Context-aware Difference Distilling for Multi-change Captioning

Yunbin Tu[1], **Liang Li**[2,5]*, **Li Su**[1]*, **Zheng-Jun Zha**[3],
Chenggang Yan[4,5], **Qingming Huang**[1]
[1]University of Chinese Academy of Sciences, Beijing, China
[2]Key Lab of Intell. Info. Process., Inst. of Comput. Tech., CAS, Beijing, China
[3]University of Science and Technology of China, Hefei, China
[4]Hangzhou Dianzi University, Hangzhou, China
[5]Lishui Institute of Hangzhou Dianzi University, Lishui, China
tuyunbin22@mails.ucas.ac.cn, liang.li@ict.ac.cn, suli@ucas.ac.cn

Abstract

Multi-change captioning aims to describe complex and coupled changes within an image pair in natural language. Compared with single-change captioning, this task requires the model to have higher-level cognition ability to reason an arbitrary number of changes. In this paper, we propose a novel context-aware difference distilling (CARD) network to capture all genuine changes for yielding sentences. Given an image pair, CARD first decouples context features that aggregate all similar/dissimilar semantics, termed common/difference context features. Then, the consistency and independence constraints are designed to guarantee the alignment/discrepancy of common/difference context features. Further, the common context features guide the model to mine locally unchanged features, which are subtracted from the pair to distill locally difference features. Next, the difference context features augment the locally difference features to ensure that all changes are distilled. In this way, we obtain an omni-representation of all changes, which is translated into linguistic sentences by a transformer decoder. Extensive experiments on three public datasets show CARD performs favourably against state-of-the-art methods. The code is available at https://github.com/tuyunbin/CARD.

1 Introduction

Change captioning aims to describe differences between a pair of similar images, which enables many important applications, such as automatic report generation about change conditions of surveillance areas (Hoxha et al., 2022) and pathological changes between medical images (Liu et al., 2021). On the other hand, this task is more challenging than image captioning (Yang et al., 2023; Rotstein et al., 2024; Zhao and Xiong, 2024). This is because machines need to understand the contents of two images simultaneously, and further reason and caption all genuine changes between them, while resisting irrelevant viewpoint/illumination changes.

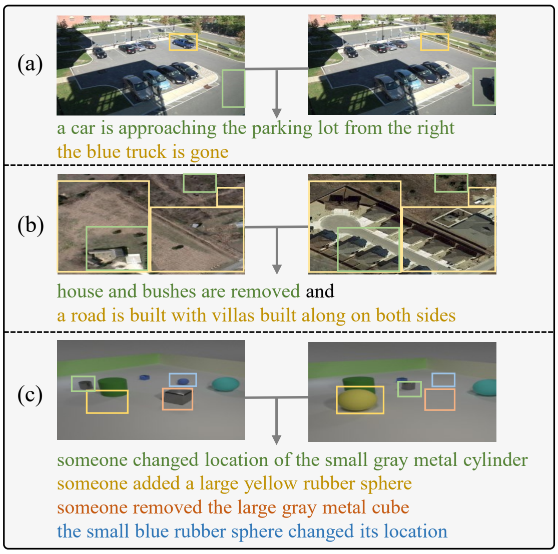

(a)
a car is approaching the parking lot from the right
the blue truck is gone
(b)
house and bushes are removed and
a road is built with villas built along on both sides
(c)
someone changed location of the small gray metal cylinder
someone added a large yellow rubber sphere
someone removed the large gray metal cube
the small blue rubber sphere changed its location

Figure 1: Three examples about multi-change captioning. (a) includes certain object changes; (b) consists of object and background changes; (c) shows both object changes and irrelevant viewpoint change. These changes are shown in colored boxes.

Recently, single-change captioning has made remarkable progress (Tu et al., 2021b; Hosseinzadeh and Wang, 2021; Yao et al., 2022; Yue et al., 2023; Tu et al., 2024). In a dynamic environment, however, the changes are usually the *many-in-one*, where multiple changes exist in an image pair. For instance, there are multiple object/background changes (Figure 1 (a) (b)). In other cases, object and viewpoint changes simultaneously appear (Figure 1 (c)). In above cases, unchanged objects commonly mingle with changed ones and even appear position misalignment under viewpoint changes. Such distractors pose a great challenge to identify and caption the genuine changes.

There are a few attempts to address multi-change captioning. The pioneer work (Jhamtani and Berg-

*Corresponding authors

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 7941–7956
August 11-16, 2024 ©2024 Association for Computational Linguistics

Kirkpatrick, 2018) computed pixel differences of two images, which is sensitive to noise. Latest works tried to capture differences at representation space: some of them (Hoxha et al., 2022; Liu et al., 2022, 2023a) computed difference features by subtraction, while the others (Qiu et al., 2021; Chang and Ghamisi, 2023) built the correlations between the patches of two images to model the change features for caption generation.

Despite progress, the above endeavors in multi-change captioning have several limitations. (1) Direct subtraction between two images generalizes poorly to unaligned image pairs under viewpoint changes (Figure 1 (c)). (2) Directly correlating two images fails to sufficiently mine locally unchanged features as multiple objects change, because such features might mingle with the features of changed objects. (3) These methods focus on modeling locally difference features, which are useful to catch conspicuous changes. Nevertheless, certain local changes with weak features might be overlooked, *e.g.,* the car occluded by its shadows in Figure 1 (a). These limitations would result in obtaining unreliable difference features for the language decoder.

We notice that the above methods capture differences between two images only based on local features, while neglecting the use of more comprehensive features. We argue that, to learn locally unchanged/changed features of two images, the model should first encapsulate their context features of commonality and difference. Such context features aggregate all similar/dissimilar semantics, termed *common/difference context features*. The former can help correlate and mine locally common features for deducing locally difference features, while the latter can augment the locally difference features to ensure all changes are distilled.

In this paper, we propose a **C**ontext-**A**ware Diffe**R**ence **D**istilling (CARD) network to learn the robust difference features under multi-change scenes. Specifically, given the featuers of two images, we first build intra-image interaction to help the model understand each image content of the pair. Then, we use CARD to decouple the common/difference context features from the image pair. Herein, the common context features of two images summarize joint semantics in between; the difference context feature in each image provides an independent space to preserve its all changed semantics. Besides, the consistency and independence constraints are designed to enforce the alignment and discrepancy of common and difference context features, respectively. Next, guided by the common context features, CARD models inter-image interaction to mine locally common features, which are removed from the pair to distill locally difference features. Subsequently, CARD augments the locally difference features via the difference context features, so as to construct an omni-representation of all changes, for generating descriptions by a transformer decoder.

Our key contributions are: **(1)** We propose CARD to first decouple common and difference context features, and then use them to facilitate modeling an omni-representation of all changes for multi-change captioning. **(2)** The consistency and independence constraints are customized to guarantee the alignment and discrepancy of decoupled common and difference context features. **(3)** Extensive experiments show our method achieves the state-of-the-art results on three public datasets.

2 Related Work

Change captioning is an emerging task in the community of multi-modal learning (Cong et al., 2022, 2023; Tu et al., 2022). In the following, we introduce the relevant works about single-change captioning and multi-change captioning, respectively.

Single-change Captioning has been widely studied by most existing methods. The prior work (Park et al., 2019) collects a dataset about geometric objects under viewpoint changes. This work computes the difference representation by direct subtraction, which generalizes poorly between two unaligned images. To remedy this limitation, M-VAM (Shi et al., 2020), VACC (Kim et al., 2021) and R^{3}Net (Tu et al., 2021a) match local features to predict difference features, which has been a classic paradigm. The latest work SCORER+CBR (Tu et al., 2023c) further improves this paradigm by maximizing cross-view contrastive alignment between two images, so as to learn a more stable difference representation. In addition, these works (Hosseinzadeh and Wang, 2021; Kim et al., 2021; Tu et al., 2023c; Yue et al., 2024) introduce the idea of cross-modal consistency constraint to improve captioning quality. Besides improving architecture, recent works (Yao et al., 2022; Guo et al., 2022) propose the strategy of pre-training to fine-tuning for facilitating change location and caption. However, it is seldom that only one change appears in a dynamic environment, so a powerful model should has the capability to describe multiple changes.

