Supplementary Material

This supplementary material provides additional implementation details of methods used in this article, including the proposed methods MCCFormers and previous methods, DUDA, and M-VAM. We also give a more detailed introduction of the CLEVR-Multi-Change dataset, including the details of the caption generation process and additional dataset examples. Additional experimental results on the CLEVR-Multi-Change dataset can be found in the last section of this material.

A. Additional Implementation Details

Feature Concatenation of MCCFormers. Here, we provide more details of the feature concatenation operation used in the Ablation Study of subsection 5.2 and Table 3 in the main paper. The MCCFormers-D (encoder) outputs g_{bef} and g_{aft} with dimension of $\mathbb{R}^{W\times H\times d_{encoder}}$, respectively. The MCCFormers-S (encoder) outputs a feature map with dimension of $\mathbb{R}^{2W\times H\times d_{encoder}}$. We then separate the output to g_{bef} and g_{aft} with dimension of $\mathbb{R}^{W\times H\times d_{encoder}}$. For both two MCCFormers, we consider two ways to concatenate g_{bef} and g_{aft} (Figure 8 (a)): concatenation over patches (Figure 8 (b)) and concatenation over feature dimension (Figure 8 (c)), before feeding features to decoders. The experimental results are given in Table 3 of the main paper.

DUDA. We implemented DUDA based on the code [1] provided by the authors of DUDA. We set the dimension of the encoder and LSTM hidden layer of DUDA to 512.

M-VAM. We implemented M-VAM following the approach introduced in the original paper of M-VAM [1]. For encoder of M-VAM, two scalars in Equation (3) in [1] are learned during training. Regarding the sentence decoder, two LSTM with hidden state dimensions of 512 are trained. The network is trained with cross-entropy loss in an end-to-end manner.

For the implementation of DUDA and M-VAM, we used the same input image features, learning rate, optimizer, learning iteration as the proposed methods introduced in Implementation Details of subsection 5.1 of the main paper.

[1]The implementation code of DUDA: https://github.com/Seth-Park/RobustChangeCaptioning

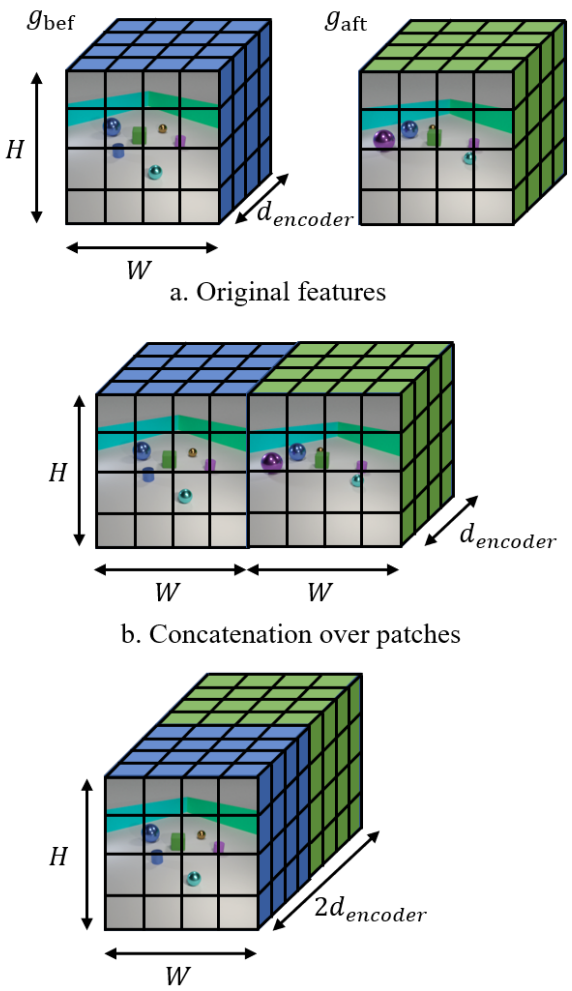


Figure 8. Visualization of feature concatenation of encoder outputs.

