Multi-View Pose-Agnostic Change Localization with Zero Labels

Supplementary Material

7. Additional Details on our Methodology

7.1. Motivation for change-specific opacity factor

As discussed in Sec. 3.4, our Change-3DGS can render both RGB images of the inference scene and change maps in parallel. To achieve this, we incorporate a separate opacity factor ($\tilde{\alpha}$) – we explain the necessity of this design decision below.

During optimization, the standard 3DGS process [13] uses the opacity factor (α) to identify when Gaussians do not contribute to the modeling and should be culled. In our change detection scenario, there can be situations where the Gaussians required to model RGB appearance versus change maps can differ. For example, consider scenarios where an object present in the reference scene is missing or has been moved in the inference scene. In the standard 3DGS process, Gaussians representing such missing/moved structures lower their opacity (α) over the training as they are not visible in the set of inference images $\mathcal{I}_{\text{inf}}$, eventually becoming transparent and being pruned. However, for change modeling, these Gaussians can be critical structures for embedding change in a change mask, carrying high change magnitudes ($\tilde{c}$). For this reason, we incorporate a separate change opacity factor into each Gaussian and consider both opacity factors (α and $\tilde{\alpha}$) when determining whether a Gaussian should be removed, applying the minimum opacity threshold ϵ_α [13]. Gaussians are only removed when both α and $\tilde{\alpha}$ fall below the culling threshold.

7.2. Motivation for initializing Change-3DGS with reference scene 3DGS

We initialize our Change-3DGS with the existing 3DGS for the reference scene for two reasons: (1) many underlying structural elements of the scene are likely to remain consistent between the two scenes, and leveraging the already built reference 3DGS can allow us to update for an inference 3DGS with less data than learning from scratch; (2) as described in Sec. 7.1, the reference scene can contain Gaussians representing structures that disappear in the inference scene and are important for modeling change – these can be challenging to learn if learning the inference 3DGS from scratch.

7.3. Visualization of Data Augmentation for Learning Change Channels

We visualize the data augmentation process described in Sec. 3.6 in Fig. 5.

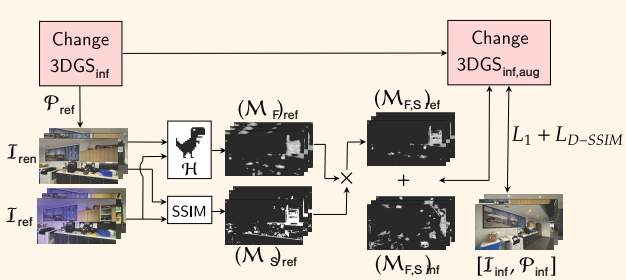


Figure 5. An overview of our data augmentation method. We concatenate the candidate masks $(\mathcal{M}_{\text{F,S}})_{\text{inf}}$ generated following Fig. 2 with candidate masks $(\mathcal{M}_{\text{F,S}})_{\text{ref}}$ obtained by considering the inference scene's representation viewed from the reference scene's poses.

7.4. Additional Implementation Details

We build the reference scene by training on $\mathcal{I}_{\text{ref}}$ and $\mathcal{P}_{\text{ref}}$ for 7000 iterations. Once initiated with a reference scene, we only train for 3000 iterations to update the representation to inference scene with $\mathcal{I}_{\text{inf}}$ and $\mathcal{P}_{\text{inf}}$ while simultaneously optimizing the change channel guided by $M_{F,P}$ (see Sec. 3.4). Once the inference scene representation is built, we fine-tune the change channel for another 3000 iterations using the augmented candidate change mask following the process described in Sec. 3.6. All the experiments were conducted on a single NVIDIA RTX 4090 GPU.

8. Additional Details on Datasets

8.1. Additional Details on MAD-Real

The MAD-Real dataset [49] has publicly released 10 scenes each containing a LEGO toy object. We illustrate each scene at the end of this Supp. Material: Bear, Bird, Elephant, Parrot, Pig, Puppy, Scorpion, Turtle, Unicorn, and Whale. During our experiments, we consider the train-set as the image set for the reference scene and the test-set as the image set for the inference scene.

8.2. Additional Details on PASLCD

We provide a breakdown of the change types and prevalence represented in PASLCD in Fig. 6. A wide range of change prevalence is tested, ranging between 0.17% and 20.12%, with an average of 3.51%.

Each figure contains a set of images from the inference scene, a set of images from the reference scene collected under similar lighting conditions to the inference images (Instance 1), and a set of images taken from the reference

Table 7. Relative performance loss (Δ) of each method when detecting changes in scenes with different lighting conditions.

Method	ΔmIoU (%) ↓	ΔF1 (%) ↓
CYWS-2D [33]	16.1	10.0
Feature Diff.	17.2	12.6
Ours	**7.2**	**4.5**

scene collected under different lighting conditions (Instance 2). The inference set is annotated with respect to Instance 1 and Instance 2.

Images were captured using an iPhone with a 16:9 aspect ratio. For each instance, a human inspector independently moved across the scene following a random trajectory, while capturing the scene with no constraints on the camera pose. Images were taken at random heights and random orientations.

We also provide additional visualizations and a description of the changes for our PASLCD dataset for each scene at the end of this Supp. Material: Cantina (see Fig. 8), Lounge (see Fig. 9), Printing area (see Fig. 10), Lunch Room (see Fig. 11), Meeting Room (see Fig. 12), Garden (see Fig. 13), Pots (see Fig. 14), Zen (see Fig. 15), Playground (see Fig. 16) and Porch (see Fig. 17).

9. Additional Experimental Results

9.1. Instance-level Results for PASLCD

Tabs. 8 and 9 show per-scene quantitative results for our PASLCD dataset under similar lighting conditions and different lighting conditions respectively. We consistently improve the change localization performance over all the baselines under both settings.

In Figs. 18 and 19 (placed towards the end of Supp. Material due to size), we show additional qualitative results for all of the methods on PASLCD under the two lighting settings.

