Region-aware Difference Distilling with Attribute-guided Contrastive Regularization for Change Captioning

Rong Li[1], Liang Li[2*], Jiehua Zhang[3], Qiang Zhao[1,4], Hongkui Wang[1,4], Chenggang Yan[1,4]

[1]Hangzhou Dianzi University, Hangzhou, China;
[2]Institute of Computing Technology, Chinese Academy of Sciences;
[3]School of Software Engineering, Xi'an Jiaotong University;
[4]Lishui Institute of Hangzhou Dianzi University
rongli@hdu.edu.cn, liang.li@ict.ac.cn, jiehua.zhang@stu.xjtu.edu.cn, qiangzhao@ieee.org, wanghk@hdu.edu.cn

Abstract

Change captioning aims to describe the differences between two similar images using natural language, significantly aiding in understanding and monitoring changes. This challenging task requires a fine-grained understanding of subtle changes while resisting disturbances like viewpoint shifts and illumination variations. Existing methods often rely solely on global difference features and lack comprehensive alignment of linguistic and visual information, leading to overlooking fine-grained details and generating semantic hallucinated sentences. To address these limitations, we propose the region-aware difference distilling (RDD) network with attribute-guided contrastive regularization (ACR). The RDD uses global difference features to progressively distill regional difference features using learnable vectors, allowing for more precise identification of changed regions. The ACR enhances comprehensive alignment between linguistic and visual information by formulating Nouns-to-Objects (N2O) and Verbs-to-Actions (V2A) alignment losses to regularize the regional difference features. Promising results on three datasets demonstrate that our method outperforms the state-of-the-art change captioning methods.

Introduction

Change captioning aims to describe the difference between two similar images by natural language, making monitoring changes highly understandable(Yan et al. 2022a; Li et al. 2024; Zhang et al. 2024a). This task tries to model the association between dynamic visual content and language(Ye, He, and Peng 2022; Yan et al. 2020a, 2021a; Zhang et al. 2024b; Yan et al. 2020b). It has many practical applications, such as real-time monitoring (Jhamtani and Berg-Kirkpatrick 2018; Sun et al. 2024), geological environment observation (Liu et al. 2022a; Bashmal et al. 2023), medical image analysis(Zhang et al. 2020) and visual reasoning(Liu et al. 2022b; Li et al. 2022; Yan et al. 2021b, 2022b), etc.

Change captioning is a challenging task that involves recognizing fine-grained area changes (Figure 1 (a)) and then building robust connections between changes and language. Additionally, it must handle disturbances such as viewpoint shifts (Figure 1 (b)) and illumination variations (Figure 1 (c)). To facilitate the recognition of fine-grained area changes, current change captioning methods often employ semantic interaction and introduce intricate positional encoding techniques (Yue et al. 2023; Tu et al. 2024b) within two individual images. For resisting the turbulence of pseudo changes, feature matching (Kim et al. 2021), including attention mechanism (Qiu et al. 2021), is often used to match similar patches before difference feature extraction. Contrastive learning is also commonly leveraged to establish cross-view consistency (Tu et al. 2024a, 2023c) between two images, enhancing the recognition of real changes.

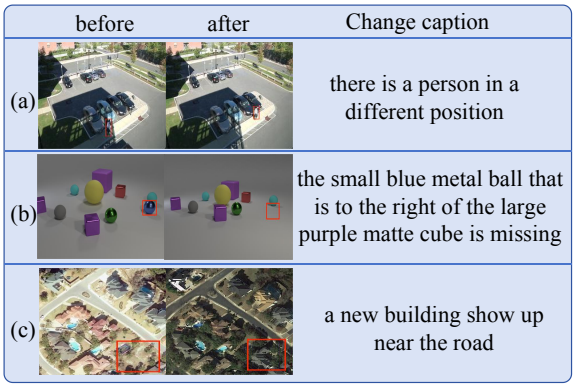

Figure 1: Three examples of change captioning: (a) Slight object movement within a scene. (b) Change with interference from viewpoint shift. (c) Change with interference from illumination variation.

Despite significant progress made by previous methods, two notable limitations remain: (1) Redundancy information unrelated to change is included due to obtaining solely global features. Specifically, previous methods directly subtract representations in the final step, extracting only global difference features. Without complete image alignment, this can include irrelevant elements like background, interfering with fine-grained change description and increasing susceptibility to pseudo-changes. (2) Semantically hallucinated difference representations are constructed by incomprehensive linguistic and visual information alignment. Previous methods rely on ground truth captions for final supervision

*Corresponding author.
Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

or use partial alignment, lacking full interaction between linguistic and visual difference information. As a result, the linguistic and final visual difference features are represented in separate representation spaces. This causes the generated visual difference representation to understand only superficial aspects of the sentence semantics, leading to the model generating semantically hallucinated captions that are structurally similar but differ in meaning. For example: 'The new houses in the forest have been built' versus 'The new houses and the forest have been removed'.

To better overcome the limitations caused by obtaining solely global difference features, the model should explore diverse fine-grained representations(Zha et al. 2019). These representations can accurately capture details of changed regions, even when the alignment between the two images is incomplete. Additionally, for the second limitation, we notice that each component in a sentence carries rich semantics and corresponds to a specific region in images. Projecting sentence components and their corresponding image areas into the same representation space can enhance the comprehensive alignment of linguistic and visual information. Furthermore, inspired by the Concreteness Effect (Paivio, Walsh, and Bons 1994), which suggests that humans remember concrete objects better than abstract actions. We assume concrete attribute information should be used first, followed by abstract ones. This process can help the model gradually deepen its understanding of changes.

