Change Entity-guided Heterogeneous Representation Disentangling for Change Captioning

Anonymous ACL submission

Abstract

Change captioning aims to describe differences between a pair of images using natural language. However, learning effective difference representations is highly challenging due to distractors such as illumination and viewpoint changes. To address this, we propose a change-entity-guided disentanglement network that explicitly learns difference representations while mitigating the impact of distractors. Specifically, we first design a change entity retrieval module to identify key objects involved in the change from a textual perspective. Then, we introduce a difference representation enhancement module that strengthens the learned features, disentangling genuine differences from background variations. To further refine the generation process, we incorporate a gated Transformer decoder, which dynamically integrates both visual difference and textual change-entity information. Extensive experiments on CLEVR-Change, CLEVR-DC and Spot-the-Diff datasets demonstrate that our method outperforms existing approaches, achieving state-of-the-art performance. The code will be released.

1 Introduction

Change captioning aims to describe the differences between two images using natural language. Unlike conventional image captioning that describes main content of a single image, change captioning requires understanding both the semantic correspondence and the differences between a pair of images. This task has garnered significant attention due to its wide-ranging applications in fields such as visual monitoring (Jhamtani and Berg-Kirkpatrick, 2018), remote sensing image analysis (Liu et al., 2024), and medical image comparison (Chen et al., 2024).

Existing methods (Park et al., 2019; Shi et al., 2020; Kim et al., 2021) mainly follow an encoder-decoder framework, which first extracts patch features from a pair of images, then models the difference features in between, and finally decodes these features to generate change captions. To accurately locate the change regions, current works (Qiu et al., 2021; Yao et al., 2022) mostly match similar features between the two images and then disentangle the difference features. Additionally, to generate higher-quality captions, some studies (Hu et al., 2024; Zhang et al., 2024) introduce large language models (LLMs) into this task. They primarily replace the LSTM/Transformer structure with pre-trained LLMs, and further fine-tune the LLMs with different strategies to make them adapt to change captioning.

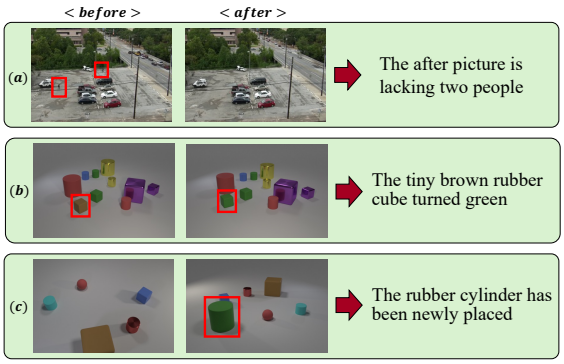

Figure 1: Examples of change captioning. (a) depicts a change occurring in a real-world scenario. (b) shows a change involving a viewpoint shift. (c) illustrates a change occurring under extreme viewpoint variation.

Despite the progress, there are two major limitations in existing approaches. First, viewpoint variation (Figure 1 (c)) between image pairs often leads to deformation of objects in the images (i.e., pseudo changes (Tu et al., 2023c)). Such pseudo changes make the distinguishing of really semantic changes more challenging. Existing works attempt to reduce the influence of irrelevant factors through introducing additional mechanisms in the visual encoder, such as using contrastive learning to align the visual features (Tu et al., 2023c, 2024a).

This approach, however, does not demonstrate significant effectiveness under extreme viewpoint changes (Figure 1 (c)), as pseudo changes in these scenarios become more pronounced. This leads to difficulties in feature matching for unchanged objects, which affects subsequent change localization. We have observed that despite the challenge of distinguishing real changes from distractions based solely on visual features, the similarity between the object representation in the image and its corresponding textual representation remains relatively unaffected. Some works (Kim et al., 2024) attempt to introduce full-sentence descriptions as prior knowledge. However, directly using an entire sentence as prior knowledge introduces a lot of redundancy and even incorrect information. In fact, it is sufficient to focus only on the change entity provided by the text (such as a red cylinder, a green cube, etc.) which indicates what has changed.

Secondly, previous studies (Qiu et al., 2021; Huang et al., 2021) typically rely solely on visual features as input to the decoder. Some existing works incorporate additional information, such as part-of-speech tags (Tu et al., 2021b), to generate higher-quality descriptions. There are also methods based on LLMs (Hu et al., 2024) that yield good results, but they come with considerable computational cost. If we can incorporate some semantic prior information to guide the model, it could improve model performance without introducing significant cost.