9.2. Robustness to Distractor Visual Changes:

In Tab. 7, we report the relative *loss in performance* of each method (methods having overall mIoU ≥ 0.2) when evaluating under different lighting conditions versus consistent lighting conditions. For both the mIoU and F1 metrics, our multi-view change masks exhibit the least performance drop under different lighting conditions, demonstrating our robustness to distractor visual changes.

9.3. Complementary Information in Feature-Aware and Structure-Aware Masks

In Fig. 7, we illustrate how combining structure-aware and feature-aware masks produces a more effective candidate mask by suppressing likely false positives. The structure-aware and feature-aware masks capture complementary information about false positive change predictions, as shown in the 3rd and 4th columns of Fig. 7. While the feature-aware mask often captures changes as blobs (over-inflating the size of the change) due to the patch-to-pixel interpolation, the structure-aware mask captures more refined change details. However the structure-aware mask suffers from its own false-positive predictions, often due to the edges of fine structures in the scene or due to reflections. Combining both masks together reduces these false change predictions in the

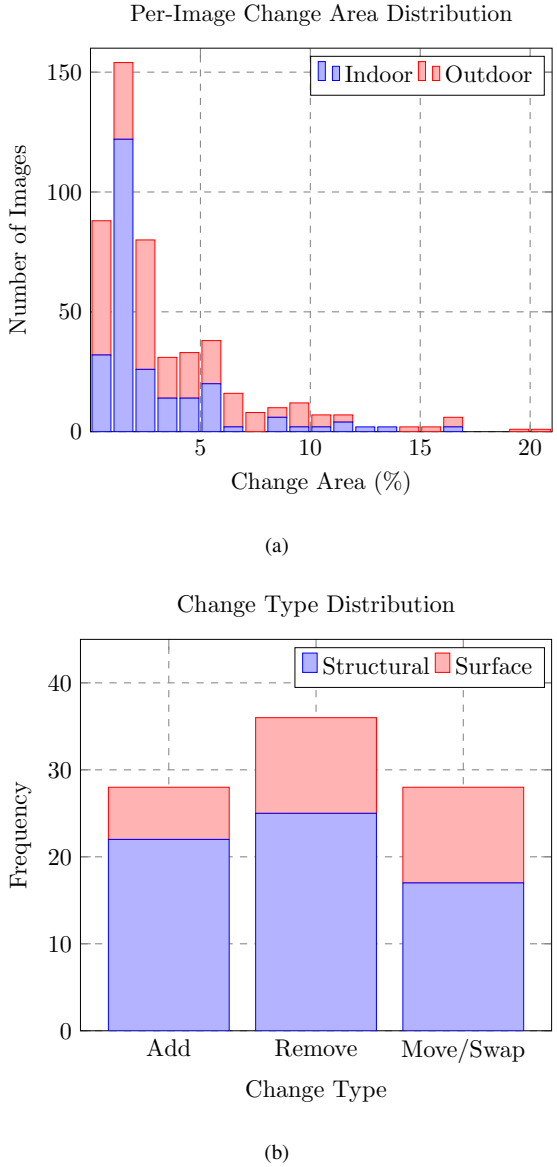


Figure 6. PASLCD dataset statistics. (a) Percentage of changed pixels across all images. (b) Distribution of change types, including structural (struct.) and surface (surf.) changes.

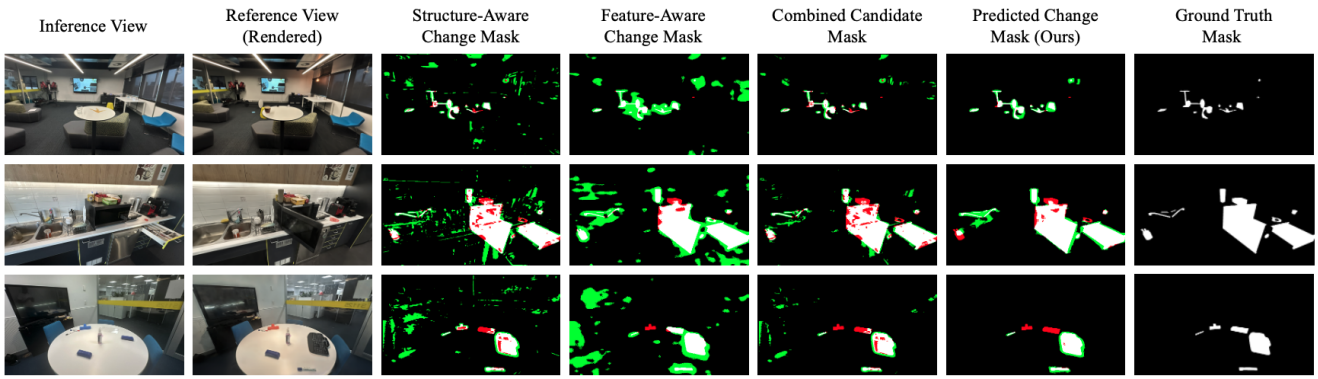


Figure 7. Qualitative visualization of change masks across two instances (under similar/different lighting conditions). From left to right: the inference view, the rendered reference view, the structure-aware change mask, the feature-aware change mask, the combined candidate mask, our predicted change mask, and the ground truth mask. The combined candidate mask effectively suppresses the distractor changes which are likely FPs (in green) by merging complementary information in structural and feature-aware masks, while our predicted change mask further refines the detection by suppressing false positives and aligning closely with the ground truth. The last row illustrates false negative failure cases discussed in Sec. 9.3 (in red). Specifically, the color change in the T-shaped structure goes undetected in the feature-aware mask, while the laminated white paper on the white table is missed in the structure-aware mask, resulting in incomplete change detection.

candidate mask (see the 5th column in Fig. 7).

However, as discussed in Sec. 5, if one of the masks fails to detect a change, it may result in missing the true change. For instance, in the 3rd row of Fig. 7, the feature-aware mask fails to capture the color change in the T-shaped structure despite the structure-aware mask flagging it, leading to an inability to fully detect the change. This highlights a potential avenue for future research: addressing the limitations of feature masks derived from pre-trained foundation models and effectively leveraging complementary information to produce a more refined change mask.

Table 8. Quantitative results for our PASLCD dataset, under similar lighting conditions, averaged across Indoor and Outdoor scenes. The best values per scene are **bolded**.