Multi-change Captioning has been explored by a few attempts. The pioneer work (Jhamtani and Berg-Kirkpatrick, 2018) proposes to describe multiple changes between the image pairs from the surveillance cameras, where it captures changes at pixel level. Recent works (Hoxha et al., 2022; Liu et al., 2022, 2023a; Chang and Ghamisi, 2023) propose to caption changes between remote sensing images, where they first compute the difference features and then use them for change detection and description. Nevertheless, the above methods only describe differences between two well-aligned images, while ignoring the cases of unaligned image pairs under varied viewpoints. To this end, Qiu *et al.* (Qiu et al., 2021) collect a dataset to caption multiple changes under viewpoint changes, where they adopt the classic paradigm of local feature matching to detect changes for caption generation. However, such a matching paradigm may fail to mine the fine-grained common features and thus learn error-prone difference features. Meanwhile, the current methods focus on modeling local difference features, which risks ignoring certain local changes with weak features. On the other hand, Qiu *et al.* (Qiu et al., 2023) develop a new synthetic dataset to describe the multiple changes and their orders. However, recording the order of changed objects is laborious in the real world, making it hard to test such a capability in a real-world scene.

In short, different from previous methods computing difference features only based on local features, CARD first decouples common and difference context features from an image pair. The common context features guide the model to fully extract locally unchanged features for computing the features of local differences, while the difference context features augment the locally difference features to construct an omni-representation of all changes, for generating accurate sentences.

3 Methodology

The overall architecture of our method is shown in Figure 2. Given a pair of images, our method is to generate linguistic sentences that detail all of the changes. Architecture-wise, our method contains three components: (a) image pair encoding; (b) context-aware difference distilling; (c) caption generation. We provide an overview of two basic components (a) and (c) in Sec. 3.1 and Sec. 3.3, while elaborating our key ingredient (b) in Sec. 3.2.

3.1 Image Pair Encoding

Formally, given a pair of "before" I_{bef} and "after" I_{aft} images, we first leverage an off-the-shell encoder (*e.g.,* ResNet-101 (He et al., 2016)) to extract n local features for each of them, and then introduce a trainable [CLS] feature to represent the global content of each image: $X_o = \{x_{cls}^o, x_1, x_2, ..., x_n\}$, where $o \in (bef, aft)$ and $x_i \in \mathbb{R}^d$. Besides, the trainable position encodings are added into the features of each image to help the model perceive the relative position changes of objects. Next, we exploit a multi-head self-attention layer (Vaswani et al., 2017) to capture the relationships among the features of each image, which helps the model sufficiently understand the image content of the pair. Through the above manner, we can obtain the relation-embedded features for each image, denoted as $\tilde{X}_o = \{\tilde{x}_{cls}^o, \tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n\}$.

3.2 Context-Aware Difference Distilling

3.2.1 Context Feature Decoupling

To obtain common and difference context features, we first devise a common encoder $\mathcal{CE}\left(\cdot;\theta_{\mathcal{C}}\right)$ and two difference encoders $\mathcal{DE}_o\left(\cdot;\theta_o\right)$, where three encoders are based on the linear projection and $o \in (bef, aft)$. The common encoder $\mathcal{CE}\left(\cdot;\theta_{\mathcal{C}}\right)$ shares the parameters $\theta_{\mathcal{C}}$ between two images, while the difference encoders $\mathcal{DE}_o\left(\cdot;\theta_o\right)$ learn the parameters θ_o for each image. Then, we feed $\tilde{x}_{cls}^o$ into these encoders to decouple the common and difference context features, respectively:

$$\begin{aligned} C_o &= \mathcal{CE}\left(\tilde{x}_{cls}^o;\theta_{\mathcal{C}}\right), \\ D_o &= \mathcal{DE}_o\left(\tilde{x}_{cls}^o;\theta_o\right), \end{aligned} \tag{1}$$

where C_{bef}, C_{aft}, D_{bef}, and $D_{aft} \in \mathbb{R}^d$.

Consistency Constraint. To make two common context features C_{bef} and C_{aft} embedded in a shared space, we tailor the consistency constraint based on contrastive learning. Given a training batch, we sample B pairs of common context features. For the common context feature in the k-th "before" image C_k^{bef}, the common context feature in the r-th "after" image $C_{r(r=k)}^{aft^+}$ is its positive, while common context features in the other "after" images $C_{r(r\neq k)}^{aft^-}$ will be the negatives in this batch. Then, we project these positive/negative pairs into a shared embedding space, normalize them by L_2-normalization, and compute their similarity. Next, we introduce the InfoNCE loss (Oord et al., 2018) to optimize their contrastive alignment, *i.e.,* pulling

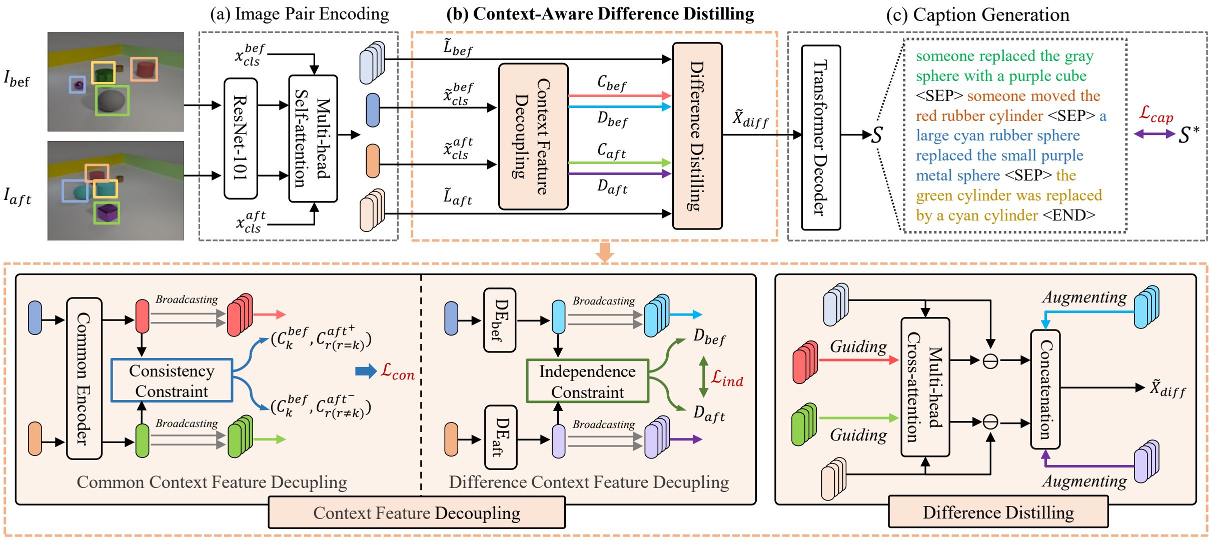


Figure 2: The overall architecture of our method, which consists of (a) Image Pair Encoding (Sec. 3.1), (b) **C**ontext-**A**ware Diffe**R**ence **D**istilling (CARD) (Sec. 3.2), and (c) Caption Generation (Sec. 3.3). Herein, CARD is the major component to learn the robust difference features by context features decoupling and context-aware difference distilling. S^* stands for ground-truth sentences.

semantically close pairs of C_k^{bef} and $C_{r(r=k)}^{aft^+}$ together and pushing away non-related pairs:

$$\mathcal{L}_{b2a} = -\frac{1}{B}\sum_{k}^{B} \log \frac{e^{\left(\text{sim}\left(C_k^{bef}, C_{r(r=k)}^{aft^+}\right)/\tau\right)}}{\sum_r^B e^{\left(\text{sim}\left(C_k^{bef}, C_r^{aft}\right)/\tau\right)}},$$

$$\mathcal{L}_{a2b} = -\frac{1}{B}\sum_{k}^{B} \log \frac{e^{\left(\text{sim}\left(C_k^{aft}, C_{r(r=k)}^{bef^+}\right)/\tau\right)}}{\sum_r^B e^{\left(\text{sim}\left(C_k^{aft}, C_r^{bef}\right)/\tau\right)}},$$

$$\mathcal{L}_{con} = \frac{1}{2}(\mathcal{L}_{b2a} + \mathcal{L}_{a2b}), \tag{2}$$

where “sim” is the dot-product function to measure the similarity between two context features. τ is the temperature hyper-parameter. Through this consistency constraint, we enforce the common context features of two images to be projected into a shared semantic space with aligned distributions.

Independence Constraint. Each decoupled difference context feature can learn the unique characteristics of its corresponding image within the image pair. These unique characteristics represent the semantic differences between the two images, so each difference context feature should be distinct from the other. To this end, we design an independence constraint, which ensures that the two difference context features are projected into separate feature spaces. Here, we opt for the Hilbert-Schmidt Independence Criterion (HSIC) (Song et al., 2007), a proven method for testing feature independence, to design the loss of the independence constraint. A lower HSIC score between the two difference context features indicates a higher independence between them: each difference context feature adequately encapsulates the semantic changes in each image.