Based on the above analysis, we propose a region-aware difference distilling network with attribute-guided contrastive regularization, which learns diverse fine-grained representations to capture subtle changes and generate captions with more precise semantics. Concretely, given a pair of images, the global difference extraction module first extracts global difference features. Then, these features are used by the **R**egion-aware **D**ifference **D**istilling(RDD) to progressively distill regional difference features, which encapsulate fine-grained interference-resistant representations by using learnable vectors. Object regional difference features are distilled first, followed by action regional difference features. To guide vectors in acquiring relevant attributes and ensure comprehensive alignment between linguistic and visual information, we design **A**ttribute-guided **C**ontrastive **R**egularization (ACR). ACR formulates nouns-to-objects (N2O) and verbs-to-actions (V2A) alignment losses. ACR first generates attribute-guided tokens from sentence attributes and then uses two losses to maximize the similarity between these attribute-guided tokens and learnable vectors. Finally, we fuse global and regional difference features to obtain detailed difference features, which are then input to the decoder.

Our main contributions are summarized as follows:

- We propose RDD with a "concrete-to-abstract" two-stage regional feature distilling, which learns two fine-grained regional representations. These representations enhance the model's ability to capture subtle changes within regions.
- We design ACR, which formulates N2O and V2A alignment losses to regularize the distilling process. This ensures comprehensive alignment between linguistic and visual differences.
- Extensive experiments demonstrate the superiority of our method, achieving state-of-the-art results on three public datasets.

Related Work

Change Captioning

Initially, the main challenge of change captioning lies in identifying and describing fine-grained changes (Jhamtani and Berg-Kirkpatrick 2018). Significant progress has been made in tackling this challenge, focusing on global feature representation. These works (Tu et al. 2023a; Yue et al. 2023) introduce adaptive position encoding into the model, while the work (Tu et al. 2021b) explores internal interaction within single image features. Additionally, the other work (Huang et al. 2021) first leverages attribute and position features to augment object features. Later, Park et al.(Park, Darrell, and Rohrbach 2019) introduce a new challenge by incorporating noticeable viewpoint shifts. To address this, attention mechanisms and feature matching methods (Kim et al. 2021; Shi et al. 2020; Qiu et al. 2021; Chang and Ghamisi 2023) enable inter-information interaction, while contrastive learning (Tu et al. 2023c, 2024a) ensures cross-view consistency. Additionally, (Liu et al. 2022a) leverages cross-semantic relationships to enhance change detection. What's more, pretraining-finetuning paradigms are explored by (Guo, Wang, and Laaksonen 2022; Yao, Wang, and Jin 2022) in change captioning, and the work (Tu et al. 2024c) disentangles context features to ensure the detection of actual changes. However, these approaches focus on global differences and neglect fine-grained representations. Our model introduces difference distilling to capture detailed and fine-grained regional features.

Cross-modal Alignment in Change Captioning

To align linguistic information with visual difference features, the method (Tu et al. 2021a) uses syntactic skeleton predictors to enhance semantic interactions between change localization and caption generation. Textual auxiliary information such as (Part of Speech) POS by (Tu et al. 2024b) and word dependency by (Tu et al. 2023b) are leveraged to help better understand complex syntax structure during training. Auxiliary tasks — the composed query image retrieval from (Hosseinzadeh and Wang 2021), are introduced to constrain the precision of the captions produced by the main network. Cross-modal contrastive regularization (Tu et al. 2024a) is introduced to maximize the alignment between globe difference features and all generated words, which is similar to our method. However, ours primarily focuses on regularizing regional difference features with the corresponding attributes of the sentence, enhancing comprehensive alignment between linguistic and visual information.

Methodology

Overview

As shown in Figure 2, our method consists of four parts: (1) The Global Difference Extraction module extracts global

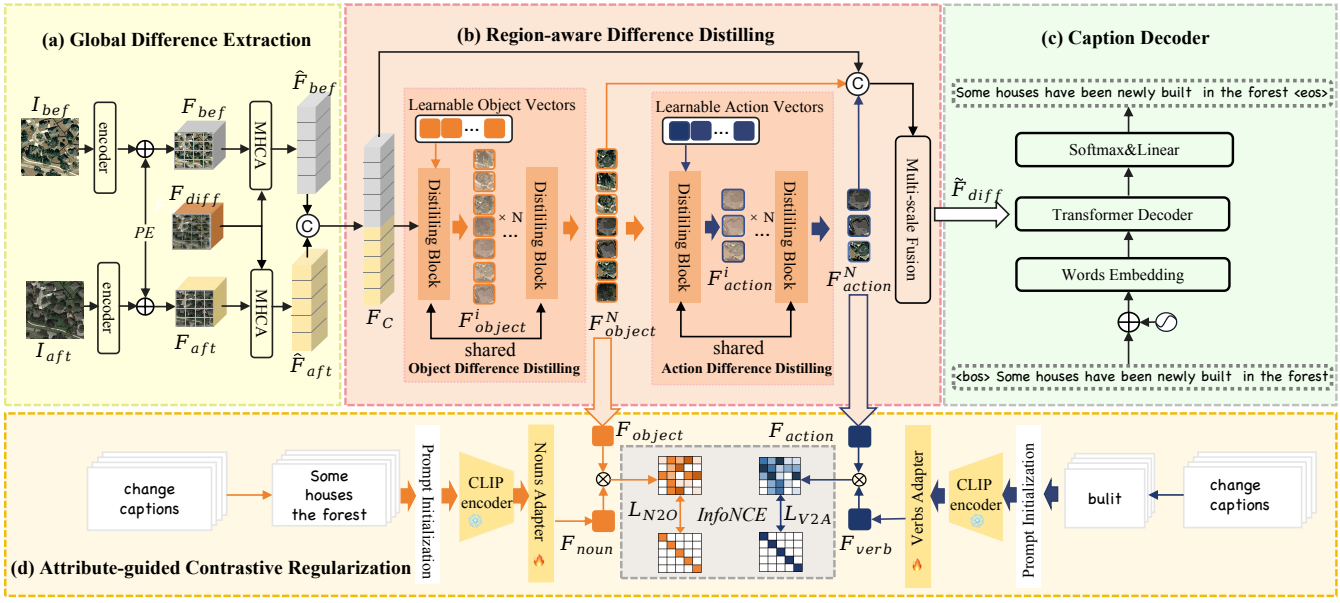

Figure 2: The overall architecture of our method, which consists of (a) Global Difference Extraction, (b) Region-aware Difference Distilling module to distill two types of regional difference features progressively, (c) Caption Decoder and (d) Attribute-guided Contrastive Regularization to generate attribute-guided tokens then use these tokens to regularize the distilling process.