In this paper, we propose a novel **CH**ange **E**ntity-guided h**E**terogeneous **R**epresentation di**S**entangling (**CHEERS**) network, which explicitly models and uses textual change entities to guide both feature disentanglement and caption generation. Specifically, we first design a Change Entity Retrieval Module, to locate what has changed based on the similarity between change entities and images. Second, we design a Heterogeneous Representation Disentangling module to decouple the genuine differences between two images and generate the representations that encapsulate the difference information. Here, we devise a Commonality Representation Enhancement module (CRE) that strengthens visual features in similar regions to decouple the difference information from the similar features. Then, we use a Difference Representation Enhancement (DRE) module, to highlight the difference regions. Meanwhile, the change entities are further used to enhance the difference features, while enforcing consistency in the enhanced regions between the image-image and image-entity pairs to constrain the model. Finally, to generate more accurate change captions, we design a gated transformer decoder that dynamically fuses the change entity textual information with the difference visual features. Through the gating mechanism, the model can adjust the fusion ratio of the textual information containing the change entity and the visual information representing the change based on context when generating the next word.

The key contributions of this work are threefold: (1) We propose a novel CHEERS that identifies changed objects from a textual perspective, providing explicit guidance for representation learning. Further, CHEERS uses HRD to effectively separate differences and similarities while mitigating viewpoint variations and enhancing subtle change perception. (2) We design a gated Transformer decoder, which dynamically adjusts the fusion of textual and visual information based on context, prioritizing textual entity information for subject descriptions and visual features for change details. (3) Extensive experiments on the three public datasets demonstrate that our approach significantly outperforms state-of-the-art change captioning models.

2 Related Work

Change Captioning is a task that aims to generate natural language descriptions of the differences between two images representing a scene before and after a change. Early works, such as Jhamtani (Jhamtani and Berg-Kirkpatrick, 2018), approach this task by approximating object-level differences through pixel-wise clustering based on the difference between images. Park (Park et al., 2019) uses dynamic attention maps to localize the changes, while Shi (Shi et al., 2020) extracts both changed and unchanged features to input into a caption decoder. However, in real-world scenarios, viewpoint variation often introduces interference, reducing the model's ability to accurately identify changes. To enhance the robustness of models to such viewpoint changes, Tu (Tu et al., 2023b) designs neighboring feature aggregation to capture contextual information and common feature distillation to learn contrastive information between images. Liao (Liao et al., 2021) attempts to model the relative spatial relationships of objects in a 3D scene to eliminate interference based on this contextual information. Tu (Tu et al., 2023c) utilizes

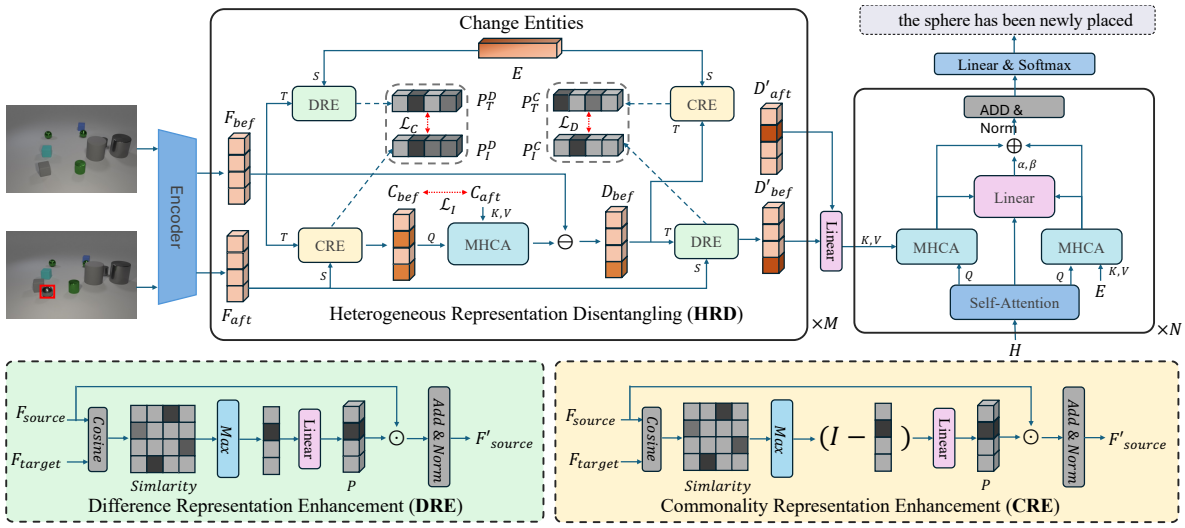

Figure 2: The overall architecture of the proposed Change Entity-guided Heterogeneous Representation Disentangling (CHEERS) network. The CHEERS primarily consists of multiple layers of HRD module and a gated transformer. Each HRD layer includes two CRE and two DRE.

