides additional analyses and results on the CLP dataset. First, we present extended implementation details for both restoration and segmentation models. We also include several supporting evaluations that extend the findings of the main paper, such as alignment verification, cross-dataset comparison with SIDL [13], and class-wise performance analysis. Furthermore, we conduct an in-depth evaluation of Grounded SAM [53] for open-vocabulary segmentation under contamination conditions. Finally, we include additional qualitative examples to illustrate CLP's visual characteristics and challenges.

A. Additional Experimental Setup

We provide extended implementation details for both restoration and segmentation models in Table A. All hyperparameters follow the settings from each model's official repository or original paper to maintain consistency with their standard training configurations. These details ensure transparency and support reproducibility on the CLP dataset.

B. Additional Experiments

B.1. Alignment Verification

Protocol. Our custom 3D-printed capture setup is designed to maintain robust geometric alignment between the original (protector-free) image and each contaminated pair. However, quantitatively verifying this alignment is difficult because heavy contaminants can distort or occlude scene details. To overcome this, we employ a feature-based verification protocol that remains stable under visual degradation. We first extract scale-invariant keypoints and descriptors using SIFT [85] from both the original and contaminated images to identify distinctive feature correspondences. These matched features form an initial set of correspondences, but some of them are incorrect because contaminants distort local image patterns. To address this, we employ the RANSAC [84] algorithm to estimate the homography matrix H, effectively rejecting these outliers arising from occlusions (*e.g.*, mud) or blur (*e.g.*, condensation). Finally, we compute the mean Euclidean reprojection error over the inlier set to quantify geometric alignment:

$$\epsilon = \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{x}_i' - H\mathbf{x}_i\|_2. \tag{1}$$

where $\mathbf{x}_i$ and $\mathbf{x}_i'$ denote the i-th matched keypoints in the original and contaminated images, respectively, and N is the number of inlier correspondences retained by RANSAC [84].

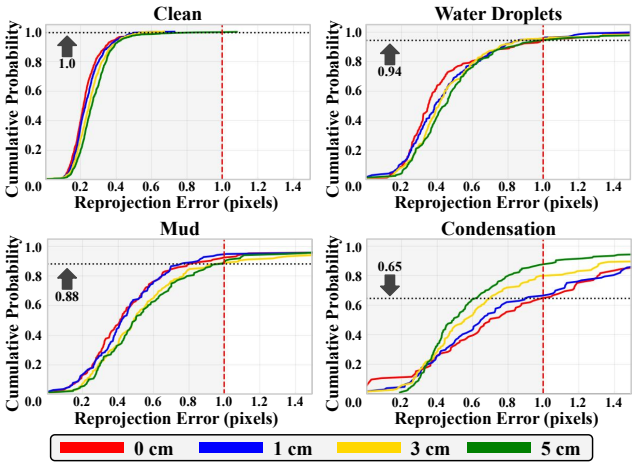

Figure A. **Cumulative distribution of reprojection errors across four lens–protector distances for each contamination type.** Each curve shows the fraction of image pairs below a given reprojection error. Higher curves indicate better alignment. The red dashed line denotes the 1.0-pixel limit, below which alignment is considered robust for pixel-level analysis.

Results. Figure A shows the cumulative distribution of reprojection errors across contamination types across different lens-to-protector distances. Following prior datasets [1, 83] that target sub-pixel alignment accuracy, we adopt the 1.0-pixel threshold (red dashed line) as the criterion for reliable geometric alignment. For the clean condition, all pairs across all distances achieve reprojection errors below 1.0 pixel, confirming that the alignment protocol performs ideally in the absence of contamination. Similarly, the water droplets and mud scenarios exhibit robust alignment performance, with over 94% and 88% of samples achieving sub-pixel accuracy (< 1.0 px), respectively. Water droplets create localized distortion, while mud causes partial occlusions. Both affect only specific image regions, allowing the feature matcher to extract distinctive keypoints from unobstructed background areas successfully. In contrast, condensation produces a global blur that uniformly degrades features across the entire image, making reliable matching more challenging. This effect leads to a lower fraction of sub-pixel reprojection errors, particularly at the 0cm and 1cm distance ($\approx 65\%$). Nevertheless, most pairs at 3cm and 5cm still achieve sub-pixel reprojection errors ($> 80\%$). This suggests that the remaining deviations are caused by the difficulty of extracting reliable features under heavy blur, rather than by any misalignment in our data capture process.