Scene	FF/360	OmniPoseAD [49]		SplatPose [16]		CSCDNet [36]		CYWS-2D [33]		Feature Diff.		Ours	
		mIoU ↑	F1 ↑	mIoU ↑	F1 ↑	mIoU ↑	F1 ↑	mIoU ↑	F1 ↑	mIoU ↑	F1 ↑	mIoU ↑	F1 ↑
Cantina	FF	0.146	0.239	0.210	0.333	0.088	0.151	0.296	0.434	0.351	0.506	**0.591**	**0.737**
Lounge	FF	0.137	0.224	0.266	0.418	0.200	0.325	0.247	0.379	0.198	0.323	**0.498**	**0.658**
Printing Area	FF	0.135	0.217	0.184	0.292	0.147	0.246	0.448	0.600	0.498	0.648	**0.637**	**0.771**
Lunch Room	360	0.144	0.224	0.146	0.234	0.037	0.065	0.108	0.183	0.103	0.176	**0.395**	**0.551**
Meeting Room	360	0.095	0.168	0.156	0.247	0.208	0.325	0.145	0.246	0.128	0.222	**0.371**	**0.531**
Garden	FF	0.297	0.440	0.228	0.357	0.245	0.389	0.347	0.510	0.265	0.410	**0.415**	**0.578**
Pots	FF	0.207	0.317	0.119	0.314	0.021	0.039	0.400	0.554	0.448	0.606	**0.569**	**0.717**
Zen	FF	0.232	0.352	0.192	0.304	0.009	0.016	0.455	0.554	0.454	0.586	**0.533**	**0.659**
Playground	360	0.074	0.121	0.096	0.155	0.131	0.213	0.054	0.100	0.041	0.078	**0.244**	**0.371**
Porch	360	0.292	0.417	0.312	0.462	0.172	0.286	0.455	0.619	0.403	0.565	**0.530**	**0.688**
Average	–	0.176	0.272	0.191	0.312	0.126	0.206	0.295	0.418	0.289	0.412	**0.478**	**0.626**

Table 9. Quantitative results for our PASLCD dataset, under different lighting conditions, averaged across Indoor and Outdoor scenes. The best values per scene are **bolded**.

Scene	FF/360	OmniPoseAD [49]		SplatPose [16]		CSCDNet [36]		CYWS-2D [33]		Feature Diff.		Ours	
		mIoU ↑	F1 ↑	mIoU ↑	F1 ↑	mIoU ↑	F1 ↑	mIoU ↑	F1 ↑	mIoU ↑	F1 ↑	mIoU ↑	F1 ↑
Cantina	FF	0.130	0.222	0.166	0.274	0.069	0.124	0.259	0.383	0.151	0.258	**0.569**	**0.720**
Lounge	FF	0.161	0.258	0.257	0.402	0.189	0.311	0.196	0.317	0.156	0.269	**0.428**	**0.593**
Printing Area	FF	0.179	0.267	0.181	0.283	0.147	0.245	0.206	0.314	0.366	0.520	**0.539**	**0.697**
Lunch Room	360	0.177	0.269	0.119	0.196	0.033	0.059	0.137	0.226	0.099	0.172	**0.382**	**0.540**
Meeting Room	360	0.118	0.196	0.104	0.175	0.218	0.335	0.130	0.220	0.115	0.200	**0.328**	**0.483**
Garden	FF	0.249	0.382	0.141	0.243	0.236	0.377	0.346	0.508	0.318	0.479	**0.456**	**0.623**
Pots	FF	0.079	0.142	0.161	0.267	0.023	0.042	0.301	0.508	0.346	0.525	**0.510**	**0.669**
Zen	FF	0.255	0.375	0.179	0.288	0.010	0.019	0.445	0.596	0.434	0.568	**0.466**	**0.606**
Playground	360	0.078	0.129	0.066	0.111	0.137	0.224	0.065	0.115	0.052	0.099	**0.254**	**0.384**
Porch	360	0.255	0.375	0.166	0.263	0.180	0.297	0.423	0.595	0.354	0.511	**0.505**	**0.664**
Average	–	0.160	0.252	0.154	0.250	0.124	0.203	0.251	0.378	0.239	0.360	**0.444**	**0.598**