Concretely, we first project the difference context feature of each image into a separate space, normalize each by L_2-normalization, and define the independence (HSIC) constraint between D_{bef} and D_{aft} as:

$$\text{HSIC}\left(D_{bef}, D_{aft}\right) = (B-1)^{-2}\,\text{tr}\left(PK_{bef}PK_{aft}\right), \tag{3}$$

where K_{bef}, $K_{aft} \in \mathbb{R}^{B\times B}$ are the Gaussian kernel matrices with $k_{bef,ij} = k_{bef}\left(D_{bef}^i, D_{bef}^j\right)$ and $k_{aft,ij} = k_{aft}\left(D_{aft}^i, D_{aft}^j\right)$. B is batch size. D_{bef}^i refers to the difference context feature in the i-th “befor” image. $P = \mathbf{I} - \frac{1}{B}ee^T$, where $P \in \mathbb{R}^{B\times B}$, $\mathbf{I}$ is an identity matrix and e is an all-one column vector. If the HSIC score between D_{bef} and D_{aft} is lower, their disparity is more significant. We define the independence loss as:

$$\mathcal{L}_{ind} = \text{HSIC}\left(D_{bef}, D_{aft}\right). \tag{4}$$

3.2.2 Difference Distilling

With the local features of each image $\tilde{L}_o = \{\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n\}$, where $o \in (bef, aft)$, we first broadcast the common context feature of each image $C_o \in \mathbb{R}^d$ to $C_o \in \mathbb{R}^{n \times d}$. Then, we concatenate it with $\tilde{L}_o$ on the channel dimension to obtain $\tilde{X}_o' \in \mathbb{R}^{n \times 2d}$, where C_o can guide inter-image interaction to mine locally unchanged features. Next, we transform $\tilde{X}_o' \in \mathbb{R}^{n \times 2d}$ to $\tilde{X}_o' \in \mathbb{R}^{n \times d}$ by a non-linear function with ReLU activation. Further, we compute the locally common features on each image by the multi-head cross-attention (MHCA) mechanism (Vaswani et al., 2017):

$$\begin{aligned} \tilde{X}_{bef}^c &= \text{MHCA}\left(\tilde{X'}_{bef}, \tilde{X'}_{aft}, \tilde{X'}_{aft}\right), \\ \tilde{X}_{aft}^c &= \text{MHCA}\left(\tilde{X'}_{aft}, \tilde{X'}_{bef}, \tilde{X'}_{bef}\right). \end{aligned} \tag{5}$$

Subsequently, we respectively subtract each $\tilde{X}_o^c$ from $\tilde{L}_o$ to compute the locally difference features of each image. These locally difference features are further augmented by difference context feature of each image, so as to distill all of the genuine changes in each image:

$$\begin{aligned} \tilde{X}_{bef}^d &= [\tilde{L}_{bef} - \tilde{X}_{bef}^c; D_{bef}], \\ \tilde{X}_{aft}^d &= [\tilde{L}_{aft} - \tilde{X}_{aft}^c; D_{aft}], \end{aligned} \tag{6}$$

where [;] is a concatenation operation. Both $\tilde{X}_{bef}^d$ and $\tilde{X}_{aft}^d$ are then concatenated as an omni-representation of all changes between two images, which is implemented by a non-linear transformation with the ReLU function:

$$\tilde{X}_d = \text{ReLU}\left(\left[\tilde{X}_{bef}^d; \tilde{X}_{aft}^d\right] W_c + b_c\right). \tag{7}$$

3.3 Caption Generation

After obtaining the omni-representation $\tilde{X}_d \in \mathbb{R}^{hw \times d}$, we use a transformer decoder (Vaswani et al., 2017) to decode it into sentences. First, we obtain the embedding features of all m words of these sentences, where each sentence is separated by a special token [SEP]. Then, we use the masked self-attention to model relationships among these word features. Next, we model the interaction between the word features and omni-representation by cross-attention, so as to locate the most related difference features during word generation. Subsequently, we feed the selected features into a feed-forward network to obtain the enhanced difference representation, denoted as $\hat{H} \in \mathbb{R}^{m \times d}$. Finally, the probability distributions of words in these sentences are calculated via a single hidden layer:

$$S = \text{Softmax}\left(\hat{H} W_s + b_s\right), \tag{8}$$

where $W_s \in \mathbb{R}^{d \times u}$ and $b_s \in \mathbb{R}^u$ are the learnable parameters. u is the dimension of vocabulary size.

3.4 Joint Training

Our method is trained in an end-to-end manner by maximizing the likelihood of the observed word sequence. Given the ground-truth words $(s_1^*, \ldots, s_m^*)$, we minimize the negative log-likelihood loss:

$$\mathcal{L}_{cap}(\theta) = -\sum_{t=1}^{m} \log p_\theta \left(s_t^* \mid s_{<t}^*\right), \tag{9}$$

where $p_\theta \left(s_t^* \mid s_{<t}^*\right)$ is computed by Eq. (8), and θ are all the learnable parameters. Our method is also self-supervised by the losses of consistency and independence constraints. Hence, the total loss is optimized as follows:

$$\mathcal{L} = \mathcal{L}_{cap} + \lambda_c(\mathcal{L}_{con} + \mathcal{L}_{ind}), \tag{10}$$

where λ_c is a trade-off parameter to balance the contribution between the caption generator and constraints, which is discussed in the appendix.

4 Experiments

4.1 Datasets

CLEVR-Multi-Change Dataset (Qiu et al., 2021) is about basic object scene with multiple changes. Since original dataset has not been released, we regenerate this dataset based on the authors' released code. The regenerated dataset has 45,044 valid image pairs/captions with viewpoint changes. Based on the official split, we split it into training, validation, and testing with a ratio of 4:1:1.

LEVIR-CC Dataset (Liu et al., 2022) is about remote sensing scene, which contains 10,077 pairs of bi-temporal images and 50,385 ground-truth captions. We use the official split with 6,815 image pairs for training, 1,333 for validation, and 1,929 for testing, respectively.

Spot-the-Diff Dataset (Jhamtani and Berg-Kirkpatrick, 2018) has 13,192 image pairs from surveillance cameras, and on an average there are 1.86 ground-truth sentences per image pair. According to the official split, we split it into training, validation, and testing with a ratio of 8:1:1.

4.2 Evaluation Metrics

We follow the existing methods of multi-change captioning to use the five metrics for evaluating the generated sentences: BLEU-4 (B) (Papineni et al., 2002), METEOR (M) (Banerjee and Lavie, 2005), ROUGE-L (R) (Lin, 2004), CIDEr (C) (Vedantam et al., 2015) and SPICE (S) (Anderson et al., 2016).

We compute all the results by the Microsoft COCO evaluation server (Chen et al., 2015).

4.3 Implementation Details

For fair-comparison, we follow previous multi-change captioning methods to use a pre-trained ResNet-101 (He et al., 2016) to extract the local features of a pair of images, with the dimension of 1024 × 14 × 14. We project them into a lower dimension of 512, while the dimension of trainable [CLS] features is also set to 512. The hidden size of the model and word embedding size are set to 512 and 300. Temperature τ in Eq. (2) is set to 0.07. We train the model to converge with 10K iterations in total. We use Adam optimizer (Kingma and Ba, 2014) to minimize the negative log-likelihood loss of Eq. (10). More details are shown in the appendix.

4.4 Performance Comparison

Results on CLEVR-Multi-Change. We compare CARD with the following SOTA methods: DUDA (Park et al., 2019), M-VAM (Shi et al., 2020), MCCFormers-S / MCCFormers-D (Qiu et al., 2021), VARD-Trans (Tu et al., 2023a), and SCORER+CBR (Tu et al., 2023c). On this regenerated dataset, we re-implement the above methods based on their papers and released codes.

The results are shown in Table 1. Our CARD performs favourably against these SOTA methods on all metrics, showing that CARD can better describe multiple changes under viewpoint changes. In addition, our CARD outperforms MCCFormers-D and MCCFormers-S by a large margin, which are classic match-based methods to directly capture inner/inter-patch correlations between two image representations. On the caption-specific metric CIDEr, CARD significantly surpasses both methods, in particular with increases of 2.2% and 3.4%.

Table 1: Comparison with the SOTA methods on CLEVR-Multi-Change. The main metric CIDEr on this dataset is highlighted.

Method	B	M	R	S	C
DUDA	41.8	36.2	53.9	64.7	283.5
M-VAM	37.1	34.0	51.5	62.2	242.9
MCCFormers-S	55.9	44.8	56.8	76.6	378.6
MCCFormers-D	56.2	44.8	57.3	76.6	383.2
VARD-Trans	48.1	41.8	55.5	72.1	344.2
SCORER+CBR	56.4	44.9	57.1	76.7	388.0
CARD (Ours)	**56.7**	**45.2**	**57.4**	**76.9**	**391.6**

Table 2: Comparison with the SOTA methods on LEVIR-CC.