Table A. Model architectures and training hyperparameters for segmentation and restoration experiments on the CLP dataset.

Model	Backbone / Size	Pretrain	Optimizer	LR	Sched.	Iter.	Loss
Segmentation Models							
Mask2Former	ResNet-50	ImageNet-1K	AdamW	5e-5	Polynomial	160k	CE + Dice
DINOv2	ViT-L/14	LVD-142M	AdamW	5e-5	Polynomial	160k	CE + Dice
SAM	ViT-H	SA-1B	AdamW	3e-5	Polynomial	160k	CE + Dice
RobustSAM	ViT-H	SA-1B + Robust-Seg	AdamW	1e-4	Polynomial	100k	CE + Dice
DAFormer	MiT-B5	ImageNet-1K	AdamW	6e-5	Poly+Warmup	40k	CE + ST + FD
MIC	MiT-B5	ImageNet-1K	AdamW	6e-5	Poly+Warmup	40k	CE+ST+FD+MIC
Rein	DINOv2-L	LVD-142M	AdamW	6e-5	Polynomial	40k	CE + BCE + Dice
SoMA	DINOv2-L	LVD-142M	AdamW	1e-4	Polynomial	40k	CE + BCE + Dice
FIFO	ResNet-152	ImageNet-1K	Adamax/SGD	—	Fixed	16k+130k	Contrastive + CE
URIE	—	ImageNet-C	—	—	—	—	—
UniRestore	SD-UNet	Stable Diffusion v1.5	AdamW	1e-4	OneCycle/Cosine	50k+50k	MSE + L1
Restoration Models							
NAFNet	NAFNet-width32	—	AdamW	1e-3	TrueCosineAnnealing	200k	PSNRLoss
Restormer	Restormer	—	AdamW	3e-4	CosineRestart	300k	L1
DiffUIR	DiffUIR-L (64-1248)	—	Adam	2e-4	Fixed	100k	L1
MambaIR	MambaIRUNet	—	AdamW	3e-4	MultiStepLR	300k	L1

Table B. Comparison of Clean *vs.* Original training performance (averaged over 0–5cm).

Method	Training GT	Water	Mud	Condensation
NAFNet	Original	27.23 / 0.8208	21.83 / 0.7512	24.46 / 0.8095
	Clean	26.87 / 0.8205	21.56 / 0.7464	24.15 / 0.7951
	Difference	+0.36 / +0.0003	+0.27 / +0.004	+0.31 / +0.0144
Restormer	Original	26.47 / 0.8170	20.51 / 0.7211	24.34 / 0.8019
	Clean	25.89 / 0.8132	20.17 / 0.6973	24.09 / 0.8014
	Difference	+0.58 / +0.0038	+0.34 / +0.0238	+0.21 / +0.0005

Table C. Segmentation performance on Clean *vs.* Original test sets with fixed training data (averaged over 0–5cm).

Method	Clean		Original		Difference	
	mIoU	pAcc	mIoU	pAcc	mIoU	pAcc
Mask2Former	40.6	69.7	41.1	70.6	+0.5	+0.9
MIC	35.6	72.3	35.6	72.9	+0.0	+0.6
SoMA	57.5	82.5	58.1	82.9	+0.6	+0.4
DINOv2	53.0	82.2	53.3	82.0	+0.3	-0.2

B.2. Clean *vs.* Original

It is crucial to confirm that the glass plate used for contamination simulation has no side effects when uncontaminated. We assess this by comparing model performance between images captured with the glass plate (clean) and without it (original). Specifically, we compared models trained with either clean or original ground truths for restoration (Table B), while assessing the performance on both test sets for segmentation (Table C). In both scenarios, the performance gap is negligible. This confirms that the glass plate is optically transparent, introducing no meaningful artifacts such as refraction or blur compared to the protector-free setup. Consequently, applying contamination to this glass plate faithfully simulates real-world lens contamination scenarios.

