Supplementary Material:
Robust Change Captioning

In this supplementary material, we provide an analysis of the performance of our Dual Dynamic Attention Model (DUDA) in terms of what change types get confused the most. We also provide additional details on how CLEVR-Change Dataset was collected, especially how change descriptions were generated, and how the data distribution in terms of difficulty measured by IoU looks like given the introduced random jitters in camera position.

1. Confusion Matrix of Change Types

In order to analyze the behavior of our method on different change types, we parse the sentences generated by our model and categorize the type of change that is detected based on the parsed results. We compare that to the ground-truth change type information, and plot the confusion matrix in Figure 1. As we have already shown (Table 3 in the main paper), the most challenging change types are TEXTURE (73% accuracy) and MOVE (45% accuracy), which are most often confused with the DISTRACTOR changes. It is interesting to note that for all change types most of the confusion comes from misidentifying scene changes as DISTRACTORs, and that such confusion is the most severe for MOVE. This is intuitive in the sense that in order to correctly distinguish MOVE from DISTRACTOR, the model has to spatially relate every other object in the scene whereas for other scene change types the changes are relatively salient and do not necessarily require understanding the spatial relationships between the objects. Moreover, MOVE is also confused with ADD and DROP, as it may be difficult to correctly establish a correspondence between all the objects in "before" and "after" scenes. Overall, the substantial amount of confusion with the DISTRACTORs demonstrates the difficulty of our problem statement, as opposed to always assuming that a scene change is present.

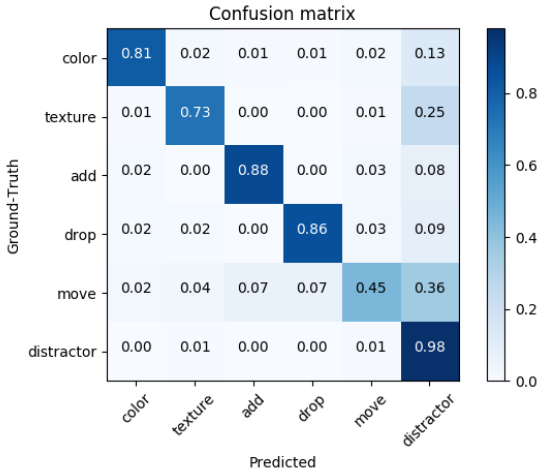

Figure 1: Confusion matrix of DUDA. The horizontal axis indicates the predicted change types of our model whereas the vertical axis indicates the actual change types.

Type	Templates	Type	Templates
COLOR	... changed to ... ... turned ... ... became ...	DROP	... has disappeared. ... is missing. ... is gone. ... is no longer there.
TEXTURE	... changed to ... ... turned ... ... became ...	MOVE	... moved. ... is in a different location. ... changed its location.
ADD	... has appeared. ... has been added. ... has been newly placed.	DISTRACTOR	no change was made. the scene is the same as before. the two scenes seem identical.

Table 1: For each change type we construct a few templates, based on which the change part of the caption is obtained.

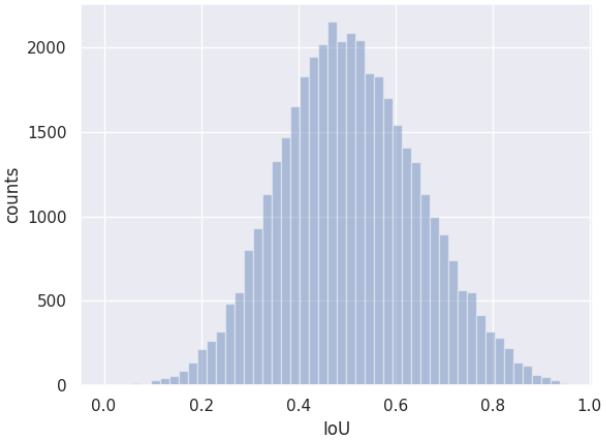

Figure 2: Histogram of CLEVR-Change Dataset based on IoU. The horizontal axis indicates the amount of viewpoint shift measured by IoU whereas the vertical axis indicates the number of data points.

2. Additional Details on CLEVR-Change Dataset

In this section, we provide details on how the captions are generated in our CLEVR-Change Dataset and how the random camera position shifts manifest themselves in the dataset distribution. Having access to all the object information in a CLEVR-rendered scene, we can easily generate multiple different sentences describing a particular change by using templates listed in Table 1. For instance once the images are generated with the desired change (e.g. COLOR), we identify the changed object in the before or after images, and extract its locations and attributes which are used to generate a referring expression (e.g. *the red metallic cube that is to the left of a big sphere*). This phrase is then combined with a randomly selected template followed by a description of how it has changed (i.e. ... *changed to yellow*).