difference features from before and after images. (2) The Region-aware Difference Distilling module first utilizes the global difference features to distill regional object difference features and then uses these features to distill the regional action difference features. Finally, it integrates all the different features from the global, object, and action stages. (3) The Caption Decoder translates the integrated difference features into the change captions. (4) The Attribute-guided Contrastive Regularization is used during training, generating two types of attribute-guided tokens based on the corresponding word segments. Then, we utilize Nouns-to-Objects and Verbs-to-Actions alignment losses to regularize the distilling process.

Global Difference Extraction

Given a pair of images, I_{bef} and I_{aft}, we first utilize a backbone to extract features from both images. Additionally, trainable positional encodings are incorporated into the features to enhance the model's acquaintance with the spatial information:

$$F_x = \text{encoder}(I_x) + PE(x) \tag{1}$$

where $x \in (\text{bef}, \text{aft})$ here. After extracting the image features, we compute the coarse global difference features by directly subtracting the before image features F_{bef} from the after image features F_{aft}:

$$F_{diff} = F_{bef} - F_{aft} \tag{2}$$

To enrich these difference features with more contextual information from the original images, we employ a multi-head cross-attention (Vaswani et al. 2017). It facilitates interaction between difference features and original features, enhancing the model's ability to capture changes. The computation process for single-head attention within the multi-head cross-attention is as follows:

$$SHCA(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}}) \tag{3}$$

$$\hat{F}_x = MHCA(F_x W^Q, F_{diff} W^K, F_{diff} W^V) \tag{4}$$

where $SHCA$ represents single-head cross-attention, and $MHCA$ represents multi-head cross-attention and $x \in (\text{bef}, \text{aft})$ here. Finally, we concatenate the enriched difference features to get the final global difference features:

$$F_c = [\hat{F}_{bef}; \hat{F}_{aft}] \tag{5}$$

where $[\cdot;\cdot]$ represents the concatenation.

Region-aware Difference Distilling

In this module, we process global difference features from "concrete-to-abstract". In this progressive way, our model can learn more fine-grained regional difference representations containing specific and interference-resistant change details. First, we use the object difference distilling module to distill the regional object difference features from the global difference features. Then, the regional action difference features are distilled by the action difference distilling module from the regional object difference feature. Finally, we employ a multi-scale fusion to integrate both global and two kinds of regional features.

Object Difference Distilling To acquire object difference features from the global features F_c, we design the object difference distilling module. First, this module initializes the

learnable object vectors F^0_{object} through Gaussian distribution. Then, to enable the different parts of global difference features to interact with each other, capturing internal correlations and dependencies. And allow the learnable vectors to interact with global information during distillation. We utilize multiple distilling blocks with the same parameters to progressively distill regional object difference features. The distilling block consists of a multi-head self-attention, a multi-head cross-attention and a fully connected layer with residual connection:

$$F^i_c = MHSA(F^{i-1}_c W^Q, F^{i-1}_c W^K, F^{i-1}_c W^V) \quad (6)$$

$$\hat{F}^i_{object} = MHCA(F^{i-1}_{object} W^Q, F^i_c W^K, F^i_c W^V) \quad (7)$$

$$F^i_{object} = FFN(\hat{F}^i_{object}) + \hat{F}^i_{object} \quad (8)$$

where $MHSA$ represents the multi-head self-attention and $i \in [1, N]$ denotes the $i-th$ distilling. F^i_c represents global difference features after passing through the multi-head self-attention of $i-th$ distilling block, and F^0_c equals F_c. F^i_{object} represents object difference features after ith distilling. The computation process of single-head attention in multi-head self-attention can be referenced from Equation (3). And FFN represents the fully connected layer.

Action Difference Distilling To extract action difference features from the object difference features F^N_{object}, we design the action difference distilling module. Its input is object difference features from the last distilling block, and its goal is to generate features representing action changes F^N_{action}. First, learnable action vectors are initialized using a Gaussian distribution. Once initialized, these vectors are fed into multiple distilling blocks to extract regional action difference features, with a structure identical to that of object difference distilling.

Multi-scale Fusion We design a multi-scale fusion strategy using a multi-head self-attention layer and a fully connected layer with residual connections to fully leverage three types of difference features. This approach enables the model to capture the interrelationships between different scales of features, enhancing their complementary and consistent expression. The input to this module is the concatenation of these three types of difference features:

$$\tilde{F}_{diff} = MF(F_c; F^N_{object}; F^N_{action}) \quad (9)$$

where MF represents the multi-scale fusion and $\tilde{F}_{diff}$ represent the final integrated difference features.

Attribute-guided Contrastive Regularization

This module is proposed to guide the distilling process and to enhance the comprehensive alignment between linguistic and visual information. Concretely, we extract verb and noun segments from the ground truth sentences. Leveraging the powerful capabilities of the CLIP model (Radford et al. 2021) in understanding and associating text with images, we extract linguistic attribute features by utilizing the CLIP model. After obtaining the final linguistic attribute-guided tokens, we apply N2O (Nouns-to-Objects) and V2A (Verbs-to-Actions) alignment losses to regularize the two kinds of regional difference features.