B. Additional Details on CLEVR-Multi-Change Dataset

Caption Generation. As introduced in section 3 of the main paper, the CLEVR-Multi-Change dataset consists of before- and after-change image pairs and captions that describe changes through language text. We record the change information during the generation of image pairs, including change type and attributes of related objects. The change captions are generated based on recorded change informa-

Change type	Caption templates
Add	"A <s> <c> <t> <z> has been added." "A <s> <c> <t> <z> shows up." "There is a new <s> <c> <t> <z>." "A new <s> <c> <t> <z> is visible." "Someone added a <s> <c> <t> <z>."
Delete	"The <s> <c> <t> <z> has disappeared." "The <s> <c> <t> <z> is no longer there." "The <s> <c> <t> <z> is missing." "There is no longer a <s> <c> <t> <z>." "Someone removed the <s> <c> <t> <z>."
Move	"The <s> <c> <t> <z> changed its location." "The <s> <c> <t> <z> is in a different location." "The <s> <c> <t> <z> was moved from its original location." "The <s> <c> <t> <z> has been moved." "Someone changed location of the <s> <c> <t> <z>."
Replace	"The <s> <c> <t> <z> was replaced by a <s1> <c1> <t1> <z1>." "A <s1> <c1> <t1> <z1> replaced the <s> <c> <t> <z>." "A <s1> <c1> <t1> <z1> is in the original position of <s> <c> <t> <z>." "The <s> <c> <t> <z> gave up its position to a <s1> <c1> <t1> <z1>." "Someone replaced the <s> <c> <t> <z> with a <s1> <c1> <t1> <z1>."

Table 8. Caption templates used in CLEVR-Multi-Change dataset. <s>, <s1>: size; <c>, <c1>: color; <t>, <t1>: material; <z>, <z1>: shape.

Models	Layers	Heads	BLEU-4 (Overall)
MCCFormers-D	1	1	59.0
		2	65.8
		4	71.0
		8	76.8
	2	1	81.4
		2	81.2
		4	82.3
		8	**82.5**
	4	1	60.3
		2	59.8
		4	64.8
		8	77.2
MCCFormers-S	1	1	58.1
		2	64.0
		4	75.8
		8	79.9
	2	1	80.0
		2	82.2
		4	**83.3**
		8	83.0

Table 9. BLEU-4 evaluation of different network designs (Layers and Heads) of MCCFormers on CLEVR-Multi-Change dataset.

tion and pre-defined sentence templates.

All templates used in the CLEVR-Multi-Change dataset are shown in Table 8. The tags "<s> <c> <t> <z>" and "<s1> <c1> <t1> <z1>" in each template are instantiated during caption generation. For example, with the template "A <s> <c> <t> <z> has been added." and an added object with attributes: small, red, metal, cube, the generated caption would be "A small red metal cube has been added."

Dataset Examples. We show additional dataset examples in Figure 9 (one-change examples), Figure 10 (two-change examples), Figure 11 (three-change examples), and Figure 12 (four-change examples).

C. Additional Experimental Results on CLEVR-Multi-Change Dataset

Additional Visualization of Examples. We show three examples with two changes on the CLEVR-Multi-Change dataset in Figure 13, Figure 14, and Figure 15. For the first two examples (Figure 13 and Figure 14), both two MCCFormers correctly generated two related sentences, while for the second example, both two MCCFormers generated a sentence with incorrect object shapes. For the third example (Figure 15), MCCFormers-D only generated one sentence, while the attention maps show that the model captured two

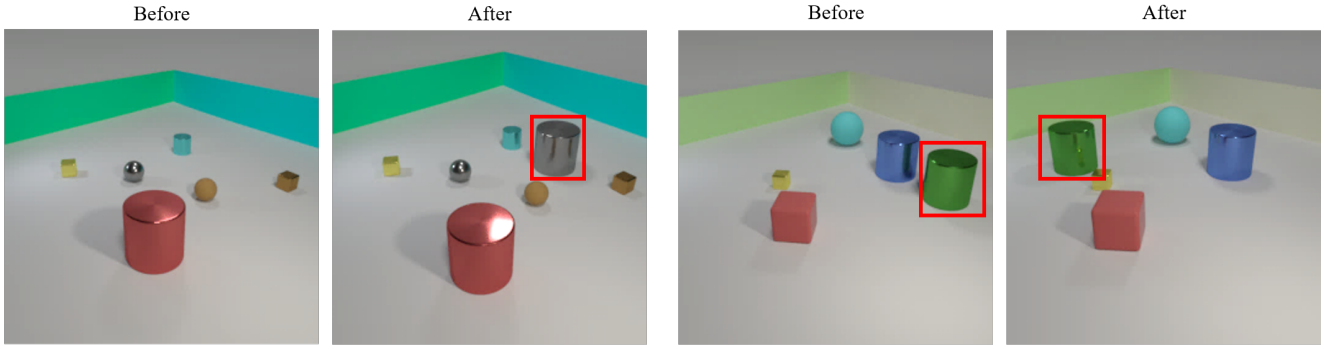
change regions.