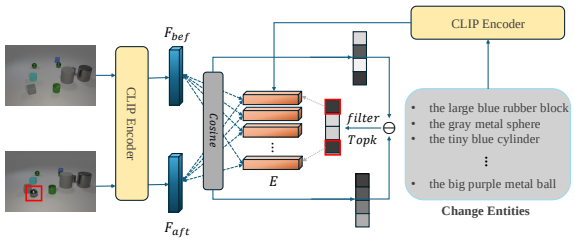

Figure 3: The process of change entity retrieval.

contrastive learning to align the representations of two images, thereby learning a stable difference representation. Additionally, to generate higher-quality captions, several works have attempted to incorporate additional information to assist the decoding process. Tu (Tu et al., 2021b) introduces part-of-speech information during decoding and uses a dynamic switch to control the fusion of this information. More recent works have leveraged large pre-trained LLM for this task. For instance, Hu (Hu et al., 2024) employs learnable query tokens that probe the multi-level encoded features of both images to effectively capture their differences and assist the LLMs in learning these differences. Zhang (Zhang et al., 2024) fine-tunes large models and incorporates a relevant corpus as additional assistance to generate more accurate captions.

Overall, previous works have primarily focused on identifying differences from visual information, generating difference representations, and then using a decoder to produce captions. In contrast, this paper shifts the focus to discovering differences from a textual perspective, leveraging additional textual information to guide the visual encoder in more accurately localizing differences. Furthermore, during the caption generation process, we fuse textual information to produce higher-quality descriptions.

3 Method

Given a pair of images (I_{bef}, I_{aft}), we first employ the Change Entity Retrieval Module, As shown in Figure 3, as to extract textual change entities, denoted as E, which provide explicit guidance for identifying key differences. Next, the Heterogeneous Representation Disentangling module processes (I_{bef}, I_{aft}) to separate difference features, denoted as (D_{bef}, D_{aft}) respectively. Finally, we utilize a gated Transformer decoder, which dynamically fuses E and D based on context to generate the final change description S_{cap}.

3.1 Change Entity Extraction and Retrieval

3.1.1 Change Entity Extraction

In change captioning, the caption typically focuses on the differences between two images, describing what has changed and how it has changed. Given a caption, the change entity generally corresponds to the subject of the sentence. In this study, we utilize SpaCy (Honnibal, 2017) to extract the subjects from captions. Our captions are collected from the corresponding training set. For instance, in experiments on the CLEVR-Change dataset (Park et al., 2019), we extract subjects from the training cap-

tions of this dataset. After obtaining all change entities, we extract semantic-level features E through pre-trained CLIP ViT-L/14 (Radford et al., 2021), which offers strong text-image alignment capabilities. These embeddings serve as the foundation for subsequent modules that localize and describe visual changes.

3.1.2 Change Entity Retrieval

After extracting the semantic embeddings E for all change entities, the next step is to retrieve the most relevant change entity that matches the given image pair (I_{bef}, I_{aft}). The goal is to find the change entity with the largest difference in cosine similarity between the two images, as this entity is likely to correspond to the true change in the scene. The main method is to identify the change entity with the largest difference in cosine similarity between the two images, as this entity generally corresponds to the true change in the scene.First, the two images are encoded by CLIP into feature representations denoted as (X_{bef}, X_{aft}). Then, the cosine similarity between each image feature and all change entity embeddings E is computed as $S_i \in \mathbb{R}^M$ where $i \in \{bef, aft\}$ represent the cosine similarity between the change entity embeddings E and the image features.Then select the most relevant change entity $\hat{E}$ by maximizing the difference between the cosine similarity scores of the change entity with the two images:

$$\hat{E} = \arg\max_{E} \left(S_{bef} - S_{aft}\right). \tag{1}$$

To ensure that the selected change entity representation is relevant, we introduce a constraint that at least one of the two cosine similarity scores S_i is higher than the average similarity $\bar{S}_i \in \mathbb{R}$.

3.2 Change Entity-guided Heterogeneous Representation Disentangling

3.2.1 Commonality Representation Enhancement Module

In the visual feature encoding stage, we design a representation enhancement module to disentangle the difference and common features between two images. The structure is illustrated in Figure 2. Given a target feature $F_{target} \in \mathbb{R}^{H\times W\times C}$ and a source feature $F_{source} \in \mathbb{R}^{H\times W\times C}$, the enhancement process is described as follows. First, the cosine similarity between each position in F_{target} and F_{source} is computed as $S \in \mathbb{R}$ denotes the similarity between position i in the target and position j in the source.Next, the similarity values are transformed into a probability map using a softmax function:

$$P(i,j) = \frac{\exp(\mathrm{Sim}(i,j))}{\sum_k \exp(\mathrm{Sim}(j,k))}. \tag{2}$$

To identify the parts of the target that have high similarity with the source, the maximum similarity across all positions in the source is computed and expanded to the same dimensions as F_{target} through a learnable linear layer:

$$\hat{P}(i) = \mathrm{Linear}(\max_j P(i,j). \tag{3}$$

Finally, a sigmoid function is introduced to control the scaling ratio, and a residual connection is added to prevent excessive information loss:

$$F'_{\text{target}} = \mathrm{LN}(\sigma(\hat{P}) \cdot F_{\text{traget}} + F_{\text{target}}), \tag{4}$$

where $\sigma(\cdot)$ denotes the sigmoid function and LN represents layer normalization. This design allows adaptive feature scaling while preserving the original visual information.