Attribute-guided Tokens Generating First, we use spaCy (Honnibal and Montani 2023) to extract all noun and verb segments from the sentence. The extracted noun segments are concatenated using *and*, and the same is done for the verb segments. Instead of using single nouns or verbs, we prioritize verb or noun phrases with adverbs or adjectives to reduce false negatives in contrastive learning. These concatenated segments are then fed into a prompt initialization module, where handcrafted templates or learnable tokens are prepended as prompts. Learnable prompts (Zhou et al. 2022a,b) adaptively generate different prompts for different segments. The prompted nouns and verbs are processed through the CLIP text encoder and an adapter with three fully connected layers, producing attribute-guided tokens F_{noun} and F_{verb}.

Constrastive Regularization To ensure that noun and verb feature tokens effectively supervise the regional object and action difference features, we employ contrastive learning to regularize the two kinds of regional visual difference features. Specifically, we first generate regional visual difference tokens suitable for contrastive learning by using average pooling:

$$F_o = \frac{1}{k} \sum F^N_o \quad (10)$$

where $o \in (\text{object}, \text{action})$ here and k represents the number of object or action difference vectors. After obtaining the attribute-guided tokens F_{noun},F_{verb}, the two regional visual difference tokens F_{object},F_{action}, we align these features. The alignment process is shown as follows.

In each training batch, we sample B object or action tokens and their corresponding noun or verb tokens. For the $k-th$ object or action token, the matching $k-th$ noun or verb token is treated as the positive sample, while all other tokens in the batch serve as negative samples. We use the cosine similarity for the similarity computation, and we maximize contrastive alignment of matched object/action tokens to noun/verb tokens by employing the InfoNCE (Oord, Li, and Vinyals 2018):

$$L_{lc} = -\frac{1}{B} \sum_k^B log \frac{exp(cos(F^k_l, F^k_c)/\tau)}{\sum_r^B exp(cos(F^k_l, F^r_c)/\tau)} \quad (11)$$

$$L_{cl} = -\frac{1}{B} \sum_k^B log \frac{exp(cos(F^k_c, F^k_l)/\tau)}{\sum_r^B exp(cos(F^k_c, F^r_l)/\tau)} \quad (12)$$

$$L_{L2C} = \frac{1}{2}(L_{lc} + L_{cl}) \quad (13)$$

where τ is the temperature hyper-parameter, $c/l \in (\text{object/noun}, \text{action/verb})$, $L/C \in (\text{N/O}, \text{V/A})$.

Caption Decoder

After obtaining the final multi-scale difference features $\tilde{F}_{diff} \in \mathbb{R}^{HW \times C}$, we use a transformer decoder (Vaswani et al. 2017) to generate sentences. A word embedding layer transfers input sentences and adds a learnable position embedding. These sentence features are processed through a masked self-attention layer. The outputs are then fed into

cross-attention layers to interact with the multi-scale difference features. Then, a feed-forward network produces the final representations, denoted as $\hat{H} \in \mathbb{R}^{n \times d}$. Finally, the probability distributions of words in these sentences are computed using a single hidden layer:

$$P = Softmax(\hat{H}W_p + b_p) \tag{14}$$

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

Joint Training

The overall model is trained by maximizing the likelihood of the observed word sequence and target ground-truth caption $S = [s_1, s_2...s_L]$, via cross-entropy loss:

$$L_{XE} = -\frac{1}{B}\sum_{i=1}^{B}\frac{1}{L}\sum_{j=1}^{L} logp_\theta(s_j^i | s_{1:j-1}^i, I_{bef}^i, I_{aft}^i) \tag{15}$$

where $logp_\theta(s_j^i | s_{1:j-1}^i, I_{bef}^i, I_{aft}^i)$ is computed by Equation (14), and θ are all the learnable parameters, and L represents the caption's length. At the training stage, the overall objective of our model is measured as the integration of the two contrastive regularization objectives of L_{N2O}, L_{V2A}, along with the cross-entropyL_{XE}.

$$L = L_{XE} + \lambda_1 L_{N2O} + \lambda_2 L_{V2A} \tag{16}$$

whereλ_1 and λ_2 are hyper-aprameters.

Experiments

Datasets

CLEVR-Change (Park, Darrell, and Rohrbach 2019) is a synthetic dataset with 79,606 image pairs and 493,735 captions featuring perspective distortions and change types like "Color", "Texture", "Add", "Drop" and "Move." It is split into 67,660 training, 3,976 validation, and 7,970 testing images. **LEVIR-CC** (Liu et al. 2022a) is a remote sensing dataset containing 10,077 image pairs and 50,385 captions, primarily focusing on changes under variations in lighting and seasons, with 6,815 images for training, 1,333 for validation, and 1,929 for testing. **Spot-the-Diff** (Jhamtani and Berg-Kirkpatrick 2018) is a surveillance dataset with 13,192 image pairs and an average of 1.86 captions per pair, divided into training, validation, and testing sets in an 8:1:1 ratio.

Implementation Details

We use pre-trained ResNet-101 (He et al. 2016) as the visual backbone to extract image features into a dimension of $14 \times 14 \times 1024$, and then these features are projected into a lower dimension of 512. The hidden size of the model and word embedding size are set to 512 and 300. We use Adam (Kingma and Ba 2014) optimizer to minimize the loss of Equation (16). For more details on the implementation, please see the supplementary material.