Overall, MCCFormers-D obtained attention maps that attend to related change regions while the MCCFormers-S tends to attend to related change regions as well as unrelated regions.

Alations of Network Design of MCCFormers (Layers and Heads). The overall BLEU-4 scores of MCCFormers-D and MCCFormers-S with different layers and heads are shown in Table 9. We found that models with two layers and four heads perform relatively well for both two methods among different network designs. Therefore, we used MCCFormers-D and MCCFormers-S with two layers and four heads in experiments described in the main paper.

References

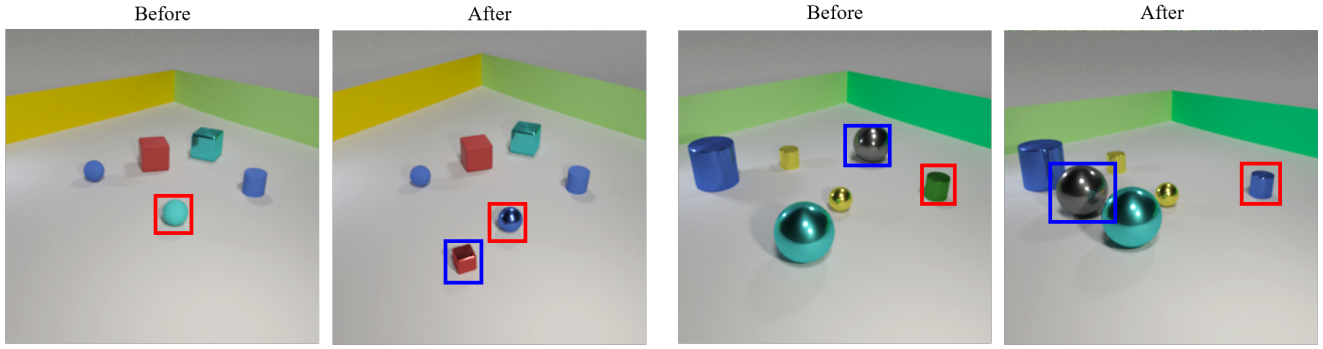
[1] Xiangxi Shi, Xu Yang, Jiuxiang Gu, Shafiq Joty, and Jianfei Cai. Finding it at another side: A viewpoint-adapted matching encoder for change captioning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 574–590, 2020. 1



Ground Truth 1: A large gray metal cylinder is visible.
Ground Truth 2: There is a new large gray metal cylinder.
Ground Truth 3: Someone added a large gray metal cylinder.
Ground Truth 4: A large gray metal cylinder has been added.
Ground Truth 5: A large gray metal cylinder shows up.

Ground Truth 1: Someone changed location of the large green metal cylinder.
Ground Truth 2: The large green metal cylinder changed its location.
Ground Truth 3: The large green metal cylinder was moved from its original location.
Ground Truth 4: The large green metal cylinder has been moved.
Ground Truth 5: The large green metal cylinder is in a different location.