Method	B	M	R	C
DUDA	57.8	37.2	71.0	124.3
MCCFormers-S	56.7	36.2	69.5	120.4
MCCFormers-D	56.4	37.3	70.3	124.4
RSICCFormer	62.8	39.6	74.1	134.1
PSNet	62.1	38.8	73.6	132.6
Prompt-CC (soft)	62.4	38.6	73.4	135.3
Prompt-CC (hard)	63.5	38.8	73.7	136.4
Chg2Cap	64.4	**40.0**	**75.1**	136.6
CARD (Ours)	**65.4**	**40.0**	74.6	**137.9**

Table 3: Comparison with the SOTA methods on Spot-the-Diff.

Method	B	M	R	S	C
DDLA	6.2	10.8	26.0	-	29.7
DUDA	5.4	10.6	-	12.9	24.8
MCCFormers-S	5.8	10.5	-	10.1	18.2
MCCFormers-D	6.2	10.2	-	**17.8**	28.8
VARD-Trans	4.1	**11.4**	22.2	11.5	13.4
SCORER+CBR	5.1	9.3	23.0	11.9	20.9
CARD (Ours)	**6.6**	10.8	**26.9**	**17.8**	**32.4**

Results on LEVIR-CC. We compare CARD with the SOTA methods: DUDA (Park et al., 2019), MCCFormers-S/D (Qiu et al., 2021), RSICCFormer (Liu et al., 2022), PSNet (Liu et al., 2023a), Prompt-CC (soft/hard) (Liu et al., 2023b), and Chg2Cap (Chang and Ghamisi, 2023).

The experimental results are shown in Table 2. Our CARD achieves the best results on all metrics. This indicates that it can detect whether there are semantic changes and what have changed between two remote sensing images. Besides, we notice that the match-based methods (MCCFormers-D / MCCFormers-S) cannot generalize well in this remote sensing scenario. Our conjecture is that there are usually most changed areas (*e.g.,* Figure 1 (b)), so directly matching two images might fail to extract fine-grained unchanged objects and thus capture the difference features with noise.

Results on the Spot-the-Diff Dataset. Most previous works (Hosseinzadeh and Wang, 2021; Huang et al., 2022; Tu et al., 2023b; Yue et al., 2023) tested the models based on single-change setup, where the models are only required to randomly describe one of the changes. Different from them, we require the model to caption all changes.

Under this setting, the compared SOTA methods are: DDLA (Jhamtani and Berg-Kirkpatrick, 2018), DUDA (Park et al., 2019), MCCFormers-S

Table 4: Ablation of common/difference context features (CCF/DCF) on CLEVR-Multi-Change and LEVIR-CC.

			CLEVR-Multi-Change					LEVIR-CC			
Ablative Variants	CCF	DCF	B	M	R	S	C	B	M	R	C
Baseline	×	×	54.7	43.6	56.7	75.6	362.3	60.7	36.3	69.7	120.0
Baseline	✓	×	56.5	45.1	57.1	76.8	385.8	63.5	38.5	72.3	130.4
Baseline	×	✓	56.5	45.0	57.1	**77.0**	385.7	60.6	37.6	71.0	125.9
Baseline	✓	✓	**56.7**	**45.2**	**57.4**	76.9	**391.6**	**65.4**	**40.0**	**74.6**	**137.9**

Table 5: Ablation of consistency/independence constraint (CC/IC) on CLEVR-Multi-Change and LEVIR-CC.

			CLEVR-Multi-Change					LEVIR-CC			
Ablative Variants	CC	IC	B	M	R	S	C	B	M	R	C
CARD	×	×	54.6	44.1	57.2	75.8	363.7	55.9	35.6	72.3	132.2
CARD	✓	×	56.2	44.8	57.1	76.8	384.2	56.2	35.8	72.6	137.6
CARD	×	✓	56.5	45.1	57.2	**77.0**	389.9	60.6	37.7	72.5	133.0
CARD	✓	✓	**56.7**	**45.2**	**57.4**	76.9	**391.6**	**65.4**	**40.0**	**74.6**	**137.9**

/ MCCFormers-D (Qiu et al., 2021), VARD-Trans (Tu et al., 2023a), and SCORER+CBR (Tu et al., 2023c). The experimental results are shown in Table 3. Our method achieves the best results on most metrics, particularly with an increase of 9.1% on CIDEr. As shown in Figure 1 (a), the changed objects are not well captured by surveillance cameras and are even occluded by shadows. Our method still achieves encouraging performance, which validates its good generalization in surveillance scenes.

Performance Analysis. On the three datasets, our CARD outperforms the existing methods by a large margin, which shows its good generalization. This superiority benefits from that decoupled context features facilitate learning difference features. Instead, the compared methods only compute the differences based on the local features, which fails to mine fine-grained common objects and thus disregards inconspicuous changes.

4.5 Ablation Study and Analysis

To figure out the contribution of each component, we conduct ablation studies on two large-scale datasets: CLEVR-Multi-Change and LEVIR-CC. The image pairs on CLEVR-Multi-Change contain basic geometric objects and are unaligned due to viewpoint changes, while the pairs on LEVIR-CC are well-aligned and from the real world.

4.5.1 Ablation Study for Context Features

We study the effectiveness of decoupled common and difference context features, denoted as CCF and DCF, respectively. The results are shown in Table 4. Baseline directly matches two image features to extract the locally common features and difference features for caption generation.

We find that on the both datasets, 1) the model's performance is enhanced when it is augmented by either CCF or DCF; 2) when we augment the model with both context features, the model's performance is significantly boosted, especially the CIDEr score is enhanced from 362.3 to 391.6. These show that 1) CCF guides the model to sufficiently interact and mine locally common features for computing locally difference features, while DCF augments the locally difference features to ensure all changes are distilled; 2) each kind of context feature not only plays its unique role, but also supplements each other for better reasoning genuine changes.

4.5.2 Ablation Study for Constrains

To study the effect of the consistency constraint (CC) and independence constraint (IC), we make the ablation study in Table 5. First, we enforce CC or IC respectively on CARD, and find that each of them improves the performance of CARD. Then, we impose both constraints and observe that the results are enhanced significantly on two datasets. The increased results indicate that both constraint losses are essential to learn the context features.

We further visualize the common/difference context features decoupled without/with the constraints, which are shown in Figure 3. Without constraints, the common/difference context features cannot be well learned on two datasets. With both constraints, the common context features are blended on CLEVR-Multi-Change, while the difference context features are learned better on LEVIR-CC. The results show that consistency constraint

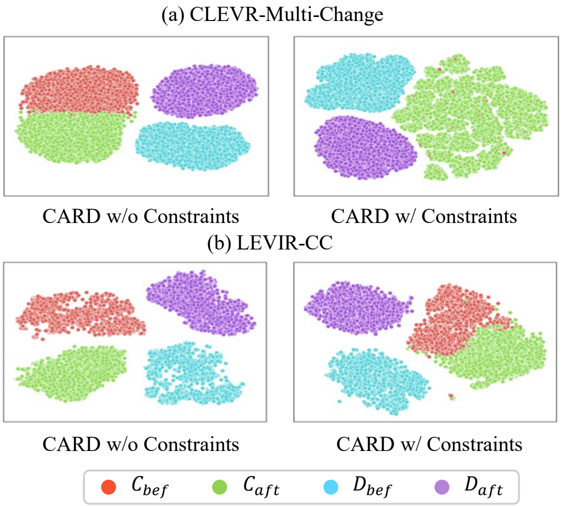


Figure 3: Visualization of context features on CLEVR-Multi-Change and LEVIR-CC. The red and green colors indicate common context features in "before" and "after" images, while blue and purple colors denote difference context features in "before" and "after" images.

helps align distributions of shared properties, while independence constraint makes difference context features of two images more separable.

4.5.3 Generalization to unaligned image pairs

To verify the generalization to unaligned image pairs under viewpoint changes, we compare CARD with two ablative variants on the CLEVR-Multi-Change dataset. (1) Direct Subtraction first performs direct subtraction between the features of two images to compute the locally difference features, which are then fed into a transformer decoder for multi-change captioning. (2) Feature Matching directly matches two image features to extract the locally common features and difference features for multi-change captioning. The experimental results are shown in Table 6.

Table 6: Verifying generalization to unaligned image pairs on CLEVR-Multi-Change.

Model	B	M	R	C	S
Direct Subtraction	53.3	42.5	56.3	350.0	74.6
Feature Matching	54.7	43.6	56.7	362.3	75.6
CARD	**56.7**	**45.2**	**57.4**	**391.6**	**76.9**

It is noted that the performance of Feature Matching is better than that of Direct Subtraction on every metric, which validates the generalization of feature matching paradigm to the unaligned image pairs under viewpoint changes. Our proposed CARD outperforms both models by a large margin. This not only indicates a better generalization of our method to unaligned image pairs, but also verifies the effectiveness of context-aware difference distilling to help capture genuine changes.