Method	B	M	R	C	S
DUDA (ICCV'19)	47.3	33.9	-	112.3	24.5
M-VAM (ECCV'20)	50.3	37.0	69.7	114.9	30.5
R3Net+SSP (EMNLP'21)	54.7	39.8	73.1	123.0	32.6
MCCFormers-D (ICCV'21)	52.4	38.3	-	121.6	26.8
MCCFormers-S (ICCV'21)	**57.4**	41.2	-	125.5	32.4
IFDC (TMM'22)	49.2	32.5	69.1	118.7	-
NCT (TMM'23)	55.1	40.2	73.8	124.1	32.9
I3N-TD (TMM'23)	55.8	40.6	73.9	125.6	32.8
VARD-Trans (TIP'23)	55.4	40.1	73.8	126.4	32.6
SCORER+CBR (ICCV'23)	56.3	41.2	74.5	126.8	33.3
MURAT+GCM (TOMM'24)	55.4	40.4	73.9	127.0	32.4
SMART (TPAMI'24)	56.1	40.8	74.2	127.0	33.4
RDD+ACR	56.1	**41.3**	**75.0**	**128.1**	**33.5**

Table 1: Comparison with the SOTA methods on CLEVR-Change Dataset.

Method	B	M	R	C
DUDA (ICCV'19)	57.8	37.2	71.0	124.3
MCCFormers-D (ICCV'21)	56.4	37.3	70.3	124.4
MCCFormers-S (ICCV'21)	56.7	36.2	69.5	120.4
RSICCformer (TGARS'22)	62.8	39.6	74.1	134.1
Chg2cap (TIP'23)	64.5	40.0	75.1	136.6
PSNet (IGARSS'23)	62.1	38.8	73.6	132.6
Prompt-CC (TGARS'23)	63.5	38.8	73.7	136.4
CARD (ACL'24)	65.4	40.0	74.6	137.9
RDD+ACR	**65.6**	**40.3**	**75.5**	**138.3**

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

Performance Comparison

We compare our model with the following SOTA methods: DUDA (Park, Darrell, and Rohrbach 2019), M-VAM (Shi et al. 2020), R3Net-SSP (Tu et al. 2021a), MCCFormers-D and MCCFormers-S (Qiu et al. 2021), IFDC (Huang et al. 2021), NCT (Tu et al. 2023b), I3N-TD (Yue et al. 2023), VARD-Trans (Tu et al. 2023a), SCORER+CBR (Tu et al. 2023c), SMART (Tu et al. 2024b), MURAT+GCM (Yue et al. 2024), RSICCformer (Liu et al. 2022a), Chg2cap (Chang and Ghamisi 2023), PSNet (Liu et al. 2023a), Prompt-CC (Liu et al. 2023b), CARD (Tu et al. 2024c), DIRL+CCR (Tu et al. 2024a).

Following the existing methods of change captioning, we use five metrics for evaluating the generated captions of our model: BLEU-4(B) (Papineni et al. 2002), METEOR(M) (Banerjee and Lavie 2005), ROUGE-L(R) (Lin 2004), CIDEr(C) (Vedantam, Lawrence Zitnick, and Parikh 2015), and SPICE(S) (Anderson et al. 2016). We use the learnable prompt to generate attribute-guided tokens for the following comparison.

Result on CLEVR-Change Dataset The experimental results of the CLEVR-Change Dataset are shown in Table 1. By comparing our method with other SOTA approaches,

Method	B	M	R	C
DDLA (EMNLP'18)	8.5	12	28.6	32.8
DUDA (ICCV'19)	9.1	11.8	29.1	32.5
R3Net-SSP (EMNLP'21)	-	13.1	32.6	36.6
MCCFormers-D (ICCV'21)	10	12.4	-	43.1
MCCFormers-S (ICCV'21)	9.8	12.3	-	41.6
SCORER+CBR (ICCV'23)	10.2	12.2	-	38.9
MURAT+GCM (TOMM'24)	10.2	13.1	**33.1**	39.4
SMART (TPAMI'24)	-	13.5	31.6	39.4
DIRL+CCR (ECCV'24)	**10.3**	13.8	32.8	40.9
RDD+ACR	9.2	**13.9**	31.0	**43.6**

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

	B	M	R	C	S
direct-sub	46.6	38.1	71.9	120.4	30.7
GDE	54.5	40.5	74.0	125.2	33.4
GDE+N	54.0	38.5	72.6	124.8	32.0
GDE+V	55.4	41.1	74.7	127.5	33.4
GDE+V&N	55.1	40.6	73.9	126.4	33.2
GDE+V+N	56.0	41.0	74.7	127.6	33.2
GDE+N+V	**56.1**	**41.3**	**75.0**	**128.1**	**33.5**

Table 4: Ablation study for each encoder modules on CLEVR-Change Dateset

it is evident that our model outperforms previous SOTA methods in most metrics. This indicates that our model performs more effectively while describing elusive changes posed by viewpoint shifts. Unlike prior methods such as SCORER, which establishes cross-view consistency through contrastive learning, or MCCFormers-D and MCCFormers-S, which focus on capturing similarities between images, our method excels by emphasizing regional fine-grained representations. Additional results related to this dataset can be found in the supplementary material.

Result on LEVIR-CC Dataset The experimental results of the LEVIR-CC Dataset are shown in Table 2. Compared to the previous methods, our model has significant advantages in capturing subtle changes and resisting disturbances caused by long sampling intervals, such as variations in lighting and seasons. Due to the extreme overhead perspective of remote sensing images, object details are often difficult to recognize. However, our method effectively leverages regional and global difference features, enabling more accurate capture of subtle changes and distinguishing between pseudo-changes and real changes.