Table D. Comparison of restoration performance when models are trained on SIDL vs. CLP (averaged over 0–5cm).

Method	Training	Condensation	Mud	Water
NAFNet	CLP	24.46 / 0.8095	21.83 / 0.7512	27.23 / 0.8208
	SIDL	22.52 / 0.7547	19.28 / 0.6866	23.83 / 0.7776
	Difference	+1.94 / +0.0548	+2.55 / +0.0646	+3.40 / +0.0432
Restormer	CLP	24.34 / 0.8019	20.51 / 0.7211	26.47 / 0.8170
	SIDL	21.50 / 0.7458	20.14 / 0.6955	24.91 / 0.7599
	Difference	+2.84 / +0.0561	+0.37 / +0.0256	+1.56 / +0.0571

B.3. SIDL *vs.* CLP

We demonstrate the domain gap between SIDL and CLP both quantitatively and qualitatively. As shown in Table D, we observe a clear quantitative performance gap in cross-dataset restoration experiments. Furthermore, Figure C provides qualitative evidence of this domain shift by visualizing representative contamination types from both datasets. While some SIDL [13] contamination types appear visually similar to CLP categories (*e.g.,* dust *vs.* mud, fingerprint *vs.* condensation), they only partially capture real-world optical phenomena. This limitation stems from SIDL's single-distance simulation, which produces limited diversity in degradation effects. In contrast, CLP introduces multiple lens-to-protector distances (0, 1, 3, and 5cm) to better represent real contamination scenarios. This multi-distance design captures distinct visual characteristics determined by lens-to-protector distance.

B.4. Class-wise Analysis

Table E groups categories into head, body, and tail by scene frequency. A key observation is that class frequency does not reliably predict robustness to contamination. Head classes such as cabinet and table drop below 42% mIoU under condensation and 52% under mud, showing that even common

Table E. Class-wise segmentation performance (mIoU %) grouped into Head (top 40%), Body (middle 40%), and Tail (bottom 20%) by scene frequency. Contamination types are denoted as condensation (C), mud (M), and water droplets (W).

	Head									Body									Tail								
Method	Wall (C)	Curtain (C)	Cabinet (C)	Chair (M)	Desk (M)	Table (M)	Floor (W)	Door (W)	Window (W)	Car sidemirror (C)	Shoe (C)	Bottle (C)	Keyboard (M)	Fire ext. (M)	Cable (M)	Cup (W)	Fruit (W)	Counter_top (W)	Mouse (C)	Calendar (C)	Kettle (C)	Light switch (M)	Power strip (M)	Pen (M)	Tape (W)	Car door handle (W)	Hinge (W)
DAFormer	63.1	88.2	23.6	49.0	47.9	40.8	83.0	42.4	**50.3**	58.2	**45.6**	54.3	73.1	75.4	18.6	74.8	76.3	58.4	69.7	80.7	36.3	60.3	13.3	0.9	72.2	48.0	61.9
MIC	63.3	91.0	34.4	50.3	50.4	49.6	61.8	53.5	46.0	68.5	40.6	46.8	54.2	67.1	8.0	60.5	71.0	45.1	61.5	76.7	49.8	62.8	15.5	0.0	61.2	26.5	22.7
Rein	73.9	86.6	19.9	61.4	**52.1**	**51.0**	**92.6**	**85.6**	47.4	73.9	19.1	64.9	87.5	84.2	26.7	**89.5**	**97.0**	**63.6**	85.4	**86.5**	**62.0**	60.8	**72.8**	1.8	**92.5**	68.6	**66.6**
SoMA	**77.5**	**94.9**	**41.5**	**85.0**	50.8	48.8	92.0	84.4	25.8	**81.6**	40.6	**69.2**	**91.0**	**85.9**	**28.2**	87.5	96.6	56.9	**87.6**	85.1	42.9	**67.8**	23.0	**41.8**	82.7	**71.7**	55.6