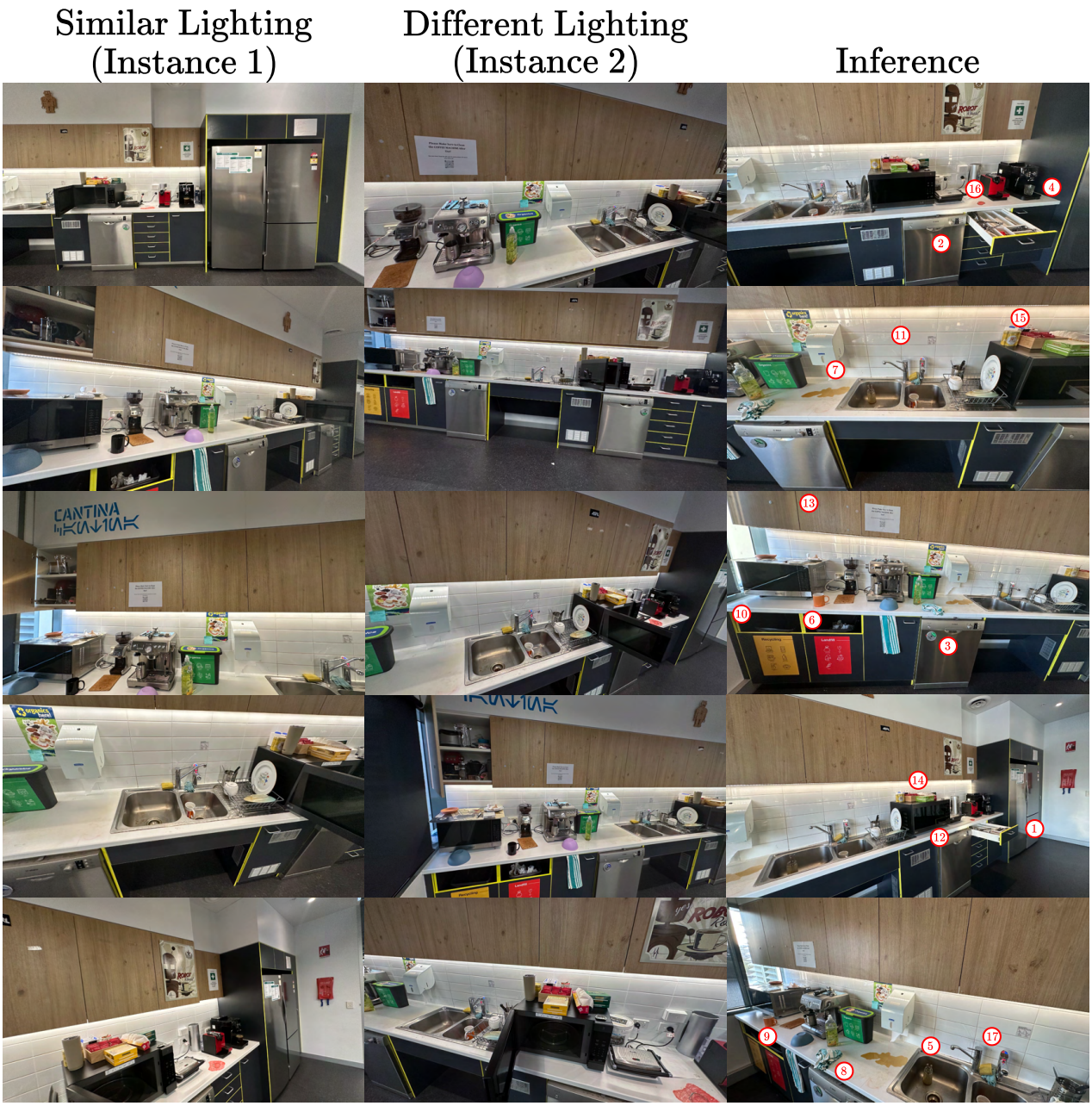


Figure 8. Cantina scene visualizations and change descriptions.

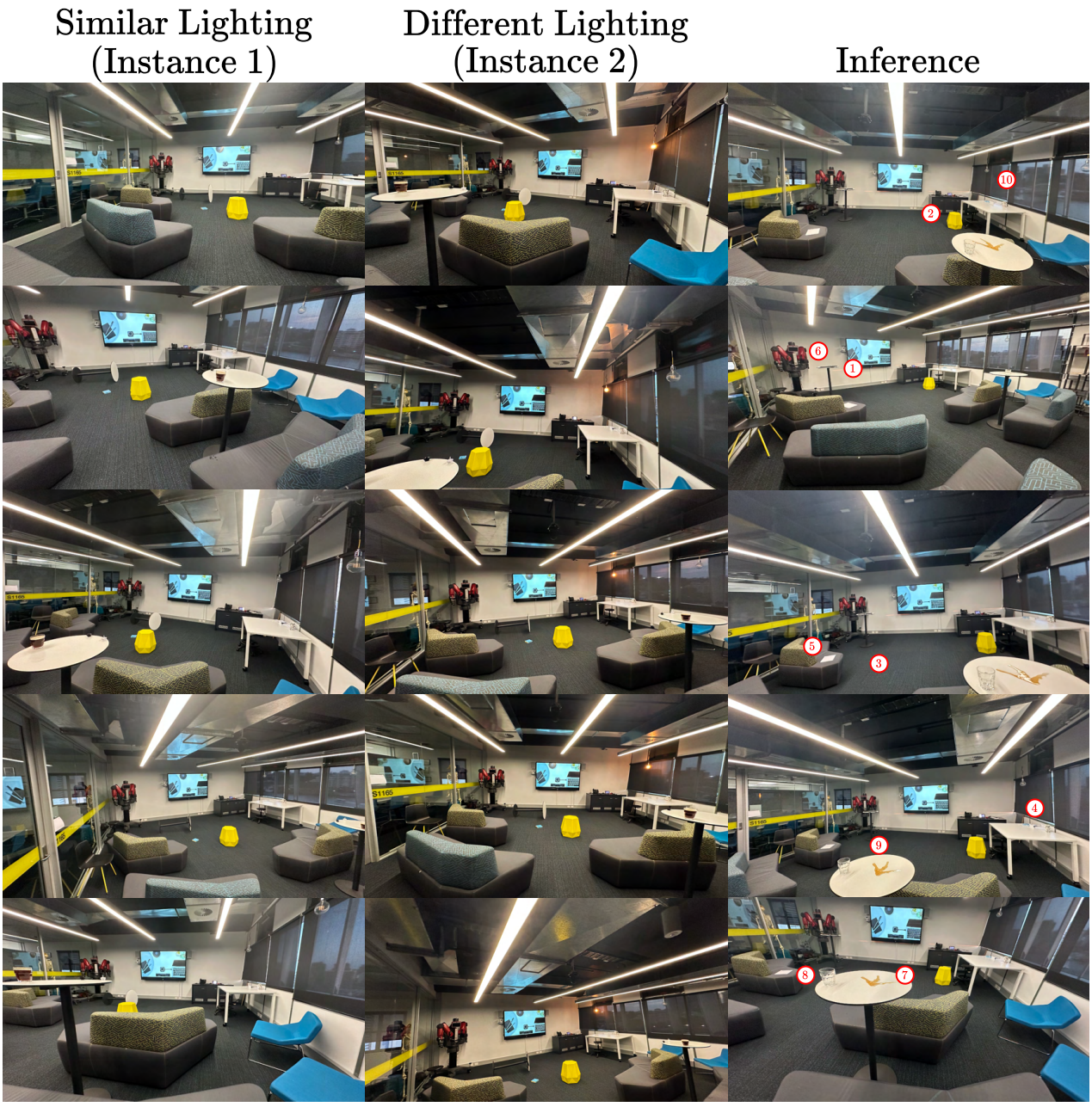


Figure 9. Lounge scene visualizations and change descriptions.