Result on Spot-the-Diff Dataset The experimental results of the Spot-the-Diff Dataset are shown in Table 3. Ours outperforms previous methods on the METEOR(M) and CIDEr(C) metrics but performs slightly worse on the BLEU(B) and ROUGE(R) metrics. Due to the small size and the variable length of captions in this dataset, we speculate that results from comprehensive alignment in the intermediate stages. It enhances the model's ability to express semantics but leads to a decrease in the degree of matching between the generated sentences and the original sentences in this dataset.

N2O	V2A	B	M	R	C	S
✗	✗	55.5	41.2	75.0	127.0	33.1
✓	✗	55.5	40.9	74.7	127.2	33.2
✗	✓	55.2	40.4	74.4	127.5	32.0
✓	✓	**56.1**	**41.3**	**75.0**	**128.1**	**33.5**

Table 5: Ablation study for contrastive regularization losses on CLEVR-Change Dataset

	B	M	R	C	S
No Init	54.7	41.2	74.5	127.7	33.4
Fixed Prompt	54.9	41.1	74.7	**128.4**	**33.5**
Learnable Prompts	**56.1**	**41.3**	**75.0**	128.1	**33.5**

Table 6: Experiments on different linguistic prompts on CLEVR-Change Dataset

Ablation Study

To determine the contribution of each component, we conduct ablation studies on two large-scale datasets: CLEVR-Change and LEVIR-CC. The results for CLEVR-Change are shown, while the ablation study results for LEVIR-CC can be found in the supplementary materials.

Ablation Study for Each Encoder Modules Table 4 shows the ablation studies of GDE and RDD in terms of total performance, using a transformer decoder for all studies. Direct-sub refers to subtracting features of two images to obtain the final difference features. GDE uses a global difference extraction module. Results indicate that interacting coarse difference features with original features improves the change representation capability of the global difference feature. Building on GDE, we add object and action regional difference features. GDE+N adds only regional object difference features, while GDE+V adds only regional action difference features. Adding regional object difference features alone leads to suboptimal performance, likely due to interference from all candidate object information during multi-scale fusion. In contrast, incorporating action difference features improves the model's understanding of abstract changes.

Ablation of Distilling Order Moreover, the last three rows of Table 4 highlight the importance of the distilling order. GDE+V&N refers to simultaneous distillation, while GDE+N+V and GDE+V+N represent two opposite orders for distilling objects and actions. The regional action difference feature should be refined from object information rather than directly from global features. Learning actions after objects helps center action representations on objects, accurately capturing action-related change regions. In contrast, simultaneous learning from global features risks focusing on pseudo changes, such as viewpoint shifts.

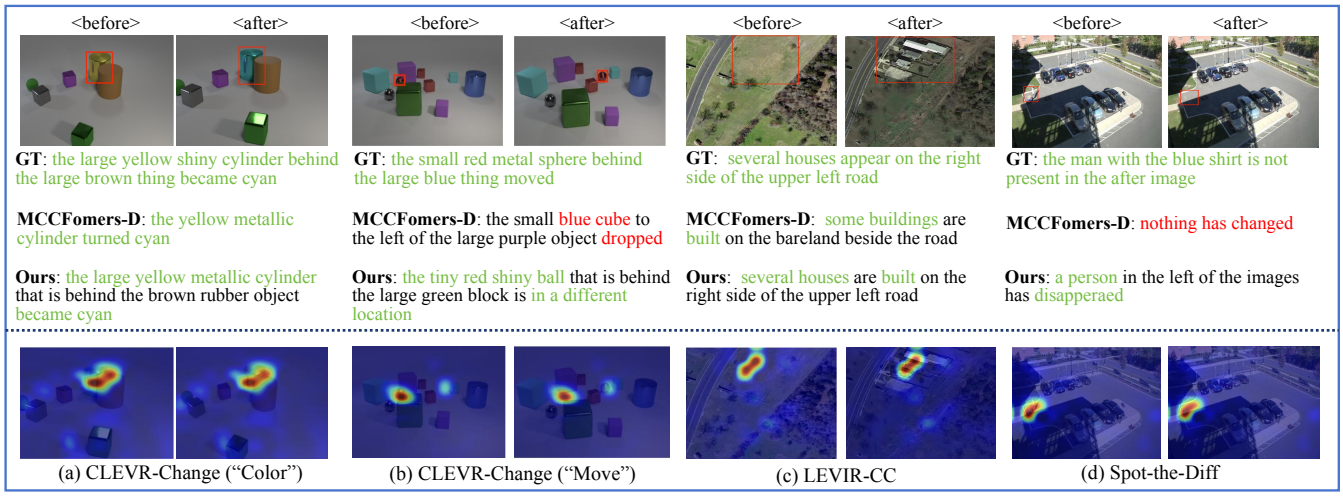

Figure 3: Our generated captions compared with MCCFormer-D's captions and the difference attention maps on three datasets.

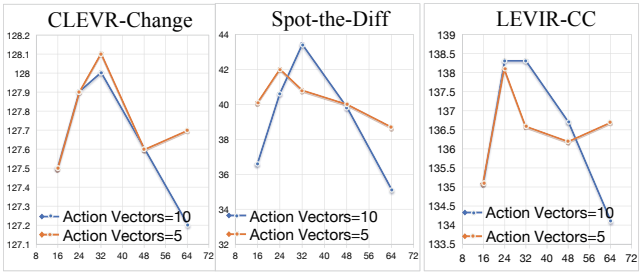

Figure 4: Effect of learnable vectors number of RDD

Ablation Study for Contrastive Regularization Table 5 shows the ablation study of contrastive regularization losses. When both contrastive regularization losses are used, the model achieves higher alignment between attribute-guided tokens and regional difference features. This improvement is better than in scenarios where only one type of contrastive regularization is applied or none.