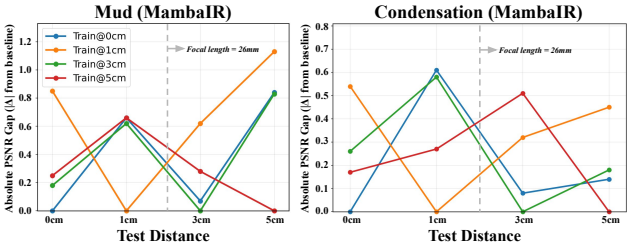

Figure B. Cross-distance performance gap analysis using MambaIR.

objects suffer from severe degradation. In contrast, visually distinctive tail classes (e.g., mouse, calendar, tape) maintain over 80% mIoU, suggesting shape and texture matter more than frequency. However, thin or textureless tail objects like pen and cable remain difficult: pen stays below 2% mIoU under mud for most models, though SoMA reaches 41.8% due to its stronger feature adaptation.

B.5. Distance-Based Analysis

Although top-performing methods benefit from strong restoration capability, domain shifts across distances highlight the need for a distance-grounded dataset. Figure B illustrates the restoration gap between models trained on specific distance data (colored lines) and those trained on the target distance (x-axis). For mud, the blue, green, and red lines show similar patterns with noticeable domain shifts, while the yellow line exhibits a distinct trend with the largest gap. For condensation, the blue and green lines share a similar trend, while the remaining lines each show different trends This difference in gap magnitude between the two types may be attributed to contamination thickness.

B.6. Open-Vocabulary Evaluation on CLP Dataset

Open-vocabulary semantic segmentation extends traditional segmentation to recognize arbitrary categories using text prompts. We evaluate Grounded SAM [53] on CLP as a representative zero-shot model, analyzing its behavior under contamination across four settings (Figure D).

(1) Standard Object Prompting. Using generic object descriptions (e.g., "book"), performance drops sharply from original to contaminated images. Even under mild blur, the model misses clearly visible objects (e.g., car parts), indicating that text–image alignment is easily disrupted by visual corruption.
(2) Contamination-Aware Prompting. Adding contamination cues (e.g., "a toy behind condensation") fails to recover reliable segmentation. The model still mislocalizes objects or produces incomplete masks, showing that textual guidance cannot compensate for severe occlusion, scattering, or blur.
(3) Grounded SAM vs. CLP Annotations. On clean images, Grounded SAM produces coarse masks that lack dense coverage. In contrast, CLP provides pixel-level manual annotations for all objects, confirming that automated segmentation alone is insufficient for precise evaluation.
(4) Effect of Contamination Intensity. As contamination intensity varies with lens–protector distance, Grounded SAM's performance degrades—consistent with the trends in Table 1 and Figure 5. These results highlight shared limitations across open-vocabulary models and the robustness methods evaluated in the main paper, underscoring the need for new strategies to handle real-world lens contamination.

B.7. Additional Qualitative Comparison

Figures E–G present additional qualitative results under different contamination types and distances for both segmentation and restoration.