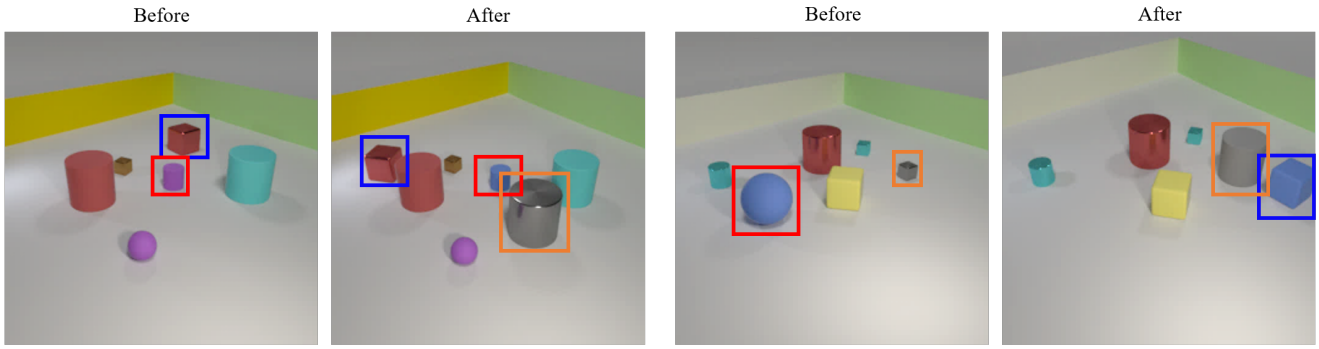
Figure 9. One-change examples from the CLEVR-Multi-Change dataset. The changed objects are highlighted by rectangles with the same color as the associated change captions.



Ground Truth 1: The small cyan rubber sphere gave up its position to a small blue metal sphere. A small red metal cube shows up.
Ground Truth 2: A small red metal cube shows up. The small cyan rubber sphere was replaced by a small blue metal sphere.
Ground Truth 3: A small red metal cube shows up. A small blue metal sphere replaced the small cyan rubber sphere.
Ground Truth 4: The small cyan rubber sphere was replaced by a small blue metal sphere. A new small red metal cube is visible.
Ground Truth 5: There is a new small red metal cube. A small blue metal sphere is in the original position of small cyan rubber sphere.

Ground Truth 1: A small blue metal cylinder replaced the small green metal cylinder. The large gray metal sphere changed its location.
Ground Truth 2: The small green metal cylinder was replaced by a small blue metal cylinder. The large gray metal sphere has been moved.
Ground Truth 3: The large gray metal sphere is in a different location. The small green metal cylinder gave up its position to a small blue metal cylinder.
Ground Truth 4: Someone replaced the small green metal cylinder with a small blue metal cylinder. The large gray metal sphere has been moved.
Ground Truth 5: Someone changed location of the large gray metal sphere. The small green metal cylinder gave up its position to a small blue metal cylinder.

Figure 10. Two-change examples from the CLEVR-Multi-Change dataset. The changed objects are highlighted by rectangles with the same color as the associated change captions.



Ground Truth 1: A small blue rubber cylinder replaced the small purple rubber cylinder. The large red metal cube has been moved. A new large gray metal cylinder is visible.

Ground Truth 2: There is a new large gray metal cylinder. The small purple rubber cylinder gave up its position to a small blue rubber cylinder. The large red metal cube was moved from its original location.

Ground Truth 3: A large gray metal cylinder shows up. The large red metal cube is in a different location. A small blue rubber cylinder is in the original position of small purple rubber cylinder.

Ground Truth 4: The large red metal cube is in a different location. Someone replaced the small purple rubber cylinder with a small blue rubber cylinder. Someone added a large gray metal cylinder.

Ground Truth 5: Someone added a large gray metal cylinder. Someone changed location of the large red metal cube. A small blue rubber cylinder replaced the small purple rubber cylinder.

Ground Truth 1: There is no longer a large blue rubber sphere. A new large blue rubber cube is visible. The small gray metal cube was replaced by a large gray rubber cylinder.