4.6 Qualitative Analysis

To obtain an overall evaluation of our method, we conduct qualitative analysis on the three datasets. The compared method MCCFormers-D (Qiu et al., 2021) performs well on the three datasets, which directly correlates two images to predict locally common and difference features. In Figure 5, we visualize the alignment of common properties between two images, to validate whether context features help mine locally common information. We find that MCCFormers-D fails to align the common properties and even misjudges changed objects as unchanged objects. Instead, our CARD can better match the joint objects. For instance, in Figure 5 (a), the unchanged object is only the brown object. MCCFormers-D wrongly identifies some changed objects as unchanged ones. By contrast, our CARD can pinpoint the unchanged brown object. This superiority benefits from the guidance of decoupled common context features, during the matching of two image features.

In Figure 4, we further visualize the captions yielded by MCCFormers-D and CARD, as well as the change localization results from CARD. In the first three cases, MCCFormers-D either only describes one of the changes or misidentifies changed objects. Contrarily, our CARD can accurately locate and describe all changed objects. Particularly, we notice that our method performs better in detecting subtle changes. For instance, in Figure 4 (c) that is from surveillance scene, the moved car and disappeared person are very tiny, and the car is occluded by the shadow of building. In this hard case, MCCFormers-D fails to recognize the moved car in the "after" image, thus generating a wrong sentence. By contrary, our CARD can locate and describe this tiny change. For the failure reason of MCCFormers-D, our conjecture is that it directly interacts two images, which cannot sufficiently identify locally unchanged features and compute the locally difference features. Besides, MCCFormers-D captures differences between two images only based on local features, which risks overlooking certain tiny changes with weak features. Compared with MCCFormers-D, the superiority of our method is mainly attributed to the

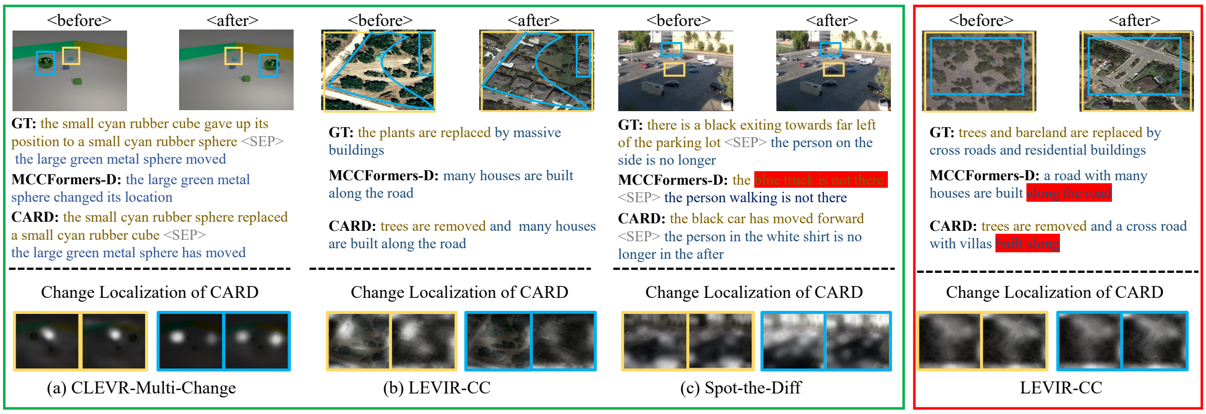

<before>
<after>
GT: the small cyan rubber cube gave up its position to a small cyan rubber sphere <SEP> the large green metal sphere moved
MCCFormers-D: the large green metal sphere changed its location
CARD: the small cyan rubber sphere replaced a small cyan rubber cube <SEP> the large green metal sphere has moved
GT: the plants are replaced by massive buildings
MCCFormers-D: many houses are built along the road
CARD: trees are removed and many houses are built along the road
GT: there is a black exiting towards far left of the parking lot <SEP> the person on the side is no longer
MCCFormers-D: the blue truck is not there <SEP> the person walking is not there
CARD: the black car has moved forward <SEP> the person in the white shirt is no longer in the after
GT: trees and bareland are replaced by cross roads and residential buildings
MCCFormers-D: a road with many houses are built along the road
CARD: trees are removed and a cross road with villas built along
Change Localization of CARD
(a) CLEVR-Multi-Change
(b) LEVIR-CC
(c) Spot-the-Diff
LEVIR-CC

Figure 4: Qualitative examples on the three datasets. For each example, we visualize the captions generated by the SOTA method MCCFormers-D (Qiu et al., 2021) and our CARD, as well as the change localization of CARD. The successful cases of CARD are shown in the green box, while the sub-optimal case is shown in the red box.

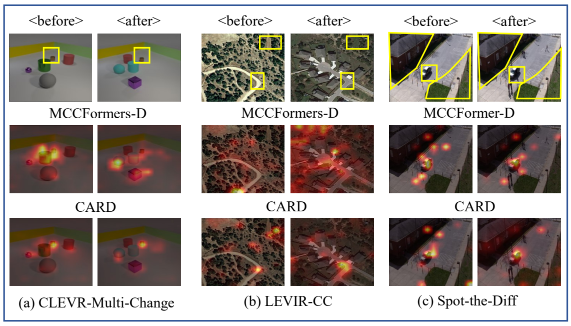


Figure 5: Visualization of alignment of common objects (shown in yellow boxes) on the three datasets, where the results are obtained by MCCFormers-D and our CARD.

guiding and augmenting of common and difference context features. Guided by the common context features, CARD models inter-image interaction to sufficiently mine locally common features and compute locally difference features. Further, the difference context features augment the locally difference features to ensure that all changes are distilled. Through this manner, the model can learn genuine changes for caption generation. More qualitative examples are shown in the appendix.

5 Conclusion

In this paper, we propose the CARD to reason and describe genuine changes under various multi-change scenarios. CARD first decouples the common and difference context features from the image pair. Then, two kinds of constraints are designed to ensure the alignment and discrepancy of the common and difference context features, respectively. Further, we use the common context features to guide the mining of locally common features for deducing locally difference features. In addition, we leverage the difference context features to augment the locally difference features, thereby constructing an omni-representation of all changes for multi-change captioning. Extensive experiments conducted on the three datasets show that the proposed CARD outperforms the current state-of-the-art methods by a large margin.

Limitations

The last case in Figure 4 shows that the trees are replaced by a cross road with villas. Our CARD successfully locates the changed objects, which validates the effectiveness of context-aware difference distilling. However, it yields a sub-optimal sentence that does not well express the change process: *trees are removed and a cross road with villas built along*. A more proper sentence should be: *trees are removed and a cross road with villas are built*. In the future work, we will try to introduce linguistic knowledge (*e.g.*, syntactic dependencies between words) that regularizes the process of sentence generation, in order to generate the optimal sentences well elaborating the change process.

Acknowledgements

This work was supported in part by National Key R&D Program of China under Grant (2023YFB4502800), National Natural Science Foundation of China: 62322211, 61931008, 62236008, 62336008, U21B2038, 62225207, Fundamental Research Funds for the Central Universities (E2ET1104), “Pionee” and “Leading Goose” R&D Program of Zhejiang Province (2024C01023, 2023C01030).

References

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *ECCV*, pages 382–398. Springer.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Shizhen Chang and Pedram Ghamisi. 2023. Changes to captions: An attentive network for remote sensing change captioning. *IEEE Transactions on Image Processing*, 32:6047–6060.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Gaoxiang Cong, Liang Li, Zhenhuan Liu, Yunbin Tu, Weijun Qin, Shenyuan Zhang, Chengang Yan, Wenyu Wang, and Bin Jiang. 2022. Ls-gan: iterative language-based image manipulation via long and short term consistency reasoning. In *ACM MM*, pages 4496–4504.

Gaoxiang Cong, Liang Li, Yuankai Qi, Zheng-Jun Zha, Qi Wu, Wenyu Wang, Bin Jiang, Ming-Hsuan Yang, and Qingming Huang. 2023. Learning to dub movies via hierarchical prosody models. In *CVPR*, pages 14687–14697.

Zixin Guo, Tzu-Jui Wang, and Jorma Laaksonen. 2022. Clip4idc: Clip for image difference captioning. In *AACL*, pages 33–42.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*, pages 770–778.

Mehrdad Hosseinzadeh and Yang Wang. 2021. Image change captioning by learning from an auxiliary task. In *CVPR*, pages 2725–2734.

Genc Hoxha, Seloua Chouaf, Farid Melgani, and Youcef Smara. 2022. Change captioning: A new paradigm for multitemporal remote sensing image analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14.

