Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

Find and Perceive: Tell Visual Change with Fine-Grained Comparison

Feixiao Lv[1,2] , **Rui Wang**[1,2*] , **Lihua Jing**[1,2] and **Lijun Liu**[1,2]
[1]Institute of Information Engineering, CAS, Beijing, China
[2]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
{lvfeixiao, wangrui, jinglihua, liulijun}@iie.ac.cn

Abstract

The goal of the image change captioning task is to capture the differences between two similar images and describe them in natural language. In this paper, we decompose this task into two sub-problems, i.e., fine-grained change feature learning and discrimination of changed regions. Compared with existing methods which only focus on change feature learning, we propose a novel change captioning learning paradigm, Find and Perceive (F&P). Our proposed F&P consists of two main ideas, i.e., the Fine-Grained Semantic Change Perception (FGSCP) module for improving the model's perception ability of subtle changes and the Weakly-Supervised Discriminator (WSD) of changed regions for improving the model's sensitivity of localising the important regions. Specifically, the FGSCP deploys a two-step manner, firstly introducing the fine-grained categorisation and then enhancing the interaction of the two paired images. And the WSD adopts the contributions of each image region for final generated captions, accurately indicating which regions are important for change captions without any extra annotations. Finally, we conduct extensive experiments on four change captioning datasets, and experimental results show that our proposed method F&P outperforms existing change caption methods and achieves new state-of-the-art performance.

1 Introduction

Image change captioning aims at telling changes before and after two images and describing them in natural language [Jhamtani and Berg-Kirkpatrick, 2018; Kim *et al.*, 2021]. Compared with traditional image captioning task [Farhadi *et al.*, 2010; Vinyals *et al.*, 2015] which focuses on describing one single static image, image change captioning not only describes the content of images but also needs to tell the changes between two similar images. Due to its wide application in practice, such as outputting logs about monitored areas and generating reports about pathological changes, image change captioning has attracted a lot of attention from both industry and scholars. With the rapid deepening of change captioning, this task is developing from describing the changes of simple geometry to describing the difference between real objects, such as birds in Figure 1. Due to the changed images being highly similar, image change captioning is challenging. The challenges mainly come from two aspects. Firstly, the model should understand the fine-grained semantic concepts. As shown in Figure 1, the change captioning task needs to focus on fine-grained changes (e.g., colour, shape, and size) and describe the differences. Secondly, the model should find the subtle differential regions. As shown in Figure 1, the differential regions marked with red boxes should be the focus of the model, while the unchanged regions marked with blue boxes should be seen as irrelevant distractions.

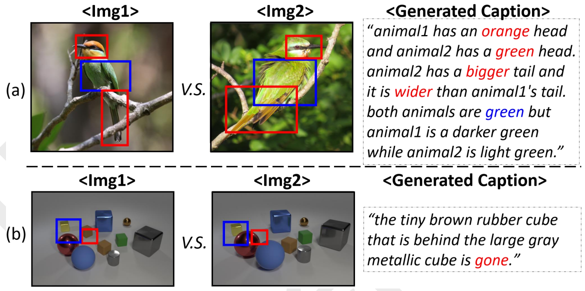

<Img1>
<Img2>
<Generated Caption>
(a)
V.S.
"animal1 has an orange head and animal2 has a green head. animal2 has a bigger tail and it is wider than animal1's tail. both animals are green but animal1 is a darker green while animal2 is light green."
<Img1>
<Img2>
<Generated Caption>
(b)
V.S.
"the tiny brown rubber cube that is behind the large gray metallic cube is gone."

Figure 1: The examples of image change captioning. (a) An example from Birds-to-Words dataset. (b)An example from CLEVR-Change dataset. The red box indicates fine-grained regions that need attention, and the blue box indicates regions that should not be noticed, respectively.

Existing image change captioning methods tend to focus on the former, i.e., improving the model's ability of perception for subtle fine-grained changes. They often regard the pre-trained paired image features as a whole to extract similarity and use these coarse features to generate comparative descriptions. For example, a recent work VACC [Kim *et al.*, 2021] obtains the overall similarity of features to pinpoint the semantic changes, which replaces the task of extracting differences with extracting features. However, for two similar images, the method of "extracting features as a whole before generating descriptions" lacks a step of "finding differences". On the contrary, as [He *et al.*, 2019] studies, when humans

[*]Corresponding author

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

look for differences, they can easily focus on different points of fine granularity on the basis of understanding the whole image. On the one hand, they can meticulously recognise the key features of each image. On the other hand, they can accurately locate discriminative regions.