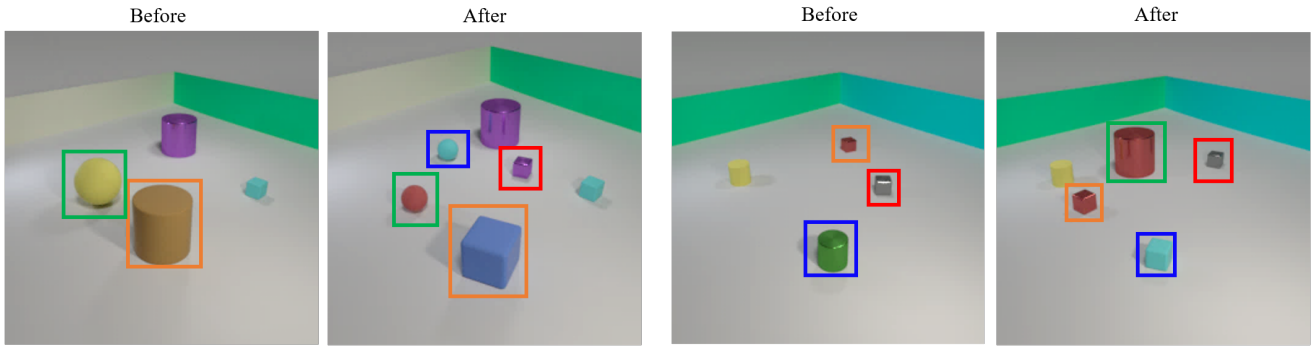
Ground Truth 2: Someone replaced the small gray metal cube with a large gray rubber cylinder. Someone removed the large blue rubber sphere. A new large blue rubber cube is visible.

Ground Truth 3: There is a new large blue rubber cube. The small gray metal cube was replaced by a large gray rubber cylinder. The large blue rubber sphere is no longer there.

Ground Truth 4: Someone added a large blue rubber cube. Someone replaced the small gray metal cube with a large gray rubber cylinder. The large blue rubber sphere is no longer there.

Ground Truth 5: The small gray metal cube was replaced by a large gray rubber cylinder. There is no longer a large blue rubber sphere. A large blue rubber cube shows up.

Figure 11. Three-change examples from the CLEVR-Multi-Change dataset. The changed objects are highlighted by rectangles with the same color as the associated change captions.



Ground Truth 1: A small purple metal cube shows up. A small cyan rubber sphere shows up. Someone replaced the large brown rubber cylinder with a large blue rubber cube. The large yellow rubber sphere gave up its position to a small red rubber sphere.

Ground Truth 2: The large yellow rubber sphere was replaced by a small red rubber sphere. A small purple metal cube shows up. Someone replaced the large brown rubber cylinder with a large blue rubber cube. A small cyan rubber sphere has been added.

Ground Truth 3: A small purple metal cube has been added. A large blue rubber cube replaced the large brown rubber cylinder. A small red rubber sphere is in the original position of large yellow rubber sphere. A small cyan rubber sphere has been added.

Ground Truth 4: Someone added a small purple metal cube. Someone replaced the large yellow rubber sphere with a small red rubber sphere. The large brown rubber cylinder gave up its position to a large blue rubber cube. A small cyan rubber sphere has been added.

Ground Truth 5: The large yellow rubber sphere was replaced by a small red rubber sphere. A large blue rubber cube replaced the large brown rubber cylinder. Someone added a small purple metal cube. A small cyan rubber sphere shows up.

Ground Truth 1: The small gray metal cube was moved from its original location. The small green metal cylinder was replaced by a small cyan rubber cube. The small red metal cube is in a different location. A large red metal cylinder has been added.

Ground Truth 2: Someone changed location of the small gray metal cube. Someone replaced the small green metal cylinder with a small cyan rubber cube. Someone changed location of the small red metal cube. A large red metal cylinder has been added.

Ground Truth 3: Someone changed location of the small red metal cube. Someone replaced the small green metal cylinder with a small cyan rubber cube. A large red metal cylinder shows up. Someone changed location of the small gray metal cube.

Ground Truth 4: The small gray metal cube has been moved. There is a new large red metal cylinder. The small green metal cylinder was replaced by a small cyan rubber cube. Someone changed location of the small red metal cube.

Ground Truth 5: The small red metal cube is in a different location. The small gray metal cube changed its location. Someone replaced the small green metal cylinder with a small cyan rubber cube. There is a new large red metal cylinder.

Figure 12. Four-change examples from the CLEVR-Multi-Change dataset. The changed objects are highlighted by rectangles with the same color as the associated change captions.