References

[83] Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In *ICCV*, 2019. 1

[84] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 1

[85] David G Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004. 1

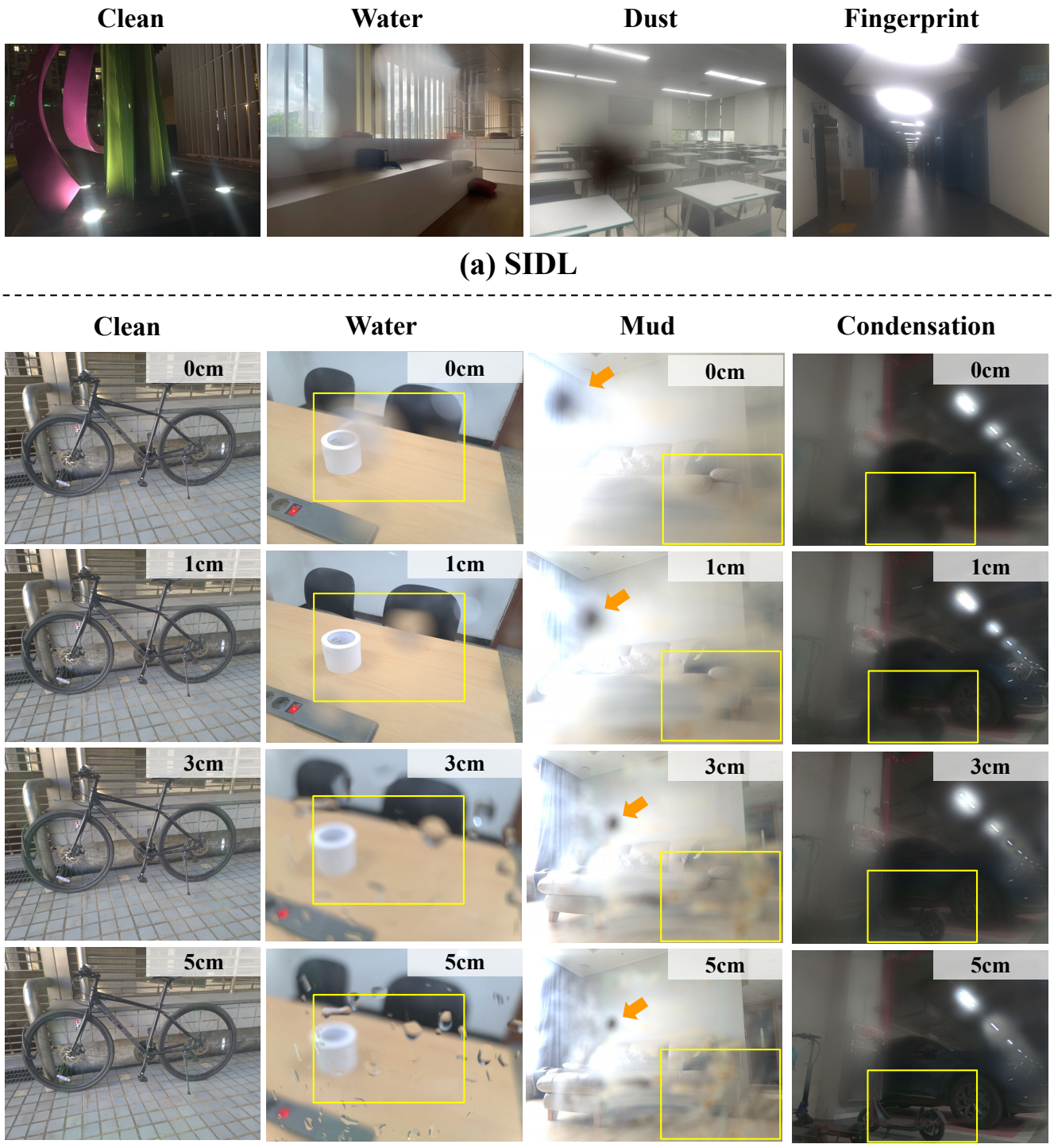

Figure C. **Qualitative comparison between SIDL [9] and CLP datasets.** SIDL provides a single contaminated view per type, lacking distance-based variation. In contrast, CLP captures contamination at multiple lens-to-protector distances (0–5 cm), which leads to clear distance-dependent changes: mud regions shift in position and extent (orange arrows), and water and condensation exhibit varying patterns and intensities within the highlighted areas (yellow boxes). These effects produce more realistic and diverse degradation characteristics that are not represented in SIDL.