Qingbao Huang, Yu Liang, Jielong Wei, Yi Cai, Hanyu Liang, Ho-fung Leung, and Qing Li. 2022. Image difference captioning with instance-level fine-grained feature representation. *IEEE Transactions on Multimedia*, 24:2004–2017.

Harsh Jhamtani and Taylor Berg-Kirkpatrick. 2018. Learning to describe differences between pairs of similar images. In *EMNLP*, pages 4024–4034.

Hoeseong Kim, Jongseok Kim, Hyungseok Lee, Hyunsung Park, and Gunhee Kim. 2021. Agnostic change captioning with cycle consistency. In *ICCV*, pages 2095–2104.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Chenyang Liu, Jiajun Yang, Zipeng Qi, Zhengxia Zou, and Zhenwei Shi. 2023a. Progressive scale-aware network for remote sensing image change captioning. *IGARSS*.

Chenyang Liu, Rui Zhao, Hao Chen, Zhengxia Zou, and Zhenwei Shi. 2022. Remote sensing image change captioning with dual-branch transformers: A new method and a large scale dataset. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–20.

Chenyang Liu, Rui Zhao, Jianqi Chen, Zipeng Qi, Zhengxia Zou, and Zhenwei Shi. 2023b. A decoupling paradigm with prompt learning for remote sensing image change captioning. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–18.

Fenglin Liu, Changchang Yin, Xian Wu, Shen Ge, Ping Zhang, and Xu Sun. 2021. Contrastive attention for automatic chest x-ray report generation. In *Findings of ACL*, pages 269–280.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318.

Dong Huk Park, Trevor Darrell, and Anna Rohrbach. 2019. Robust change captioning. In *ICCV*, pages 4624–4633.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, pages 8024–8035.

Yue Qiu, Yanjun Sun, Fumiya Matsuzawa, Kenji Iwata, and Hirokatsu Kataoka. 2023. Graph representation for order-aware visual transformation. In *CVPR*, pages 22793–22802.

Yue Qiu, Shintaro Yamamoto, Kodai Nakashima, Ryota Suzuki, Kenji Iwata, Hirokatsu Kataoka, and Yutaka Satoh. 2021. Describing and localizing multiple changes with transformers. In *ICCV*, pages 1971–1980.

Noam Rotstein, David Bensaïd, Shaked Brody, Roy Ganz, and Ron Kimmel. 2024. Fusecap: Leveraging large language models for enriched fused image captions. In *WACV*, pages 5689–5700.

Xiangxi Shi, Xu Yang, Jiuxiang Gu, Shafiq Joty, and Jianfei Cai. 2020. Finding it at another side: A viewpoint-adapted matching encoder for change captioning. In *ECCV*, pages 574–590.

Le Song, Alex Smola, Arthur Gretton, Karsten M Borgwardt, and Justin Bedo. 2007. Supervised feature selection via dependence estimation. In *ICML*, pages 823–830.

Yunbin Tu, Liang Li, Li Su, Junping Du, Ke Lu, and Qingming Huang. 2023a. Viewpoint-adaptive representation disentanglement network for change captioning. *IEEE Transactions on Image Processing*, 32:2620–2635.

Yunbin Tu, Liang Li, Li Su, Shengxiang Gao, Chenggang Yan, Zheng-Jun Zha, Zhengtao Yu, and Qingming Huang. 2022. I2 transformer: Intra-and inter-relation embedding transformer for tv show captioning. *IEEE Transactions on Image Processing*, 31:3565–3577.

Yunbin Tu, Liang Li, Li Su, Ke Lu, and Qingming Huang. 2023b. Neighborhood contrastive transformer for change captioning. *IEEE Transactions on Multimedia*, pages 1–12.

Yunbin Tu, Liang Li, Li Su, Zheng-Jun Zha, and Qingming Huang. 2024. Smart: Syntax-calibrated multi-aspect relation transformer for change captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Yunbin Tu, Liang Li, Li Su, Zheng-Jun Zha, Chenggang Yan, and Qingming Huang. 2023c. Self-supervised cross-view representation reconstruction for change captioning. In *ICCV*, pages 2805–2815.

Yunbin Tu, Liang Li, Chenggang Yan, Shengxiang Gao, and Zhengtao Yu. 2021a. Rˆ3Net:relation-embedded representation reconstruction network for change captioning. In *EMNLP*, pages 9319–9329.

Yunbin Tu, Tingting Yao, Liang Li, Jiedong Lou, Shengxiang Gao, Zhengtao Yu, and Chenggang Yan. 2021b. Semantic relation-aware difference representation learning for change captioning. In *Findings of ACL*, pages 63–73.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*, pages 5998–6008.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *CVPR*, pages 4566–4575.

Xu Yang, Jiawei Peng, Zihua Wang, Haiyang Xu, Qinghao Ye, Chenliang Li, Songfang Huang, Fei Huang, Zhangzikang Li, and Yu Zhang. 2023. Transforming visual scene graphs to image captions. In *ACL*, pages 12427–12440.

Linli Yao, Weiying Wang, and Qin Jin. 2022. Image difference captioning with pre-training and contrastive learning. In *AAAI*.

Shengbin Yue, Yunbin Tu, Liang Li, Shengxiang Gao, and Zhengtao Yu. 2024. Multi-grained representation aggregating transformer with gating cycle for change captioning. *ACM Transactions on Multimedia Computing, Communications and Applications*.

Shengbin Yue, Yunbin Tu, Liang Li, Ying Yang, Shengxiang Gao, and Zhengtao Yu. 2023. I3n: Intra- and inter-representation interaction network for change captioning. *IEEE Transactions on Multimedia*, pages 1–14.

Kai Zhao and Wei Xiong. 2024. Cooperative connection transformer for remote sensing image captioning. *IEEE Transactions on Geoscience and Remote Sensing*.

A Appendix

In this appendix, we provide more details about the caption generation and more experimental results, as well as the qualitative analyses.

A.1 Caption Generation

After obtaining the representation of all changes $\tilde{X}_d \in \mathbb{R}^{hw\times d}$, we use a transformer decoder to decode it into target sentences. First, we obtain the embedding features of all m words of these sentence $E[S] \in \mathbb{R}^{m\times d}$, where each sentence is separated by a special token [SEP]. Then, we use the masked self-attention (Vaswani et al., 2017) to model relationships among these word features, which is defined as:

$$\hat{E}[S] = \text{LN}\left(E[S] + \text{MHSA}\left(E[S], E[S], E[S]\right)\right), \tag{11}$$

where LN is short for layer normalization (Ba et al., 2016). Next, we model the interaction between these word features and difference representation $\tilde{X}_d$ by multi-head cross-attention (MHCA) (Vaswani et al., 2017), so as to locate the most related difference features $\tilde{H}$:

$$\tilde{H} = \text{LN}\left(E[\hat{S}] + \text{MHCA}\left(E[\hat{S}], \tilde{X}_d, \tilde{X}_d\right)\right). \tag{12}$$

Subsequently, we feed the selected features $\tilde{H}$ into a feed-forward network to obtain the enhanced difference representation, denoted as $\hat{H} \in \mathbb{R}^{m\times d}$.

$$\hat{H} = \text{LN}((\tilde{H} + \text{FFN}(\tilde{H})). \tag{13}$$

Finally, the probability distributions of words in these sentences are calculated via a single hidden layer:

$$S = \mathrm{Softmax}\left(\hat{H}W_s + b_s\right), \tag{14}$$

where $W_s \in \mathbb{R}^{d\times u}$ and $b_s \in \mathbb{R}^u$ are the learnable parameters. u is the dimension of vocabulary size.

A.2 Experiments

A.2.1 Implementation Details

For fair-comparison, we follow the previous multi-change captioning methods (Jhamtani and Berg-Kirkpatrick, 2018; Qiu et al., 2021; Liu et al., 2022) to use a pre-trained ResNet-101 (He et al., 2016) to extract the local features of a pair of images, with the dimension of 1024 × 14 × 14. We project them into a lower dimension of 512, while the dimension of trainable [CLS] features is also set to 512. The hidden size of overall model and word embedding size are set to 512 and 300, respectively. Temperature τ in Eq. (2) of main paper is set to 0.07. The attention layers in CARD is set to 1 on the CLEVR-Multi-Change and Spot-the-Diff datasets; and 3 on the LEVIR-CC dataset. We train the model with PyTorch (Paszke et al., 2019) on a single RTX 3090 GPU, and use Adam optimizer (Kingma and Ba, 2014) to minimize the negative log-likelihood loss of Eq. (10) in the main paper. The training details about batch size and learning rate are shown in Table 7. The used training resources about time and GPU memory are shown in Table 8. We find that training CARD does not require much more time and GPU memory. Hence, it would be easily re-implemented by other researchers and be a strong baseline for the future works.