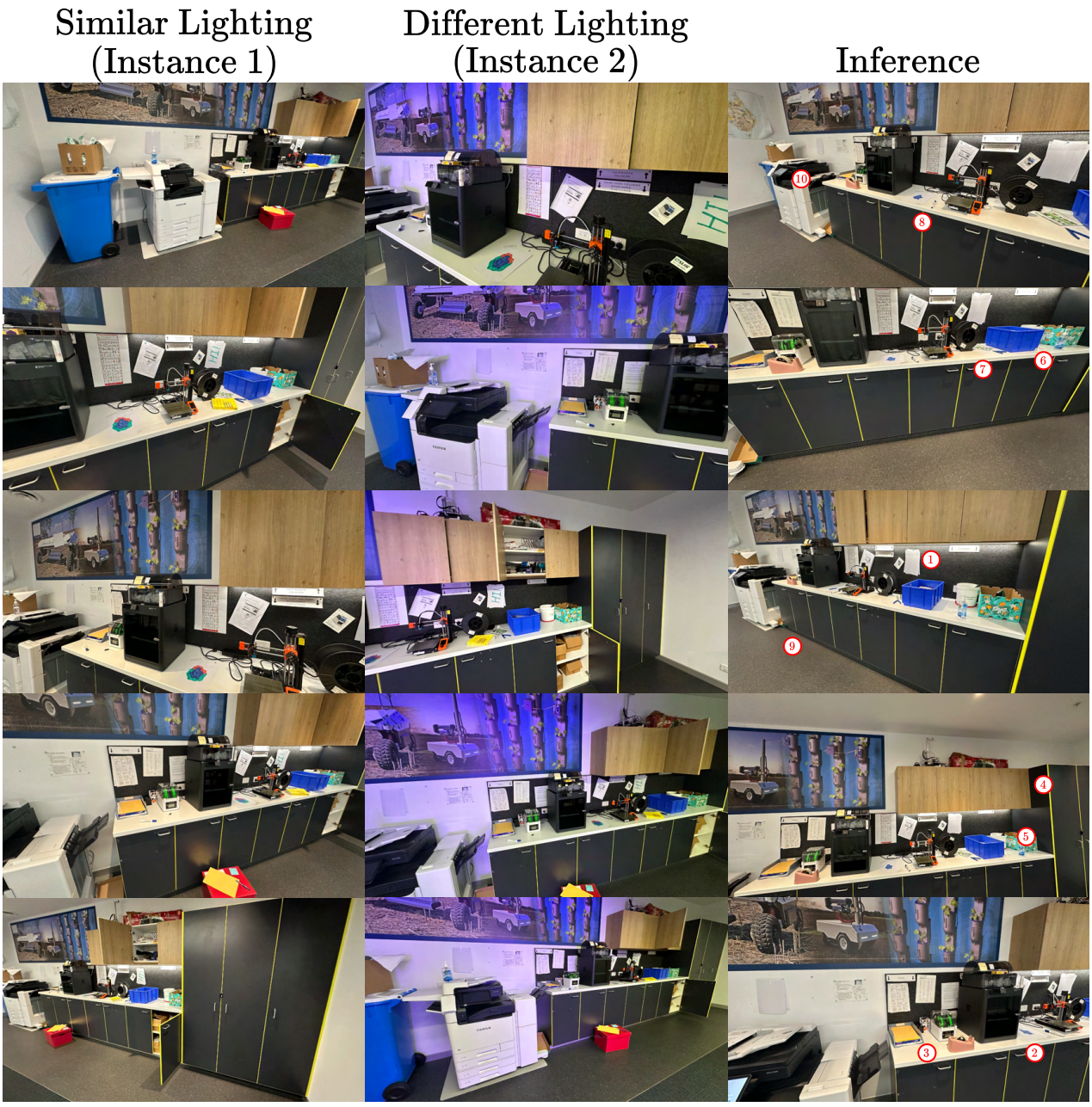


Figure 10. Printing area scene visualizations and change descriptions.

Figure 11. Lunch room scene visualizations and change descriptions.