3.2.2 Difference Representation Enhancement Module

The difference enhancement module follows the same basic structure and operational process as the aforementioned CRE framework, with the only difference being the computation process of $\hat{P}$:

$$\hat{P}(i) = \mathrm{Linear}(I - \max_j P(i,j)). \tag{5}$$

This method emphasizes the difference between the two representations rather than the similar parts.

3.2.3 Heterogeneous Representation Disentangling

During the visual encoding process, we primarily use the aforementioned representation enhancement module to decouple and highlight the difference features. The structure is illustrated in Figure 2. The input images I_{bef}, $I_{aft} \in \mathbb{R}^{C\times H\times W}$ are first processed by a ResNet backbone to extract the raw feature representations F_{bef}, $F_{aft} \in \mathbb{R}^{C'\times H'\times W'}$.Then, we feed the two raw features into a CRE to highlight the common parts between them and we also indirectly enhance image representations using the change entities through the DRE:

$$\begin{aligned} C_i &= \mathrm{CRE}(F_j, F_i), \\ C'_i &= \mathrm{DRE}(E, F_i), \end{aligned} \tag{6}$$

where $C_i, C_i' \in \mathbb{R}^{C\times H\times W}$. However, instead of directly using the output features, we apply the probability matrix $\hat{P}$ in module to enforce consistency on the CRE as follows:

$$\mathcal{L}_C = \text{MSE}(\hat{P_C}, \hat{P_C'}), \tag{7}$$

where $\hat{P_C}$ donates the probability matrix in CRE and $\hat{P_C'}$ donates the probability matrix in DRE. After highlight their common parts, multi-head cross-attention (Vaswani, 2017) is applied between the enhanced common features to model interactions between the two images:

$$\tilde{C}_i = \text{MHCA}(C_i, C_j), \tag{8}$$

where MHCA represents the multi-head cross-attention. Inspired by previous works (Tu et al., 2023c, 2024b), contrastive learning is introduced during the computation of cross-attention to further help it obtain stable change representations, with the loss function being InfoNCE loss:

$$\mathcal{L}_I = -\log \frac{\exp(\text{sim}(q, k^+)/\tau)}{\sum_{i=1}^{N} \exp(\text{sim}(q, k_i)/\tau)} \tag{9}$$

Then the difference between the attended features and the original raw features is computed and further enhanced by DRE. In a similar manner, we use the change entity to further enhance the differences:

$$\begin{aligned} D_i &= \text{DRE}(F_i, F_i - \tilde{C}_i). \\ D_i' &= \text{CRE}(E, F_i - \tilde{C}_i), \end{aligned} \tag{10}$$

then enforce consistency on the DRE through the probability matrix:

$$\mathcal{L}_D = \text{MSE}(\hat{P_D}, \hat{P_D'}). \tag{11}$$

Finally, we fuse the two difference representations through a linear layer:

$$F_{diff} = \text{Linear}(\text{Concat}[D_{bef}; D_{aft}]), \tag{12}$$

3.3 Gated Transformer Decoder

After obtaining the visual difference features F_{diff}, the gated mechanism is applied to dynamically combine them with the textual change entity information E during caption generation. The decoder first processes its hidden states H through a self-attention mechanism:

$$H_{n-1}' = \text{SelfAttention}(H_{n-1})$$

where H_{n-1}' represents the updated decoder hidden states after self-attention.Next, the updated hidden statesH'Hare used in multi-head attention mechanisms with both the textual change entity features E and the visual difference features F_{diff} . Specifically, we compute:

$$\begin{aligned} H_{n-1}^T &= \text{MHCA}(H_{n-1}', E, E) \\ H_{n-1}^V &= \text{MHCA}(H_{n-1}', F_{diff}, F_{diff}) \end{aligned} \tag{13}$$

These operations allow the model to attend to both the textual and visual information based on the updated hidden states from the self-attention. Then, the attention outputs H^T and H^V are combined through a learnable weighting mechanism. We use a linear layer to generate two parameters, α and β , that control the importance of each attention output and the final feature representation H_{n-1}^F is then computed as a weighted sum of H^T and H^V :

$$\alpha = \text{Linear}(\text{Concat}([H_{n-1}'; H_{n-1}^T])) \tag{14}$$

$$\beta = \text{Linear}(\text{Concat}([H_{n-1}'; H_{n-1}^V])) \tag{15}$$

$$H_{n-1}^F = \alpha \cdot H_{n-1}^T + \beta \cdot H_{n-1}^V \tag{16}$$

Finally, this combined feature is added to the residual connection and passed through a normalization layer to produce the updated hidden states:

$$H_n = \text{LN}(H_{n-1}^F + H_{n-1}') \tag{17}$$

The final output at each decoding step is then computed by passing through a Linear layer and a softmax layer to predict the next word in the caption.