Table 7: The training details of CARD on the three datasets.

	batch size	learning rate
CLEVR-Multi-Change	128	2×10^{-4}
LEVIR-CC	64	1×10^{-4}
Spot-the-Diff	32	2×10^{-4}

Table 8: The used training resources of CARD on the three datasets.

	Training Time	GPU Memory
CLEVR-Multi-Change	71 minutes	9.2 GB
LEVIR-CC	120 minutes	8 GB
Spot-the-Diff	20 minutes	3.9 GB

Table 9: Effects of λ_c on CLEVR-Multi-Change.

Model	λ_c	B	M	R	S	C
CARD	0	56.2	45.3	57.3	76.9	387.8
CARD	0.1	56.6	45.2	**57.4**	**77.0**	391.2
CARD	0.2	56.6	**45.3**	57.3	**77.0**	390.7
CARD	0.3	**56.7**	45.2	**57.4**	76.9	**391.6**
CARD	0.4	56.5	45.2	57.0	76.8	388.8
CARD	0.5	56.6	**45.3**	**57.4**	76.9	390.7

Table 10: Effects of λ_c on the LEVIR-CC dataset.

Model	λ_c	B	M	R	C
CARD	0	55.9	35.6	72.3	132.2
CARD	0.1	**65.4**	**40.0**	**74.6**	**137.9**
CARD	0.2	63.7	39.3	73.3	132.9
CARD	0.3	54.8	34.4	72.7	137.5
CARD	0.4	63.2	38.3	73.6	137.5
CARD	0.5	59.5	36.9	72.9	135.1

A.3 Study on the Trade-off Parameter λ_c

We discuss the influence of trade-off parameter λ_c in Eq. (10). The results under varied values are shown in Table 9-11. Note that λ_c=0 means there are no constraints upon the decoupled results. We observe that imposing the constraint losses does improve the model's performance. Besides, there is a trade-off between the captioning loss and constraint losses, because too large λ_c may lead to deterioration in caption performance. Based on the results, we choose the value as 0.3 on CLEVR-Multi-Change, 0.1 and 0.001 on the LEVIR-CC and Spot-the-Diff, respectively.

A.4 Qualitative Analysis

In this appendix, we will show more qualitative examples on the CLEVR-Multi-Change, LEVIR-CC, and Spot-the-Diff datasets, which are shown in Figure 6-9. To intuitively understand whether the common context features help mine reliable common properties, we visualize the alignment of common properties between two images on the three datasets, which are shown in Figure 6-7. The compared method is MCCFormers-D (Qiu et al.,

Table 11: Effects of λ_c on the Spot-the-Diff dataset.

Model	λ_c	B	M	R	S	C
CARD	0	4.3	10.6	23.3	13.0	15.8
CARD	0.001	**6.6**	10.8	**26.9**	**17.8**	**32.4**
CARD	0.002	4.3	9.8	23.9	15.1	23.0
CARD	0.003	6.2	9.5	25.7	15.7	28.4
CARD	0.004	5.0	11.0	24.4	15.2	20.5
CARD	0.005	5.1	**11.4**	24.0	15.6	18.3

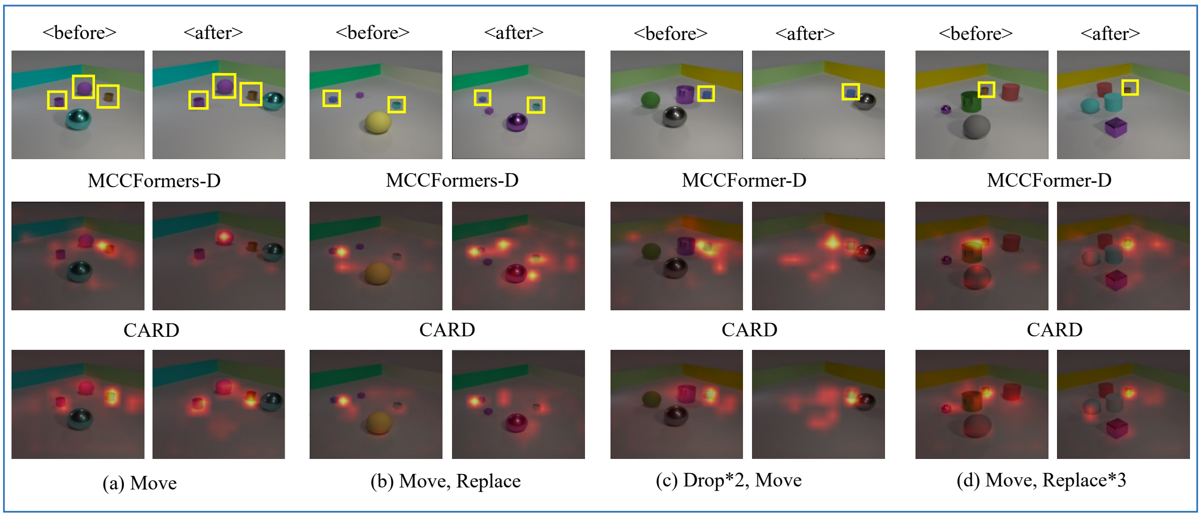


Figure 6: Visualization of common objects matching on the CLEVR-Multi-Change dataset under one-to-four changes. For each example, we visualize the matching results by the state-of-the-art method MCCFormers-D (Qiu et al., 2021) and our CARD. The common objects are shown in the yellow boxes.

2021), which has the stable performance on the three datasets. From these examples, we can observe that MCCFormers-D is unable to align the common properties properly and even misjudges changed objects as unchanged objects. Compared with it, the proposed CARD can better match the common objects, so as to validate the effectiveness of the decoupled common context features.

Further, in Figure 8-9, we visualize the captions yielded by MCCFormers-D and CARD, as well as the change localization results from CARD on the three datasets. In Figure 8, on the CLEVR-Multi-Change dataset under one-to-four changes, we find that MCCFormers-D either only describes partial changes or misidentifies changed objects. Instead, the proposed CARD is able to accurately locate and describe all changed objects. In Figure 9, on the LEVIR-CC and Spot-the-Diff datasets, it is noted that MCCFormers-D fails to describe all the changes within each image pair that is from the real-word environment. Different it, our CARD is capable of locating and describing all changed objects, which show a good generalization and robustness of our method. The superiority benefits from that the proposed CARD can provide the model with a overview of potential changed/unchanged semantics within an image pair, which helps learn genuine changes for caption generation.