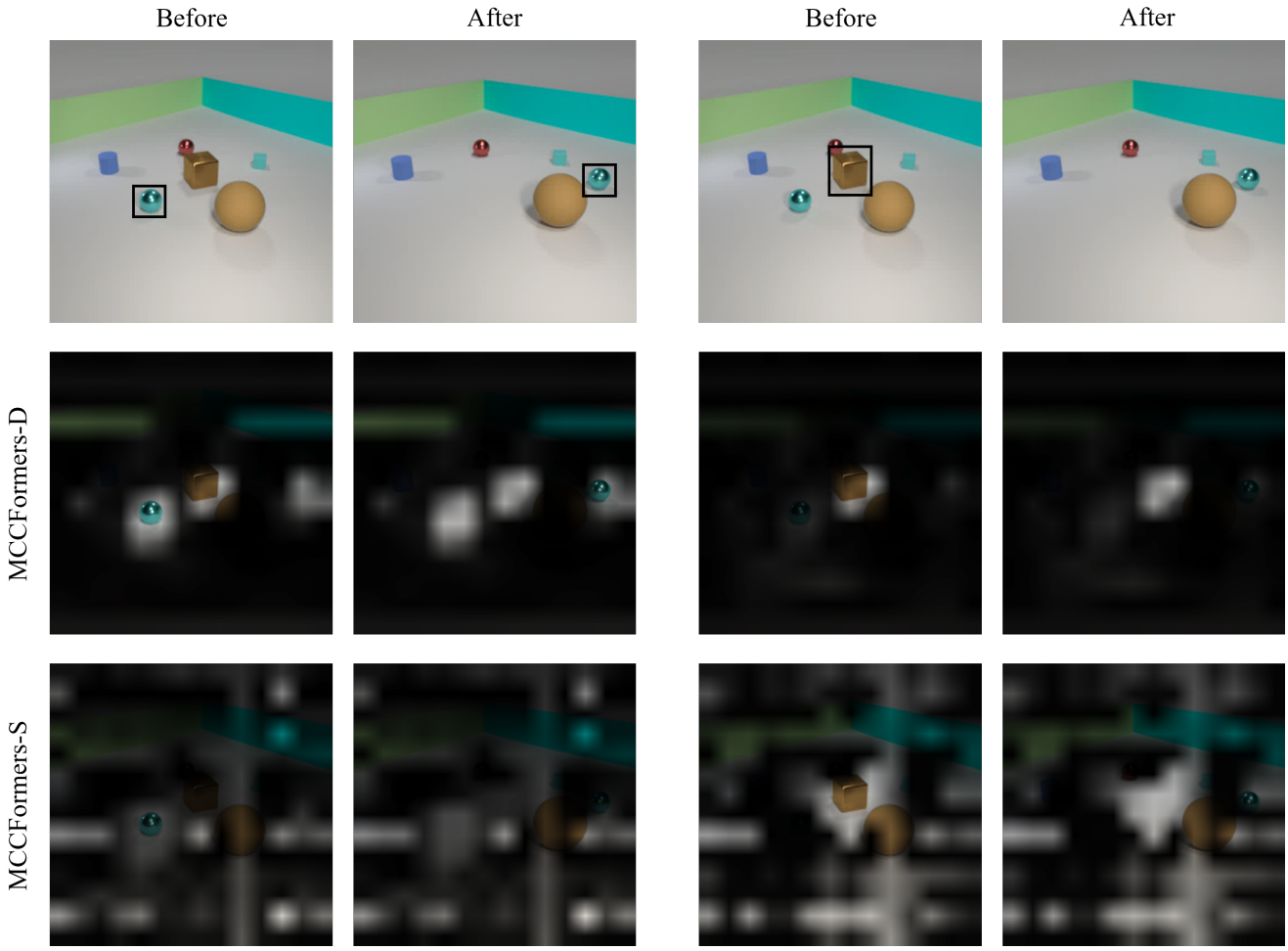