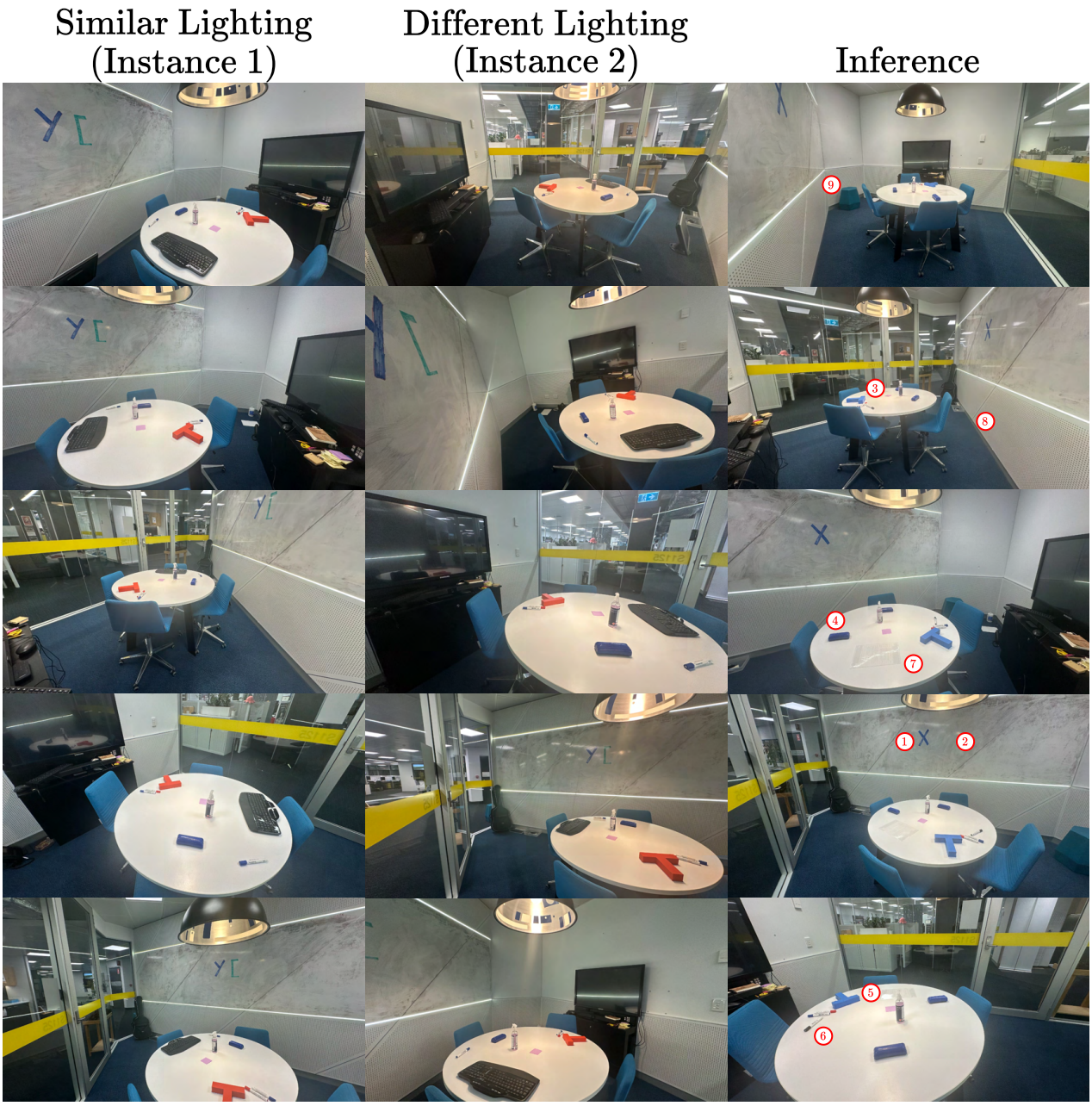


Figure 12. Meeting room scene visualizations and change descriptions.