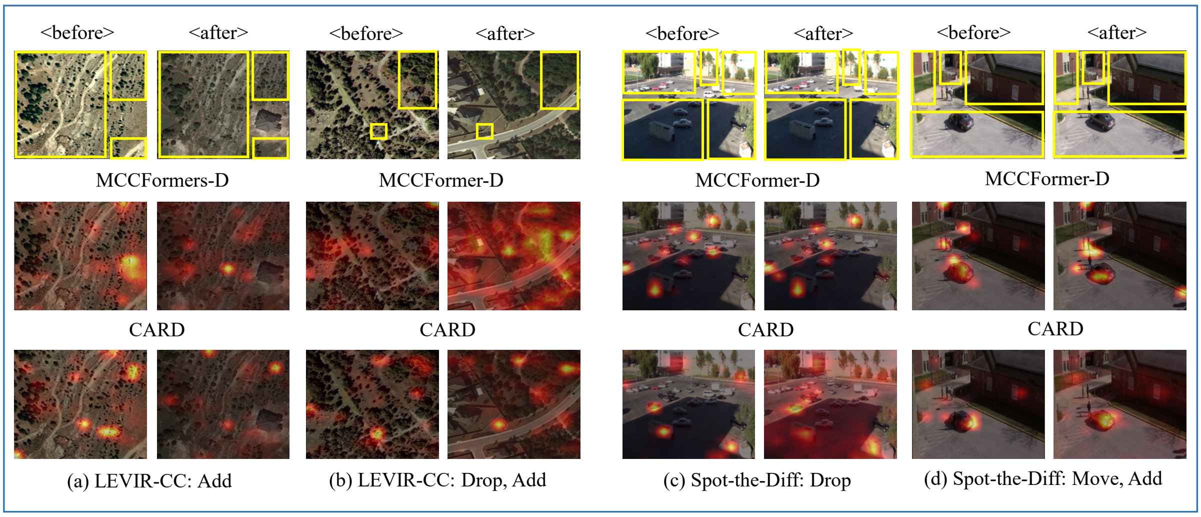
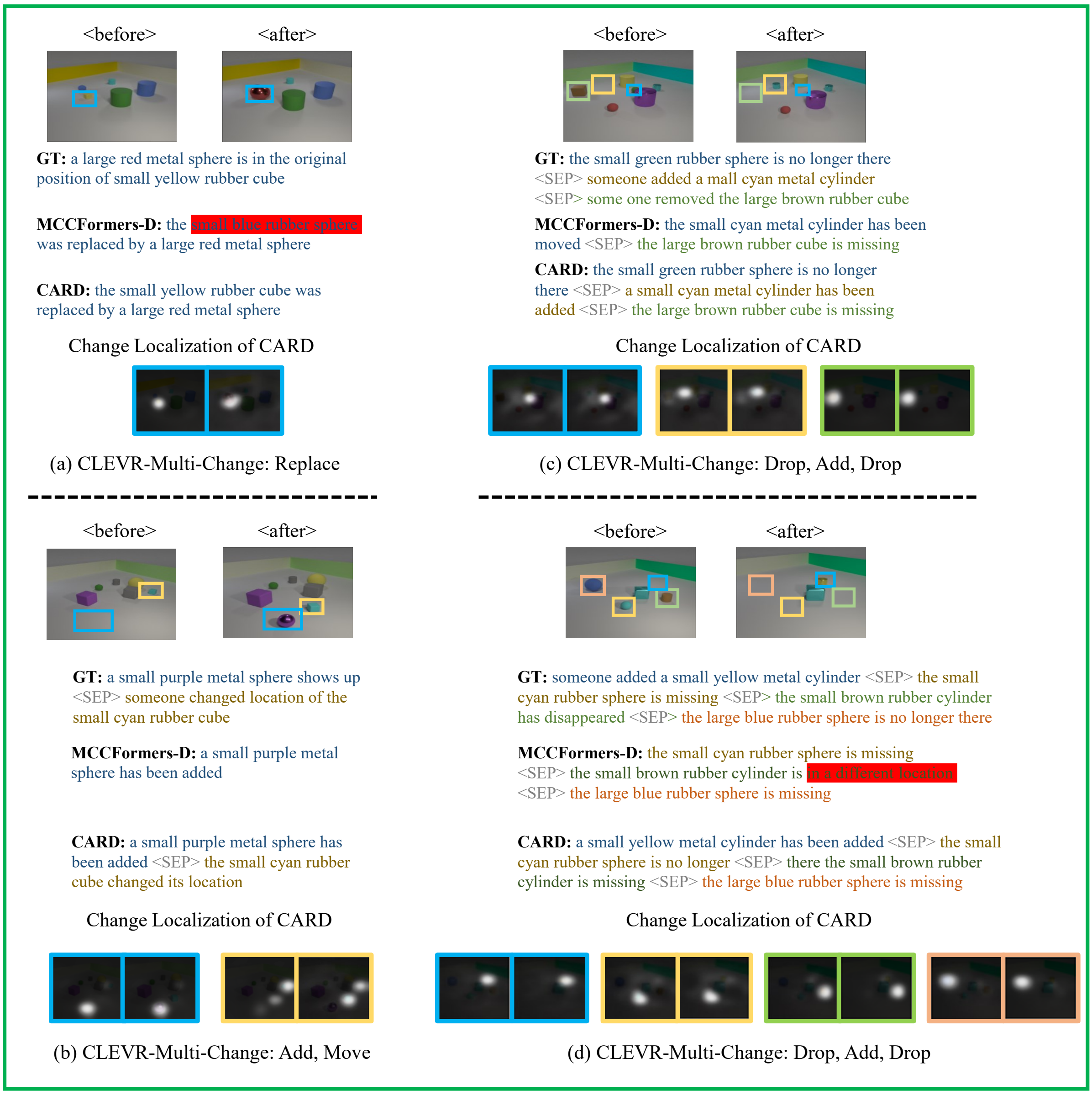


Figure 7: Visualization of common objects matching on the LEVIR-CC and Spot-the-Diff datasets under varied changes. For each example, we visualize the matching results by the state-of-the-art method MCCFormers-D (Qiu et al., 2021) and our CARD. The common objects are shown in the yellow boxes.


<before>
<after>
Change Localization of CARD
<before>
<after>
Change Localization of CARD
<before>
<after>
Change Localization of CARD
<before>
<after>
Change Localization of CARD

GT: a large red metal sphere is in the original position of small yellow rubber cube

MCCFormers-D: the small blue rubber sphere was replaced by a large red metal sphere

CARD: the small yellow rubber cube was replaced by a large red metal sphere

(a) CLEVR-Multi-Change: Replace

GT: the small green rubber sphere is no longer there <SEP> someone added a mall cyan metal cylinder <SEP> some one removed the large brown rubber cube

MCCFormers-D: the small cyan metal cylinder has been moved <SEP> the large brown rubber cube is missing

CARD: the small green rubber sphere is no longer there <SEP> a small cyan metal cylinder has been added <SEP> the large brown rubber cube is missing

(c) CLEVR-Multi-Change: Drop, Add, Drop

GT: a small purple metal sphere shows up <SEP> someone changed location of the small cyan rubber cube

MCCFormers-D: a small purple metal sphere has been added

CARD: a small purple metal sphere has been added <SEP> the small cyan rubber cube changed its location

(b) CLEVR-Multi-Change: Add, Move

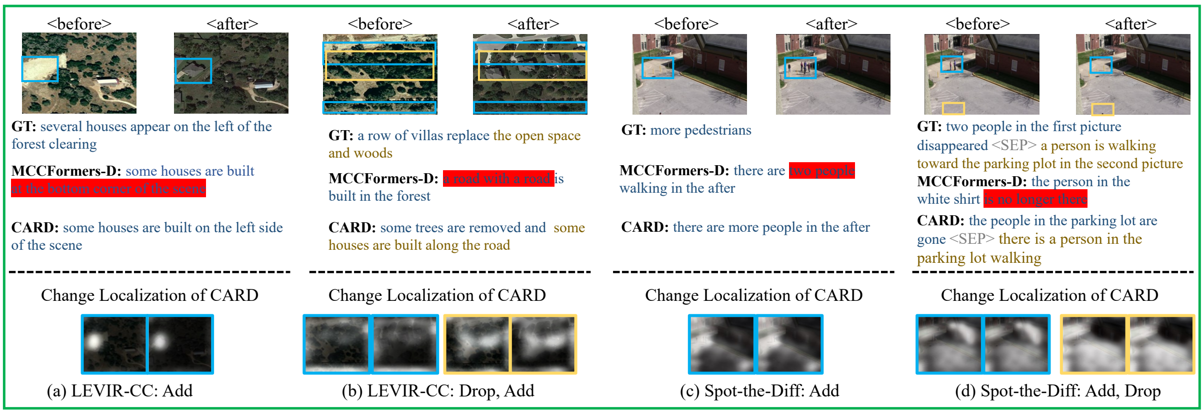
GT: someone added a small yellow metal cylinder <SEP> the small cyan rubber sphere is missing <SEP> the small brown rubber cylinder has disappeared <SEP> the large blue rubber sphere is no longer there

MCCFormers-D: the small cyan rubber sphere is missing <SEP> the small brown rubber cylinder is in a different location <SEP> the large blue rubber sphere is missing

CARD: a small yellow metal cylinder has been added <SEP> the small cyan rubber sphere is no longer <SEP> there the small brown rubber cylinder is missing <SEP> the large blue rubber sphere is missing

(d) CLEVR-Multi-Change: Drop, Add, Drop

Figure 8: Qualitative examples on the CLEVR-Multi-Change dataset under one-to-four changes. For each example, we visualize the captions generated by the state-of-the-art method MCCFormers-D (Qiu et al., 2021) and our CARD, as well as the change localization of CARD. The changed objects are shown in the colored boxes.


<before>
<after>
GT: several houses appear on the left of the forest clearing
MCCFormers-D: some houses are built at the bottom corner of the scene
CARD: some houses are built on the left side of the scene
Change Localization of CARD
<before>
<after>
GT: a row of villas replace the open space and woods
MCCFormers-D: a road with a road is built in the forest
CARD: some trees are removed and some houses are built along the road
Change Localization of CARD
<before>
<after>
GT: more pedestrians
MCCFormers-D: there are two people walking in the after
CARD: there are more people in the after
Change Localization of CARD
<before>
<after>
GT: two people in the first picture disappeared <SEP> a person is walking toward the parking plot in the second picture
MCCFormers-D: the person in the white shirt is no longer there
CARD: the people in the parking lot are gone <SEP> there is a person in the parking lot walking
Change Localization of CARD

(a) LEVIR-CC: Add

(b) LEVIR-CC: Drop, Add

(c) Spot-the-Diff: Add

(d) Spot-the-Diff: Add, Drop

Figure 9: Qualitative examples on the LEVIR-CC and Spot-the-Diff datasets under varied changes. For each example, we visualize the captions generated by the state-of-the-art method MCCFormers-D (Qiu et al., 2021) and our CARD, as well as the change localization of CARD. The changed objects are shown in the colored boxes.