3.4 Joint Training

The overall training of the proposed network follows an end-to-end approach, where the goal is to maximize the likelihood of generating the correct word sequence. Given the ground-truth sequence of words $(w_1, \ldots, w_m)$, the network is trained by minimizing the negative log-likelihood loss function:

$$L_{\text{S}}(\theta) = -\sum_{t=1}^{m} \log p_\theta(w_t^* | w_{<t}^*). \tag{18}$$

In this equation, $p_\theta(w_t^* | w_{<t}^*)$ is the predicted probability for the t-th word given all the previous words. Here, θ represents the parameters of the network. In addition to this standard captioning loss, the model incorporate two alignment losses and a contrastive loss. These losses help the model learn better feature alignments between visual and textual

representations. The final loss function combines the captioning loss with these contrastive losses:

$$L = L_S + \lambda_v(L_C + L_D) + \lambda_m L_I, \quad (19)$$

where λ_v and λ_m are scalar trade-off parameters that control the relative importance of the losses, which are explained in further detail in the supplementary material.

4 Experiments

4.1 Datasets

CLEVR-Change: This large-scale dataset (Park et al., 2019) focuses on moderate viewpoint changes. It consists of 79,606 image pairs across five change types: "Color", "Texture", "Add", "Drop", and "Move". We use the official dataset split, with 67,660 pairs for training, 3,976 for validation, and 7,970 for testing.

CLEVR-DC: A more challenging dataset (Kim et al., 2021) that includes extreme viewpoint shifts. It contains 48,000 image pairs with the same change types as CLEVR-Change. The official split is used, with 85% for training, 5% for validation, and 10% for testing.

Spot-the-Diff: A dataset (Jhamtani and Berg-Kirkpatrick, 2018) of 13,192 aligned image pairs taken from surveillance cameras. Following standard practices, we evaluate our model on a single-change task. The dataset is split into training (80%), validation (10%), and testing (10%).

4.2 Evaluation Metrics

We evaluate the quality of the generated sentences using five standard metrics: BLEU-4 (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ROUGE-L (Lin, 2004), CIDEr (Vedantam et al., 2015), and SPICE (Anderson et al., 2016). All results are computed through the Microsoft COCO evaluation server (Chen et al., 2015), providing a consistent and standardized evaluation across different models.

4.3 Implementation Details

For feature extraction, we utilize ResNet-101 (He et al., 2016) pre-trained on the Imagenet dataset (Deng et al., 2009). Specifically, we extract features from the convolutional layers, yielding a tensor of size 14×14. To handle these features, we set the hidden dimension of our model to 512.

During training, we adjust the minibatch sizes based on the dataset: 128 for CLEVR-Change, 128 for CLEVR-DC and 96 for Spot-the-Diff. We employ the Adam optimizer (Kingma, 2014) with different learning rates for each dataset, specifically 1×10^{-3} for CLEVR-Change, 1×10^{-3} for CLEVR-DC and 5×10^{-4} for Spot-the-Diff. In the inference phase, we adopt a greedy decoding strategy to generate captions from the model outputs. All experiments are carried out using PyTorch (Paszke et al., 2019) and run on one RTX3090 GPU to ensure efficient training and testing.

4.4 Performance Comparison

4.4.1 Results on CLEVR-Change

In this experiment, we compare our approach with existing state-of-the-art methods and the results are summarized in Table 1. It is evident that our method consistently outperforms existing Transformer-based decoder models across all evaluation metrics, particularly in B, M, R metrics, indicating that our model is effective at decoupling differences and similarities. Compared to the LLM-based FINER-MLLM, our method significantly outperforms it in B, M, R, S metrics, highlighting that our approach can more accurately pinpoint differences, even without relying on large models.

In the case of semantic changes alone, it can be observed that, our model outperforms existing models across all metrics. This is primarily due to the use of DRE, where the change entity acts as a guide to strengthen the representation of differences, allowing our model to more accurately locate differences even under the interference introduced by changes in perspective.