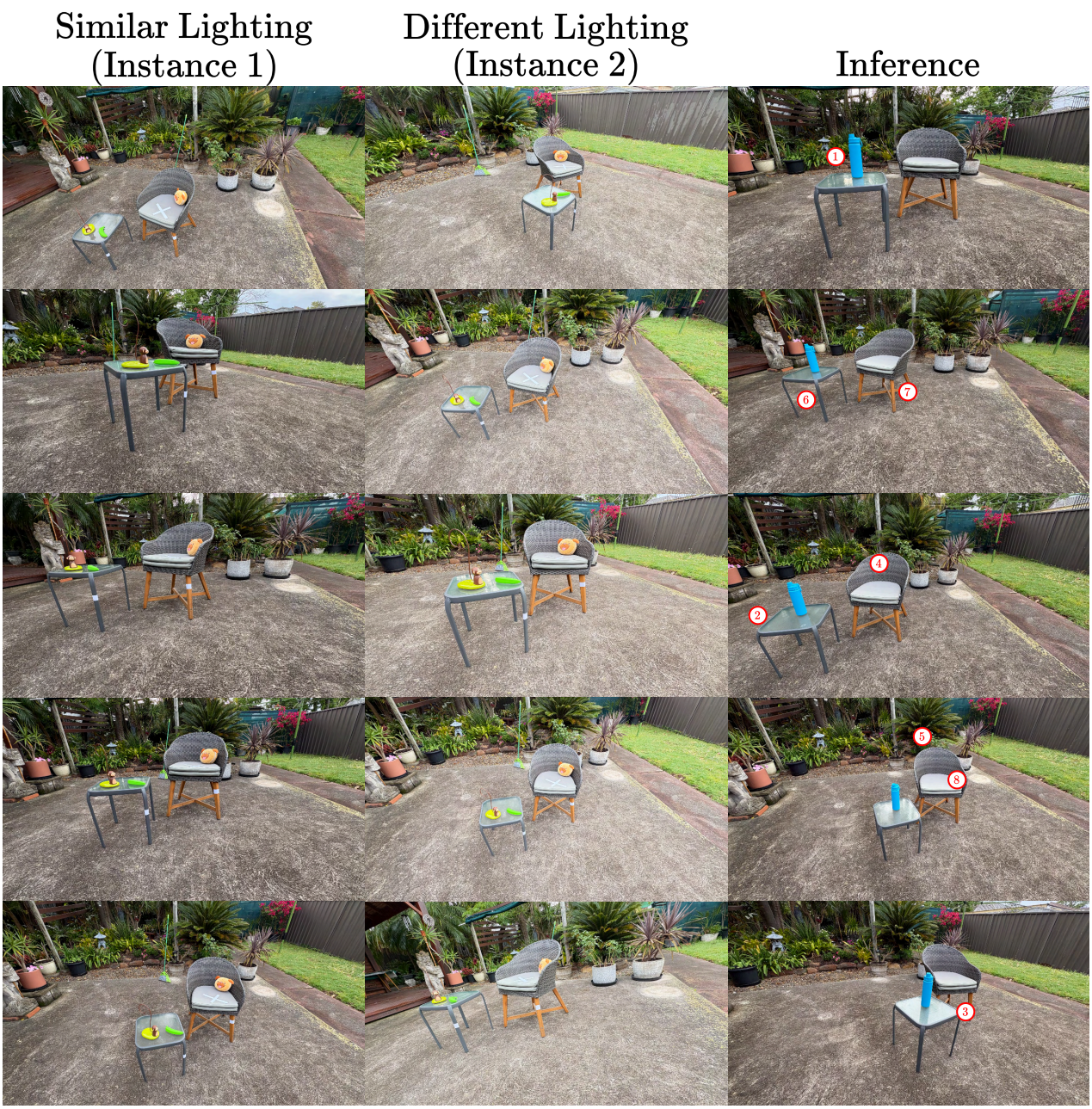


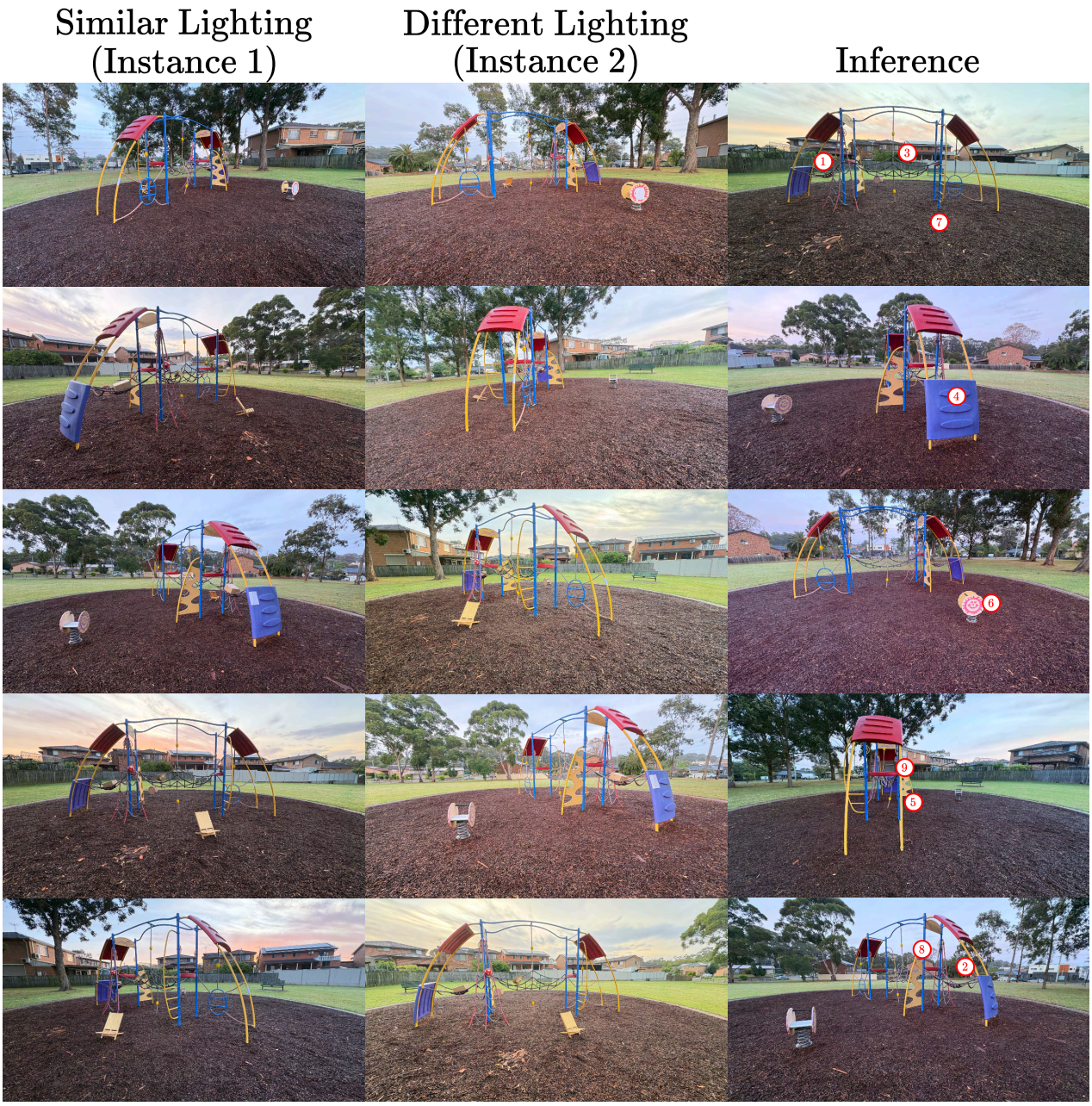
Figure 13. Garden scene visualizations and change descriptions.

1. Paper added on fence
2. Brown pot added on top row of stand
3. Gray pot added on bottom row of stand
4. Black bottle added in front of stand
5. Purple pot added in front of stand

Figure 14. Pots scene visualizations and change descriptions.

1. Red speaker added
2. Black Halloween basket added
3. Brick added
4. Dark blue pot added
5. A4 sheet added on rock
6. A4 sheet removed from bench

Figure 15. Zen scene visualizations and change descriptions.



1. Red soccer ball removed
2. Cardboard box removed
3. Hat removed
4. A4 sheet removed from purple surface
5. A4 sheet unwrapped from yellow rail
6. A4 sheet removed from spring rocker
7. Wooden chair removed
8. Boot on red platform removed
9. Boot on other red platform removed

Figure 16. Playground scene visualizations and change descriptions.

Figure 17. Porch scene visualizations and change descriptions.