Experiments on Different Linguistic Prompts We use three different prompt initialization methods: No Init, which extracts verb or noun segments from captions and connects them with "and"; Fixed Prompt, which uses a manually designed prompt; and Learnable Prompt, which adds 4 learnable prompt tokens before the verbs or nouns. The results shown in Table 6 indicate that adding a prompt for verb or noun segments is more effective than not adding one. Additionally, the learnable prompts are better at adapting to different words, leading to overall better performance than the manually designed fixed prompt.

Effect of Learnable Vectors Number of RDD We investigate the effect of the number of vectors in Figure 4. Our findings show that increasing the number of vectors on three datasets does not improve performance due to more non-changing interferences being distilled. Conversely, too few vectors result in insufficient change information. Experiments reveal that 32 learnable object vectors yield the best results. For CLEVR-Change, LEVIR-CC, and Spot-the-Diff datasets, setting learnable action vectors to 5, 10, and 10, respectively, achieves optimal performance.

Qualitative Analysis

To better evaluate our model, we conduct a qualitative analysis. Figure 3 compares our method with MCCFormers-D, a transformer-based feature matching approach, and also shows the visualization for change localization. In these attention maps, areas with more intense red coloring indicate greater attention from the model. The results demonstrate that our model can effectively locate changes even under illumination variations and viewpoint shifts. For the generated captions, we observe that on the CLEVR-Change dataset, methods based on feature matching often rely solely on global information, making them prone to missing subtle changes caused by viewpoint alterations. For example, as shown in Figure 3(b), MCCFormers-D incorrectly identifies changes to the small red metal sphere obscured by a large green block. Similarly, in Figure 3(d) for the Spot-the-Diff dataset, MCCFormers-D overlooks small character changes. In contrast, our method effectively utilizes attribute-guided regional information to detect these subtle changes.

Conclusion

This paper introduces a novel region-aware difference distilling (RDD) network with attribute-guided contrastive regularization (ACR) to address the challenges and limitations in change captioning tasks. The RDD leverages global difference features to progressively distill regional difference features, thereby enhancing the accuracy in identifying change areas while resisting pseudo changes. ACR regularizes the regional visual difference features for comprehensive cross-modal alignment through Noun-to-Object (N2O) and Verb-to-Action (V2A) losses. Extensive experiments demonstrate that our method achieves SOTA performance in the change captioning task.

Acknowledgements

This work was supported by National Natural Science Foundation of China (U21B2024, 62322211, 62336008), "Pioneer" and "Leading Goose" R&D Program of Zhejiang Province (2024C01023, 2023C01046).

References

Anderson, P.; Fernando, B.; Johnson, M.; and Gould, S. 2016. Spice: Semantic propositional image caption evaluation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, 382–398. Springer.

Banerjee, S.; and Lavie, A. 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*, 65–72.

Bashmal, L.; Bazi, Y.; Melgani, F.; Al Rahhal, M. M.; and Al Zuair, M. A. 2023. Language Integration in Remote Sensing: Tasks, datasets, and future directions. *IEEE Geoscience and Remote Sensing Magazine*, 11(4): 63–93.

Chang, S.; and Ghamisi, P. 2023. Changes to captions: An attentive network for remote sensing change captioning. *IEEE Transactions on Image Processing*.

Guo, Z.; Wang, T.-J. J.; and Laaksonen, J. 2022. CLIP4IDC: CLIP for image difference captioning. *arXiv preprint arXiv:2206.00629*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Honnibal, M.; and Montani, I. S. 2023. 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. 2017. *Unpublished software application. https://spacy. io*.

Hosseinzadeh, M.; and Wang, Y. 2021. Image change captioning by learning from an auxiliary task. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2725–2734.

Huang, Q.; Liang, Y.; Wei, J.; Cai, Y.; Liang, H.; Leung, H.-f.; and Li, Q. 2021. Image difference captioning with instance-level fine-grained feature representation. *IEEE transactions on multimedia*, 24: 2004–2017.

Jhamtani, H.; and Berg-Kirkpatrick, T. 2018. Learning to describe differences between pairs of similar images. *arXiv preprint arXiv:1808.10584*.

Kim, H.; Kim, J.; Lee, H.; Park, H.; and Kim, G. 2021. Agnostic change captioning with cycle consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2095–2104.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Li, L.; Gao, X.; Deng, J.; Tu, Y.; Zha, Z.-J.; and Huang, Q. 2022. Long short-term relation transformer with global gating for video captioning. *IEEE Transactions on Image Processing*, 31: 2726–2738.

Li, Z.; Wang, P.; Wang, Z.; and Zhan, D.-c. 2024. Flowgananomaly: Flow-based anomaly network intrusion detection with adversarial learning. *Chinese Journal of Electronics*, 33(1): 58–71.

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

Liu, C.; Chen, K.; Qi, Z.; Zhang, H.; Zou, Z.; and Shi, Z. 2023a. Pixel-level change detection pseudo-label learning for remote sensing change captioning. *arXiv preprint arXiv:2312.15311*.

Liu, C.; Zhao, R.; Chen, H.; Zou, Z.; and Shi, Z. 2022a. 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.

Liu, C.; Zhao, R.; Chen, J.; Qi, Z.; Zou, Z.; and Shi, Z. 2023b. A decoupling paradigm with prompt learning for remote sensing image change captioning. *IEEE Transactions on Geoscience and Remote Sensing*.

Liu, X.; Li, L.; Wang, S.; Zha, Z.-J.; Li, Z.; Tian, Q.; and Huang, Q. 2022b. Entity-enhanced adaptive reconstruction network for weakly supervised referring expression grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3): 3003–3018.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Paivio, A.; Walsh, M.; and Bons, T. 1994. Concreteness effects on memory: When and why? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20(5): 1196.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Park, D. H.; Darrell, T.; and Rohrbach, A. 2019. Robust change captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4624–4633.