In section 4 of the main paper, it is described that different viewpoint and illumination are introduced via a random shift in camera (x, y, z) location ranging between -2.0 to 2.0 in coordinate points. As a way to understand how this translates to an actual difference between before and after images, we plot a histogram of the entire dataset based on the IoU heuristics explained in section 5.2 of the main paper. As can be seen from Figure 2, the random camera jitters form a reasonable distribution of data points in terms of viewpoint shift difficulty. To show a better sense of what the IoU means, we provide relatively difficult (i.e. low IoU of 0.17 - 0.18) and easy (i.e. high IoU of 0.81 - 0.92) examples in Figure 3. We notice that depending on the viewpoint shift, the task can become significantly difficult even for a simple scene. For instance in the top-left example of Figure 3, where there are only three objects, we see that it becomes hard to localize the changed object as it escapes the scene due to significant camera movement. On the other hand, for a more complex scene like the bottom-left example in Figure 3, localizing change is easier with a small viewpoint shift.

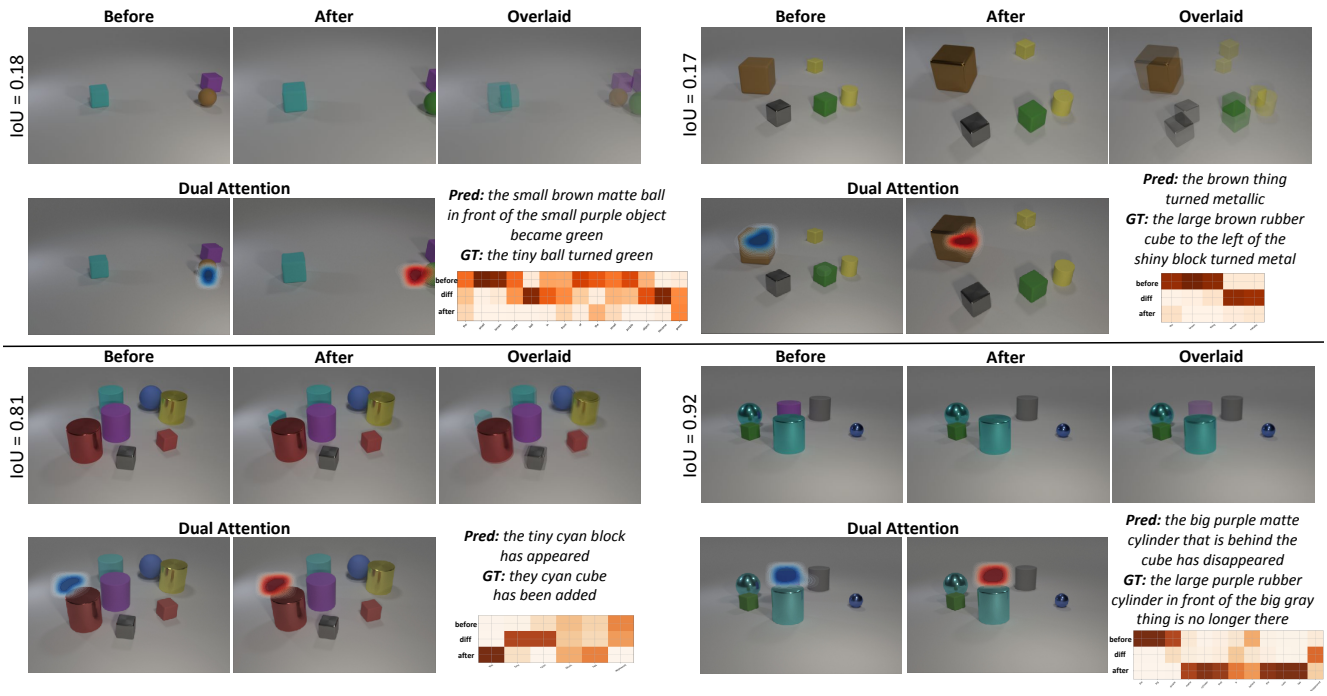

Figure 3: Difficult and easy examples chosen via IoU-based heuristics. The examples at the top are the difficult ones, where the viewpoint shift is noticeable. The examples at the bottom are the easy ones, where the viewpoint change is not significant. We also show the corresponding attention and sentences generated by our model, as well as the ground-truth descriptions.

3. DUDA vs. DDLA [1] on CLEVR-Change

The official implementation of DDLA, the change captioning model proposed in [1], has not been publicly released by the authors yet, so we have reimplemented their model to the best of our understanding. As can be seen in Table 2, DUDA outperforms DDLA in both captioning and localization metrics. We observe that the clustering algorithm used in DDLA becomes unstable when pixels are not aligned between the before and after images, thereby leading to less robustness to viewpoint shit and weaker performance.

	Captioning (Total)				Localization (Total)
Approach	B	C	M	S	Acc
DDLA (reimp.)	44.7	103.3	31.7	22.8	37.16
DUDA (Ours)	**47.3**	**112.3**	**33.9**	**24.5**	**48.10**

Table 2: Evaluation on our CLEVR-Change dataset.

References

[1] Harsh Jhamtani and Taylor Berg-Kirkpatrick. Learning to describe differences between pairs of similar images. 2018. 3