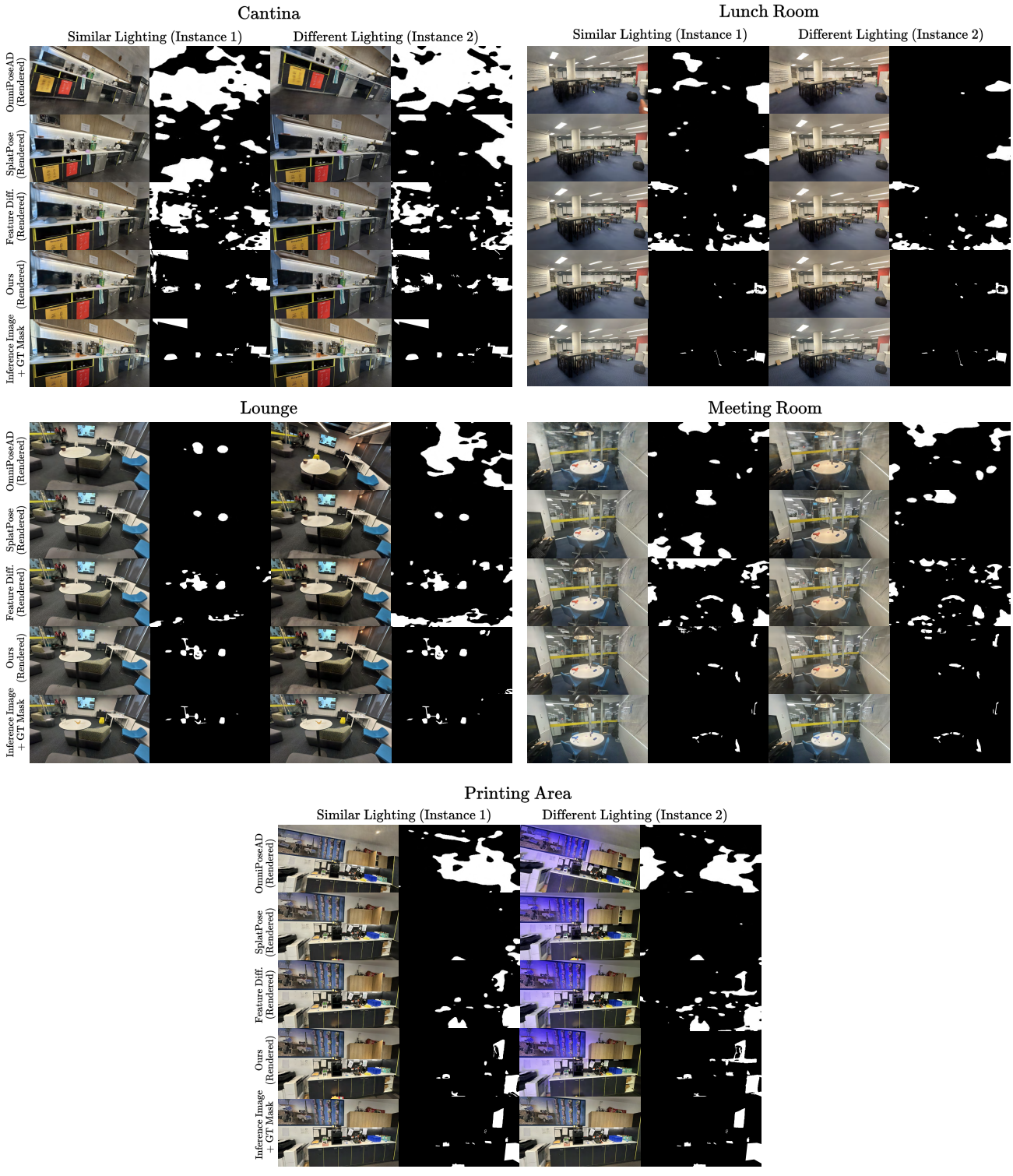


Figure 18. Qualitative results of each method for the indoor scenes of our dataset PASLCD.

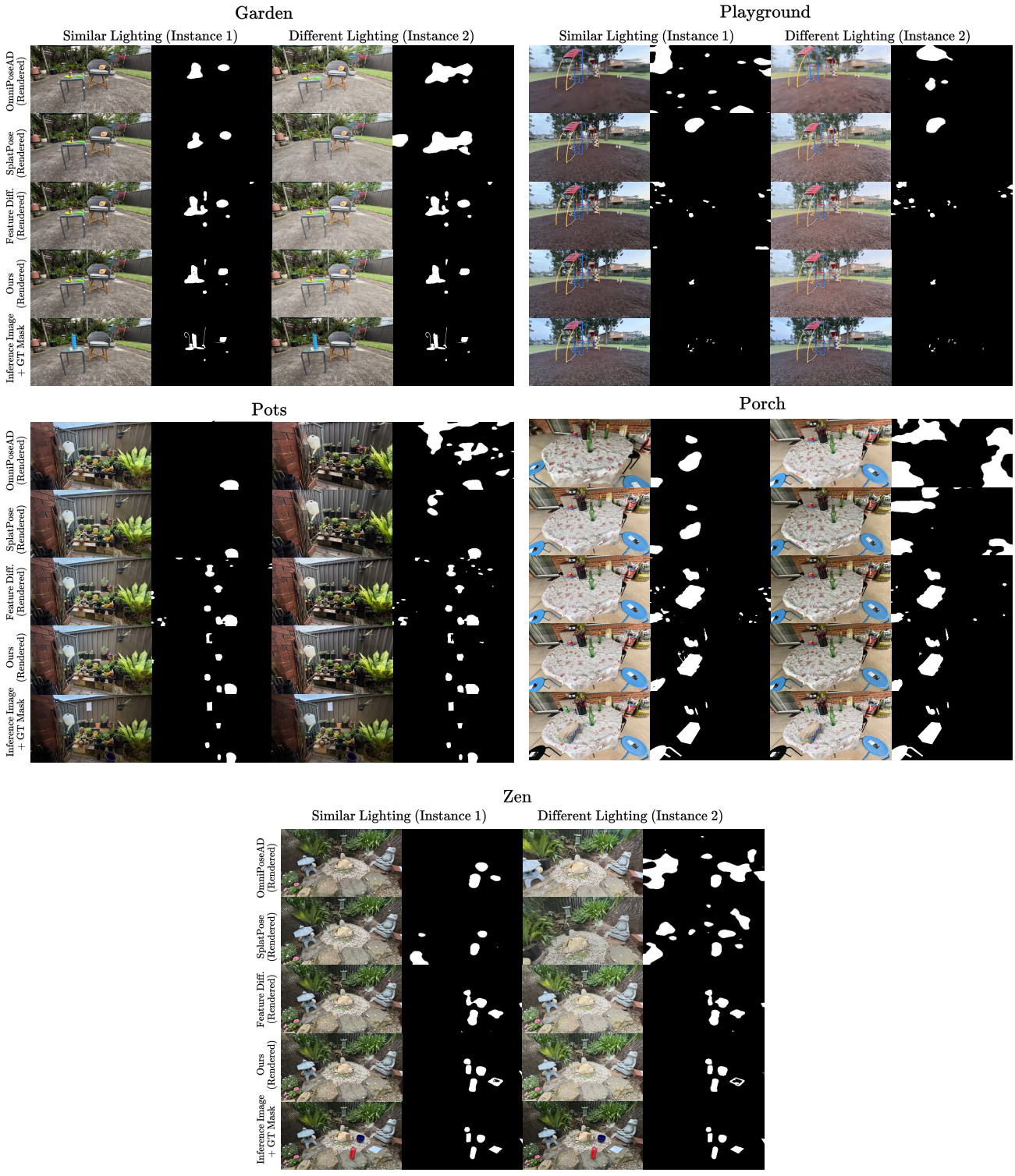


Figure 19. Qualitative results of each method for the outdoor scenes of our dataset PASLCD.