4.4.2 Results on CLEVR-DC

To evaluate the model's performance under extreme viewpoint changes, we conduct experiments on the recently released CLEVR-DC dataset, which primarily consists of image pairs with significant viewpoint variations. In this experiment, we compare our approach with state-of-the-art methods and the results are summarized in Table 2. It is clear that our model significantly outperforms existing methods in the R, C, S metric, demonstrating stronger robustness to viewpoint changes compared to prior works. This improvement can be largely attributed to the DRE guided by the change entities, which effectively emphasizes the difference features between two images. Additionally, the decoder, which integrates textual information, enhances the generation of more accurate captions. This combination enables our model to capture and

Method	Total Performance					Semantic Change				
	B	M	R	C	S	B	M	R	C	S
DUDA (Park et al., 2019)	47.3	33.9	-	112.3	24.5	42.9	29.7	-	94.6	19.9
IFDC (Huang et al., 2021)	49.2	32.5	69.1	118.7	-	47.2	29.3	63.7	105.4	-
DUDA+ (Hosseinzadeh and Wang, 2021)	51.2	37.7	70.5	115.4	31.1	49.9	34.3	65.4	101.3	27.9
VACC (Kim et al., 2021)	52.4	37.5	-	114.2	31.0	-	-	-	-	-
SRDRL (Tu et al., 2021b)	54.9	40.2	73.3	122.2	32.9	52.7	36.4	69.7	114.2	30.8
R^3Net (Tu et al., 2021a)	54.7	39.8	73.1	123.0	32.6	52.7	36.2	69.8	116.6	30.3
BiDiff (Sun et al., 2022)	54.2	38.3	-	118.1	31.7	-	-	-	-	-
IDC-PCL (Yao et al., 2022)	51.2	36.2	71.7	128.9	-	-	-	-	-	-
NCT (Tu et al., 2023b)	55.1	40.2	73.8	124.1	32.9	53.1	36.5	70.7	118.4	30.9
VARD (Tu et al., 2023a)	55.2	40.8	74.1	124.1	33.3	53.6	36.7	71.0	119.1	30.5
SCORER (Tu et al., 2023c)	56.3	41.2	74.5	126.8	33.3	54.4	37.6	71.7	122.4	31.6
SMART (Tu et al., 2024b)	56.1	40.8	74.2	127.0	33.4	54.3	37.4	71.8	123.6	32.0
DIRL+CCR (Tu et al., 2024a)	-	-	-	-	-	**54.6**	38.1	71.9	123.6	31.8
FINER (Zhang et al., 2024)	55.6	36.6	72.5	**137.2**	26.4	-	-	-	-	-
CHEERS (Ours)	**57.1**	**42.2**	**75.8**	130.0	**33.9**	54.1	**39.0**	**73.6**	**127.4**	**33.2**

Table 1: Comparing with state-of-the-art methods on CLEVR-Change dataset.

Method	B	M	R	C	S
DUDA (Park et al., 2019)	40.3	27.1	-	56.7	16.1
VAM (Shi et al., 2020)	40.9	27.1	-	60.1	15.8
VACC (Kim et al., 2021)	45.0	29.3	-	71.7	17.6
NCT (Tu et al., 2023b)	47.5	32.5	65.1	76.9	15.6
VARD (Tu et al., 2023a)	48.3	32.4	-	77.6	15.4
SCORER (Tu et al., 2023c)	49.4	**33.4**	66.1	83.7	16.2
DIRL+CCR (Tu et al., 2024a)	51.4	32.3	66.3	84.1	16.8
CHEERS (Ours)	**51.6**	32.7	**66.8**	**86.9**	**17.0**

Table 2: Comparing with state-of-the-art methods on CLEVR-DC dataset.

Method	B	M	R	C	S
DUDA (Park et al., 2019)	8.1	11.8	29.1	32.5	-
VAM (Shi et al., 2020)	10.1	12.4	31.3	38.1	-
VACC (Kim et al., 2021)	9.7	12.6	32.1	**41.5**	-
VARD (Tu et al., 2023a)	-	12.5	29.3	30.3	17.3
SCORER (Tu et al., 2023c)	10.2	12.2	-	38.9	18.4
DIRL+CCR (Tu et al., 2024a)	10.3	13.8	32.8	40.9	**19.9**
CHEERS (Ours)	**10.5**	**12.9**	**33.1**	41.0	19.6

Table 3: Comparing with state-of-the-art methods on Spot-the-Diff dataset.

highlight subtle changes in images, making it superior to previous approaches.

4.4.3 Results on Spot-the-Diff

To evaluate the expressive capability of our model in real-world scenarios, we conduct experiments on the recently released Spot-the-Diff dataset, which primarily consists of well-aligned image pairs without any viewpoint changes. In this setup, we compare our approach against state-of-the-art methods. As shown in Table 3 our model achieves improvements across various metrics compared to these

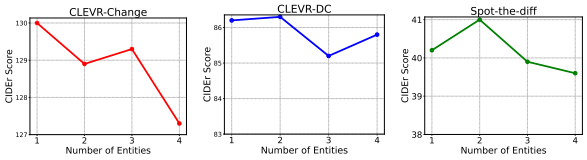

Figure 4: Effect of number of entities on three datasets.

models. This indicates that our model can still accurately describe differences in more complex scenes. Since the dataset contains diverse statements but is not large, the models struggle to learn the semantic information of less frequent words. However, with the help of prior semantic information about the change entity, CHEERS can more accurately describe the changes present in such scenarios.