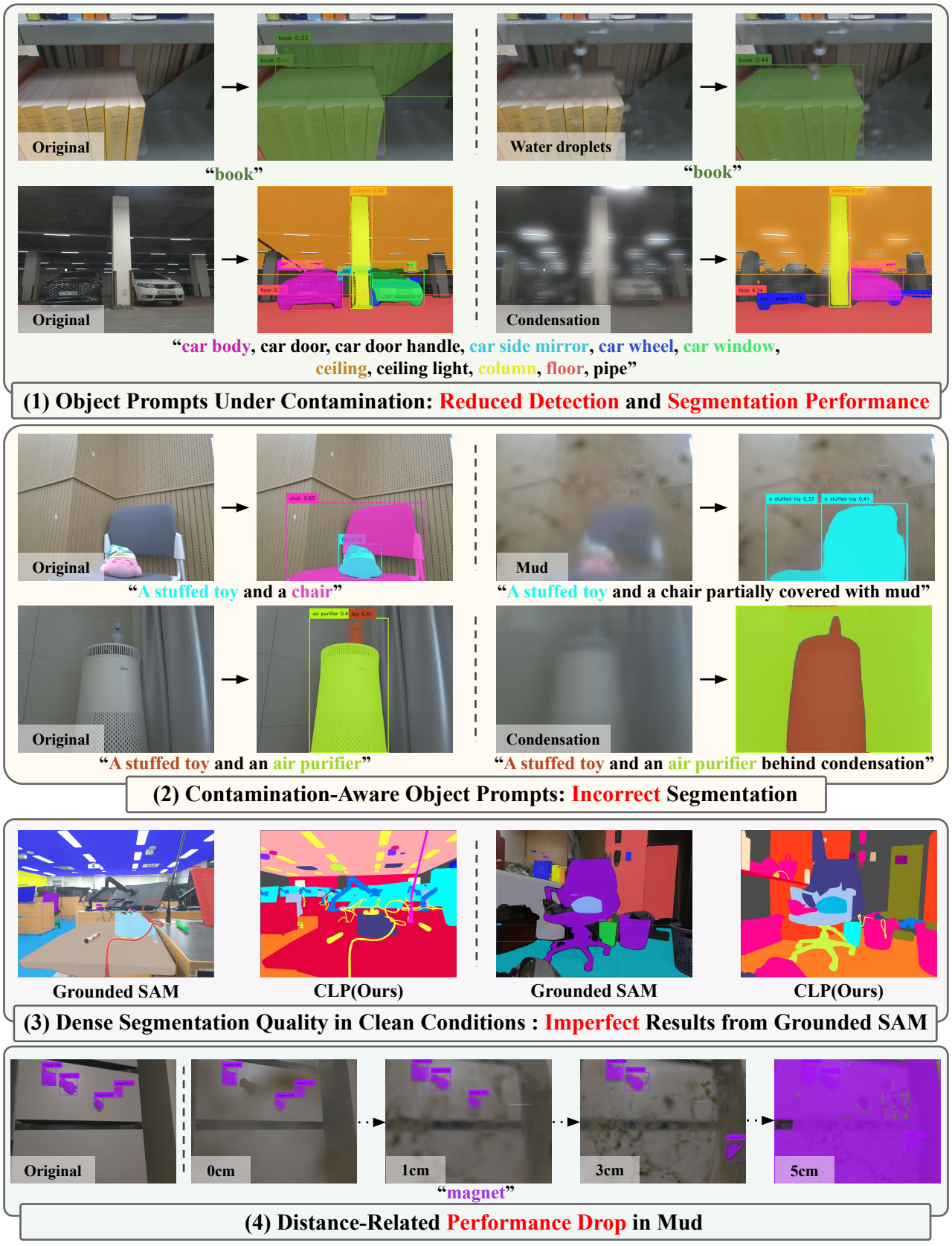

Figure D. Grounded SAM failure cases across prompt types, dense segmentation cases, and distance variations, highlighting limited robustness under real-world contamination.