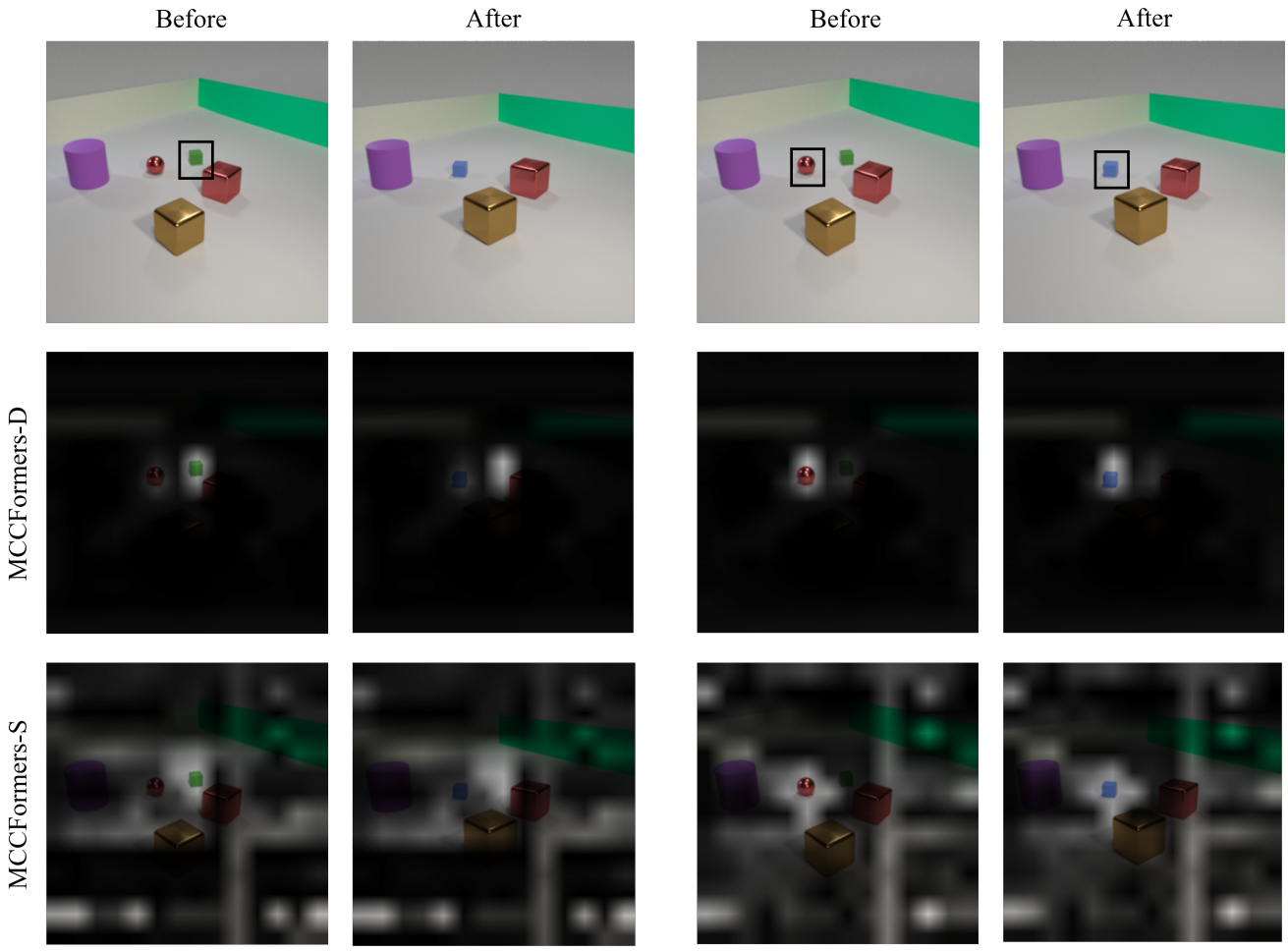