Method	B	M	R	C	S
Baseline	42.2	34.7	67.5	100.1	28.6
HDR	56.3	41.2	74.3	125.3	32.9
GATE	43.9	36.2	69.8	104.4	29.1
CHEERS	**57.1**	**42.2**	**75.8**	**130.0**	**33.9**

Table 4: Ablation study of each module on CLEVR-Change dataset.

4.5 Ablation studies

Ablation Study of Each Module.To assess the contribution of each module, we conduct the following ablation studies on CLEVR-Change. Table 4 shows the overall performance of each component of the proposed method across the entire dataset and only scene changes. It is evident that each

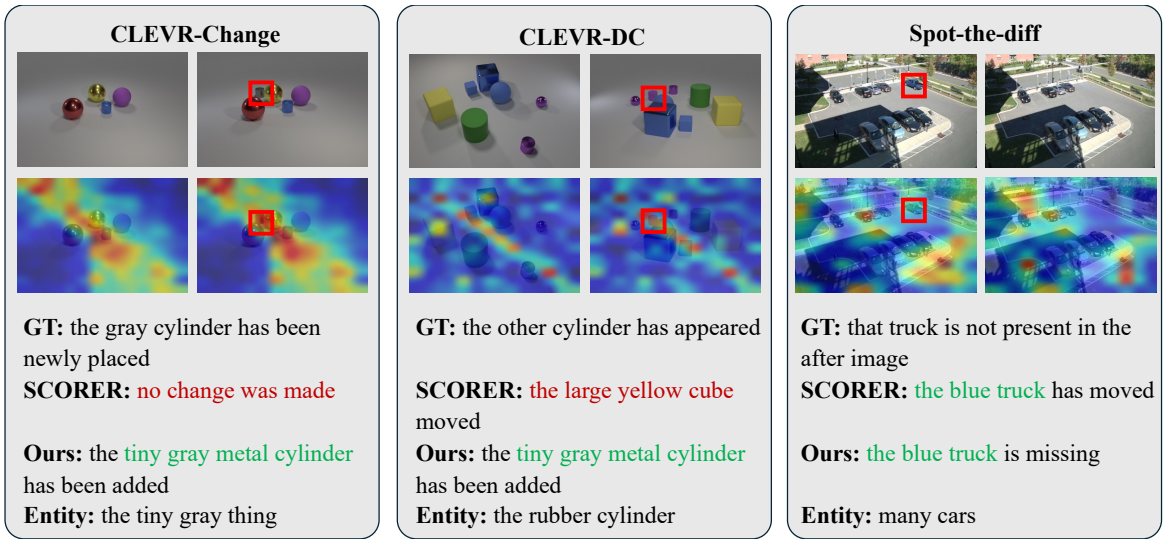

Figure 5: Qualitative analysis between the state-of-the-art method SCORER (Tu et al., 2023c) and our CHEERS on the CLEVR-Change, CLEVR-DC, and Spot-the-diff datasets.

module contributes to enhancing the baseline performance. Furthermore, the best performance is achieved when all modules are combined, demonstrating that each component not only fulfills its unique role but also complements the others. This indicates that, with the guidance of the change entity, the model can more accurately pinpoint differences and generate higher-quality captions.

Ablation Study of Number of Entities. We conduct an ablation study on the number of entities used in our model, as illustrated in Figure 4 Across three different datasets, we observe that simply increasing the number of entities does not lead to significant performance improvements. In fact, having too many entities can make it difficult for the model to focus on the true change targets, as the increased variation in the entities may distract the model from capturing the most relevant changes. Based on these findings, we set the number of entities to optimal values of 3 allowing the model to focus on the true change targets without being distracted by irrelevant variations.

4.5.1 Qualitative Analysis

Figure 5 presents three representative examples from CHEERS, evaluated against the baseline model SCORER, across three different datasets. Each example shows the ground-truth change captions alongside those generated by CHEERS and SCORER, with changed regions highlighted by red boxes. Besides, we also present the change entity retrieved by our model to demonstrate its correlation with the real changes. Additionally, to observe whether the model can focus on the difference regions, we illustrate the attention distributions and visualize them as a heatmap.

Upon reviewing the descriptions generated by both methods in Figure 5, it becomes clear that our model outperforms SCORER in recognizing subtle differences, and it demonstrates greater robustness in handling extreme viewpoint changes. The heatmap analysis reveals that our model effectively focuses on the different objects across the paired images, highlighting its attention to key details. Moreover, it can be observed that the extracted entities have a high correlation with the ground truth, which further validates the approach of using entities to guide the model in identifying the differences, reinforcing the practicality and effectiveness of this strategy.