Qiu, Y.; Yamamoto, S.; Nakashima, K.; Suzuki, R.; Iwata, K.; Kataoka, H.; and Satoh, Y. 2021. Describing and localizing multiple changes with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1971–1980.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Shi, X.; Yang, X.; Gu, J.; Joty, S.; and Cai, J. 2020. Finding it at another side: A viewpoint-adapted matching encoder for change captioning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, 574–590. Springer.

Sun, Y.; Qiu, Y.; Khan, M.; Matsuzawa, F.; and Iwata, K. 2024. The STVchrono Dataset: Towards Continuous Change Recognition in Time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14111–14120.

Tu, Y.; Li, L.; Su, L.; Du, J.; Lu, K.; and Huang, Q. 2023a. Adaptive representation disentanglement network for change captioning. *IEEE Transactions on Image Processing*, 32: 2620–2635.

Tu, Y.; Li, L.; Su, L.; Lu, K.; and Huang, Q. 2023b. Neighborhood contrastive transformer for change captioning. *IEEE Transactions on Multimedia*, 25: 9518–9529.

Tu, Y.; Li, L.; Su, L.; Yan, C.; and Huang, Q. 2024a. Distractors-Immune Representation Learning with Cross-modal Contrastive Regularization for Change Captioning. *arXiv preprint arXiv:2407.11683*.

Tu, Y.; Li, L.; Su, L.; Zha, Z.-J.; and Huang, Q. 2024b. Smart: Syntax-calibrated multi-aspect relation transformer for change captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Tu, Y.; Li, L.; Su, L.; Zha, Z.-J.; Yan, C.; and Huang, Q. 2023c. Self-supervised cross-view representation reconstruction for change captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2805–2815.

Tu, Y.; Li, L.; Su, L.; Zha, Z.-J.; Yan, C.; and Huang, Q. 2024c. Context-aware Difference Distilling for Multi-change Captioning. *arXiv preprint arXiv:2405.20810*.

Tu, Y.; Li, L.; Yan, C.; Gao, S.; and Yu, Z. 2021a. R ^{3} Net: Relation-embedded Representation Reconstruction Network for Change Captioning. *arXiv preprint arXiv:2110.10328*.

Tu, Y.; Yao, T.; Li, L.; Lou, J.; Gao, S.; Yu, Z.; and Yan, C. 2021b. Semantic relation-aware difference representation learning for change captioning. In *Findings of the association for computational linguistics: ACL-IJCNLP 2021*, 63–73.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4566–4575.

Yan, C.; Gong, B.; Wei, Y.; and Gao, Y. 2020a. Deep multi-view enhancement hashing for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(4): 1445–1451.

Yan, C.; Hao, Y.; Li, L.; Yin, J.; Liu, A.; Mao, Z.; Chen, Z.; and Gao, X. 2021a. Task-adaptive attention for image captioning. *IEEE Transactions on Circuits and Systems for Video technology*, 32(1): 43–51.

Yan, C.; Li, Z.; Zhang, Y.; Liu, Y.; Ji, X.; and Zhang, Y. 2020b. Depth image denoising using nuclear norm and learning graph model. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 16(4): 1–17.

Yan, C.; Meng, L.; Li, L.; Zhang, J.; Wang, Z.; Yin, J.; Zhang, J.; Sun, Y.; and Zheng, B. 2022a. Age-invariant face recognition by multi-feature fusionand decomposition with self-attention. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 18(1s): 1–18.

Yan, C.; Sun, Y.; Zhong, H.; Zhu, C.; Zhu, Z.; Zheng, B.; and Zhou, X. 2022b. Review of omnimedia content quality evaluation. *J. Signal Process.*, 38(6): 1111–1143.

Yan, C.; Teng, T.; Liu, Y.; Zhang, Y.; Wang, H.; and Ji, X. 2021b. Precise no-reference image quality evaluation based on distortion identification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 17(3s): 1–21.

Yao, L.; Wang, W.; and Jin, Q. 2022. Image difference captioning with pre-training and contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 3108–3116.

Ye, Z.; He, X.; and Peng, Y. 2022. Unsupervised Cross-Media Hashing Learning via Knowledge Graph. *Chinese Journal of Electronics*, 31(6): 1081–1091.

Yue, S.; Tu, Y.; Li, L.; Gao, S.; and Yu, Z. 2024. Multi-grained Representation Aggregating Transformer with Gating Cycle for Change Captioning. *ACM Transactions on Multimedia Computing, Communications and Applications*.

Yue, S.; Tu, Y.; Li, L.; Yang, Y.; Gao, S.; and Yu, Z. 2023. I3n: Intra-and inter-representation interaction network for change captioning. *IEEE Transactions on Multimedia*, 25: 8828–8841.

Zha, Z.-J.; Liu, D.; Zhang, H.; Zhang, Y.; and Wu, F. 2019. Context-aware visual policy network for fine-grained image captioning. *IEEE transactions on pattern analysis and machine intelligence*, 44(2): 710–722.

Zhang, B.; Li, L.; Wang, S.; Cai, S.; Zha, Z.-J.; Tian, Q.; and Huang, Q. 2024a. Inductive state-relabeling adversarial active learning with heuristic clique rescaling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhang, D.; Zhang, H.; Tang, J.; Hua, X.-S.; and Sun, Q. 2020. Causal intervention for weakly-supervised semantic segmentation. *Advances in Neural Information Processing Systems*, 33: 655–666.

Zhang, Z.; Li, L.; Cong, G.; Yin, H.; Gao, Y.; Yan, C.; Hengel, A. v. d.; and Qi, Y. 2024b. From speaker to dubber: movie dubbing with prosody and duration consistency learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 7523–7532.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022a. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16816–16825.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022b. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337–2348.