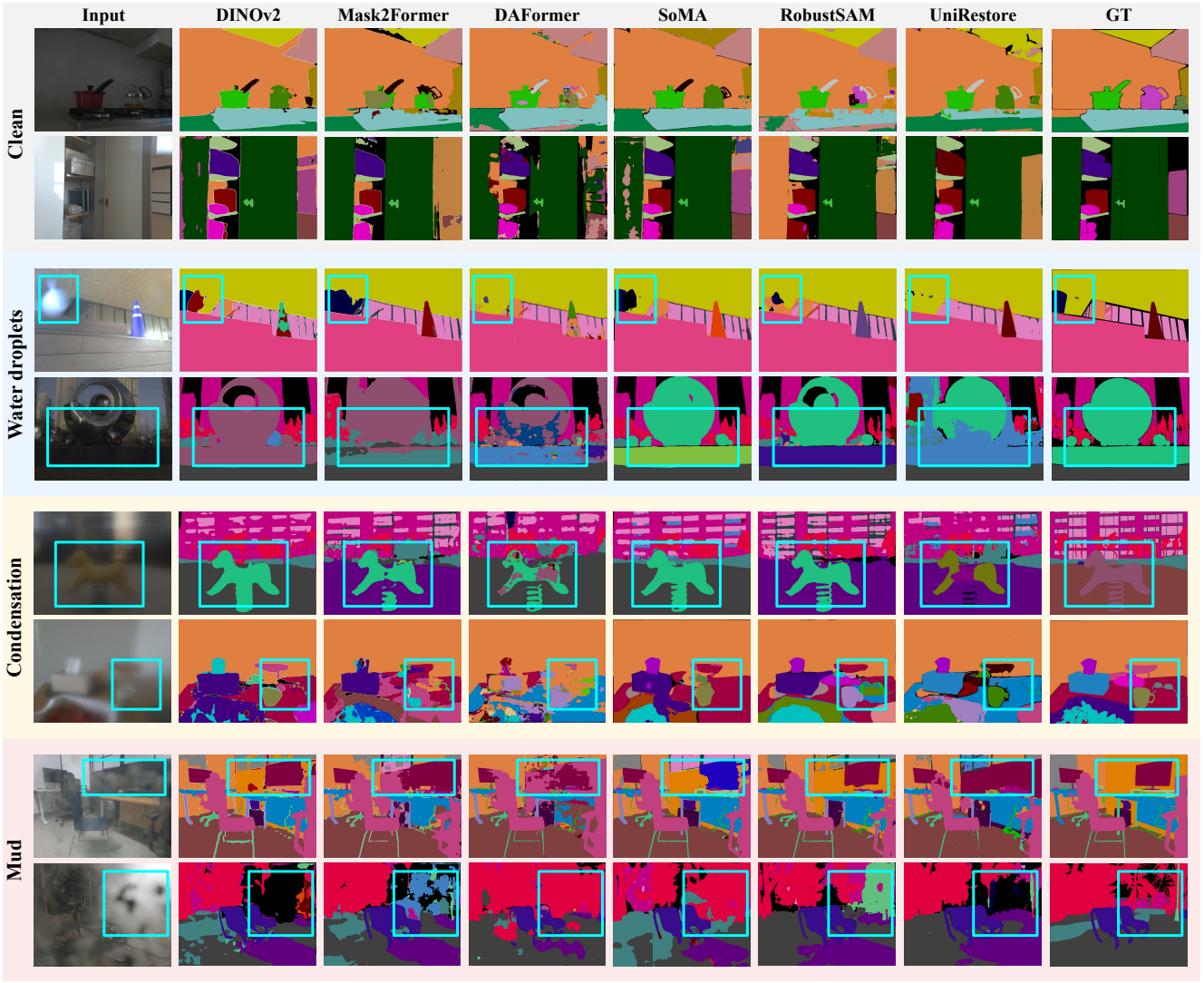

Figure E. **Additional qualitative comparison of segmentation results under CLP contamination.** Highlighted regions show where real-world contaminants in CLP cause models to miss objects, fragment boundaries, or produce incorrect segmentations. The examples illustrate how different contamination types—ranging from localized distortions to global blur and heavy occlusion—lead to various forms of prediction failure, demonstrating the difficulty of maintaining reliable segmentation under CLP.