Figure 13. Visualization of an example from the CLEVR-Multi-Change dataset. We highlighted changed regions in black rectangles.



Ground Truth: The small green rubber cube is missing.

MCCFormers-D: The small green rubber cylinder is no longer there.

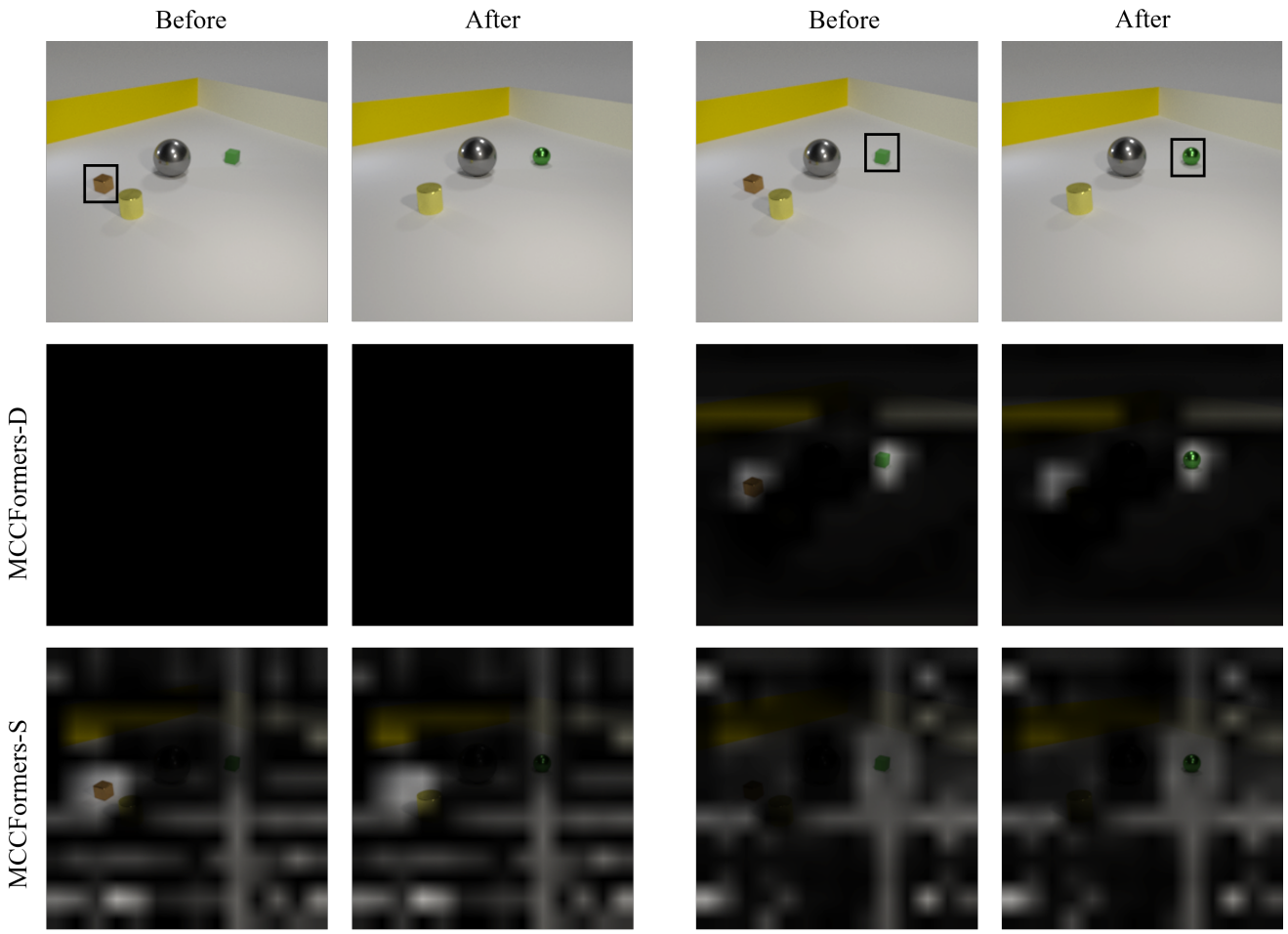
MCCFormers-S: The small green rubber cylinder is missing.

Ground Truth: Someone replaced the small red metal sphere with a small blue rubber cube.

MCCFormers-D: A small blue rubber cube replaced the small red metal sphere.

MCCFormers-S: The small red metal sphere gave up its position to a small blue rubber cube.

Figure 14. Visualization of an example from the CLEVR-Multi-Change dataset. Incorrect captions are in red font. We highlighted changed regions in black rectangles.



Ground Truth: The small brown metal cube has disappeared.

MCCFormers-D: (No sentence was generated.)

MCCFormers-S: The small brown metal cube is missing.

Ground Truth: The small green rubber cube gave up its position to a small green metal sphere.

MCCFormers-D: The small green rubber cube gave up its position to a small green metal sphere.

MCCFormers-S: A small green metal sphere replaced the small green rubber cube.

Figure 15. Visualization of an example from the CLEVR-Multi-Change dataset. Incorrect captions are in red font. We highlighted changed regions in black rectangles.