5 Conclusion

This paper proposes CHEERS, which leverages change entities to guide difference localization and caption generation. CHEERS first determines the change entity by maximizing the similarity difference between two images and candidate subjects. Then, guided by the change entity, a representation enhancement mechanism is applied to disentangle difference features from distraction. Additionally, we design a gated transformer that dynamically fuses visual difference information with the textual change entity features. Extensive experiments show that CHEERS achieves state-of-the-art results on multiple benchmark datasets, demonstrating its effectiveness in various change scenarios.

Limitations

We propose a novel model, CHEERS, aimed at generating higher-quality text and having stronger robustness in the change captioning task. Although our model achieves state-of-the-art performance on several public datasets, there is still room for improvement. In the entity retrieval and encoding stage, we primarily use CLIP, which is not sensitive to numerical and spatial relationships. More powerful models could be used for entities retrieval or to generate prior textual information.

References

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 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*, pages 382–398. Springer.

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.

Wenting Chen, Linlin Shen, Jingyang Lin, Jiebo Luo, Xiang Li, and Yixuan Yuan. 2024. Fine-grained image-text alignment in medical imaging enables explainable cyclic image-report generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9494–9509.

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*.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Matthew Honnibal. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *(No Title)*.

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

Erdong Hu, Longteng Guo, Tongtian Yue, Zijia Zhao, Shuning Xue, and Jing Liu. 2024. Onediff: A generalist model for image difference captioning. In *Proceedings of the Asian Conference on Computer Vision*, pages 2439–2455.

Qingbao Huang, Yu Liang, Jielong Wei, Yi Cai, Hanyu Liang, Ho-fung Leung, and Qing Li. 2021. 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. *arXiv preprint arXiv:1808.10584*.

Hoeseong Kim, Jongseok Kim, Hyungseok Lee, Hyunsung Park, and Gunhee Kim. 2021. Agnostic change captioning with cycle consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2095–2104.

Minkuk Kim, Hyeon Bae Kim, Jinyoung Moon, Jinwoo Choi, and Seong Tae Kim. 2024. Do you remember? dense video captioning with cross-modal memory retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13894–13904.

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

Zeming Liao, Qingbao Huang, Yu Liang, Mingyi Fu, Yi Cai, and Qing Li. 2021. Scene graph with 3d information for change captioning. In *Proceedings of the 29th ACM international conference on multimedia*, pages 5074–5082.

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

Chenyang Liu, Keyan Chen, Bowen Chen, Haotian Zhang, Zhengxia Zou, and Zhenwei Shi. 2024. Rscama: Remote sensing image change captioning with state space model. *IEEE Geoscience and Remote Sensing Letters*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Dong Huk Park, Trevor Darrell, and Anna Rohrbach. 2019. Robust change captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 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. *Advances in neural information processing systems*, 32.

Yue Qiu, Shintaro Yamamoto, Kodai Nakashima, Ryota Suzuki, Kenji Iwata, Hirokatsu Kataoka, and Yutaka Satoh. 2021. Describing and localizing multiple changes with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1971–1980.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 574–590. Springer.

Yaoqi Sun, Liang Li, Tingting Yao, Tongyv Lu, Bolun Zheng, Chenggang Yan, Hua Zhang, Yongjun Bao, Guiguang Ding, and Gregory Slabaugh. 2022. Bidirectional difference locating and semantic consistency reasoning for change captioning. *International Journal of Intelligent Systems*, 37(5):2969–2987.

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

Yunbin Tu, Liang Li, Li Su, Ke Lu, and Qingming Huang. 2023b. Neighborhood contrastive transformer for change captioning. *IEEE Transactions on Multimedia*, 25:9518–9529.

Yunbin Tu, Liang Li, Li Su, Chenggang Yan, and Qingming Huang. 2024a. Distractors-immune representation learning with cross-modal contrastive regularization for change captioning. In *European Conference on Computer Vision*, pages 311–328. Springer.

Yunbin Tu, Liang Li, Li Su, Zheng-Jun Zha, and Qingming Huang. 2024b. 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 *CVF International Conference on Computer Vision (ICCV) pp*, pages 2793–2803.

Yunbin Tu, Liang Li, Chenggang Yan, Shengxiang Gao, and Zhengtao Yu. 2021a. R^3Net: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 the association for computational linguistics: ACL-IJCNLP 2021*, pages 63–73.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

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

Linli Yao, Weiying Wang, and Qin Jin. 2022. Image difference captioning with pre-training and contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3108–3116.

Xian Zhang, Haokun Wen, Jianlong Wu, Pengda Qin, Hui Xue', and Liqiang Nie. 2024. Differential-perceptive and retrieval-augmented mllm for change captioning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 4148–4157.