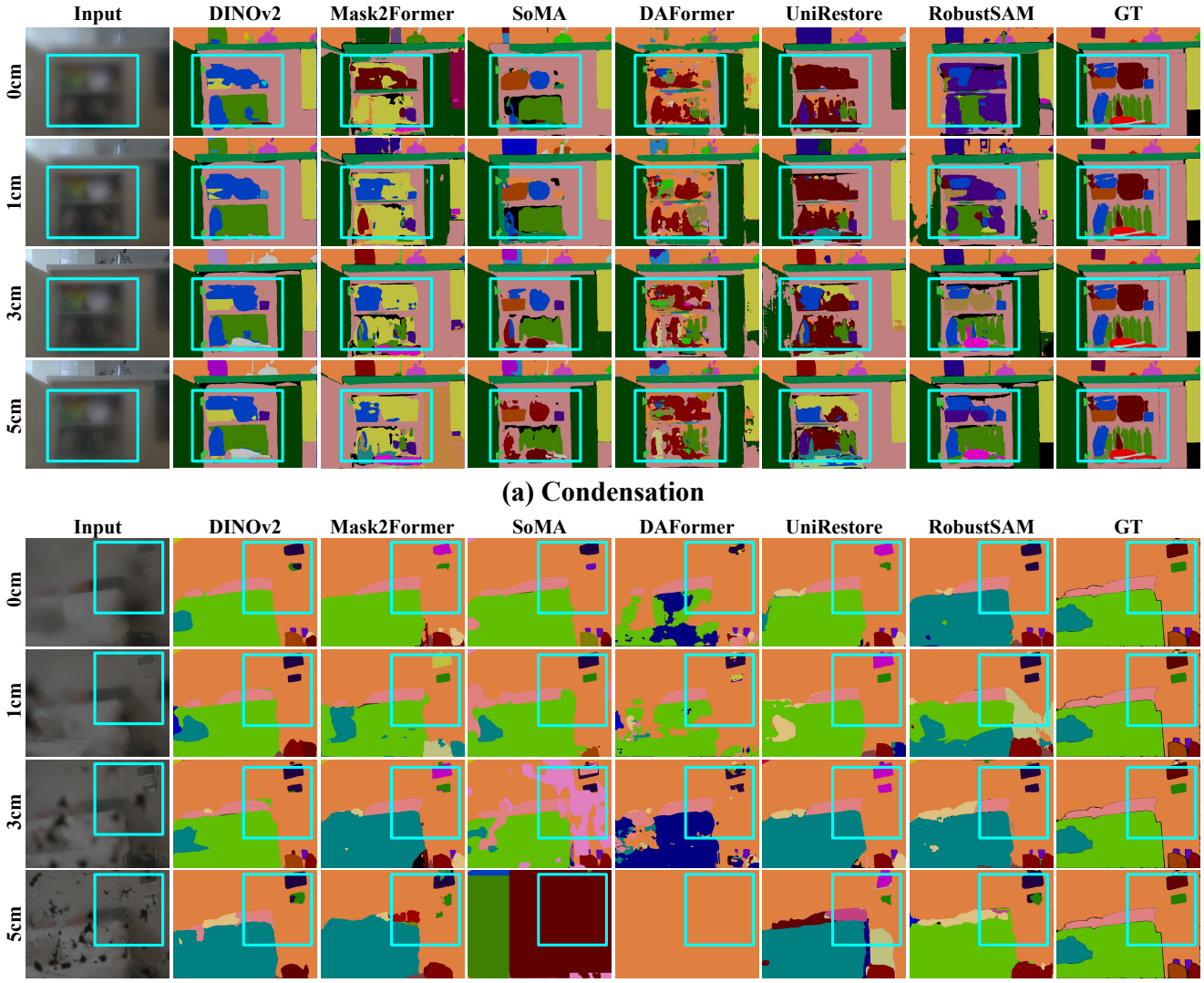

Figure F. **Additional qualitative comparison of segmentation performance across varying lens-to-protector distances (0cm, 1cm, 3cm, 5cm).** Highlighted regions show that contamination changes in position and severity across distances, causing objects that were previously segmented correctly to become missed or incorrectly predicted.

Figure G. **Additional qualitative comparisons of restoration results on CLP.** The diffusion-based model achieves the best scores but still leaves noticeable artifacts, showing that CLP contamination remains challenging to remove fully.