Inspired by the "tell changes" process of human-being, we decompose the image change captioning into two sub problems, i.e., fine-grained change feature learning and discrimination of changed regions, and propose a novel change captioning learning paradigm, Find and Perceive (F&P). Our proposed method (F&P) consists of two main components: the Fine-Grained Semantic Change Perception module and the Weakly-Supervised Discriminator for changed regions. Specifically, the proposed *Fine-Grained Semantic Change Perception* module improves the model's perception ability of subtle changes with two steps. The first step is introducing fine-grained categorisation to deploy the Fine-Grained Feature Learning, which helps the model learn fine-grained subtle patterns. Then, the second step is modelling the visual comparison process with the Difference Enhance module, which improves the model's ability of describing changes by interaction between fine-grained features of the two paired images. To help the change captioning model find the changed regions, we proposed the *Weakly-Supervised Discriminator* for localising the changed regions (i.e., we only use the text change label without any visual localisation labels such as bounding boxes, key points, and so on). Specifically, we utilise the contributions of each patch for final generated captions, constructing the pseudo labels to train the Weakly-Supervised Discriminator. By doing this, the Weakly-Supervised Discriminator could accurately tell the model which regions have changed, and therefore the generated change captions are improved. The main contributions of our paper can be summarised as follows:

- We decompose the image change captioning into two sub-problems, i.e., fine-grained change feature learning and discrimination of changed regions, and propose a novel change captioning learning paradigm Find and Perceive (F&P).
- We propose the Fine-Grained Semantic Change Perception module, which improves the model's perception ability of subtle changes with two steps.
- We propose the Weakly-Supervised Discriminator for localising the changed regions, without using any extra localisation labels.
- We conduct extensive experiments on the two commonly used Brids-to-words and CLEVR-Change datasets, and experimental results show that our proposed F&P surpasses existing image change captioning methods and achieves new state-of-the-arts.

2 Related Works

Our work is related to two lines of research, namely, image change captioning and vision-language pre-training. We briefly review relevant work on these topics.

2.1 Image Change Captioning

Image change captioning is a new task of describing differences between similar image pairs with natural language. Compared with the image captioning task, image change captioning promotes a fine-grained understanding of image vision and language, making it more challenging. DDLA [Jhamtani and Berg-Kirkpatrick, 2018] first proposes the change captioning task and publishes the Spot-the-diff dataset extracted from the VIRAT dataset. They propose a change caption model to capture visual saliency by aligning clusters of different pixels with output sentences using potential variables. To address the viewpoint change problem in the change captioning task, DUDA [Park *et al.*, 2019] proposes a robust model for distractions in the sense that it can distinguish relevant scene changes from viewpoint changes. In addition, they build a synthetic dataset with a viewpoint change between each image pair based on the CLEVR engine. Besides, more recent work [Hosseinzadeh and Wang, 2021] formulates a training scheme to improve the training accuracy of the existing change captioning network by means of an auxiliary task. And an extra level of supervision is also provided through their training scheme. With the rise of attention mechanism [Vaswani *et al.*, 2017], Transformer-based method is becoming more and more popular. To enhance the adaptability of change captioning models to complex scenarios, the method [Qiu *et al.*, 2021] proposes a Transformer-based network to describe more than one change between the image pair. VARD-Transformer [Tu *et al.*, 2023a] assists the network to perceive the change information by proposing a novel position encoding mode. NCT [Tu *et al.*, 2023b] proposes a Transformer-based architecture to enhance the discernment of the network by generating new attention states for each token. Thus, it performs well in both change captioning and localisation.

However, they all tend to focus on learning the changed features by regarding the image as a whole without selection, ignoring finding the changed regions is important for change captioning. In this paper, we not only improve the model's understanding of fine-grained patterns but also propose a weakly-supervised mechanism to improve model's sensitivity to changed regions.

2.2 Vision and Language Pre-training

Since the birth of ViLBERT [Lu *et al.*, 2019] and LXMERT [Tan and Bansal, 2019], vision and language pre-training (VLP) models have demonstrated powerful performance in multi-task and multi-modal learning. The popularity of visual-language pre-training models is gradually rising [Addepalli *et al.*, 2024; Hu *et al.*, 2022]. Prominent examples include UNITER [Chen *et al.*, 2020], OSCAR [Li *et al.*, 2020c], UNIMO [Li *et al.*, 2020b], and PaLM-E [Driess *et al.*, 2023]. Especially in the field of image captioning, along the journey of VLP, researchers have investigated different training strategies [Zeng *et al.*, 2023], robustness [Li *et al.*, 2020a], discrimination [Dessì *et al.*, 2023], and the extension [Hirota *et al.*, 2023]. Due to the lack of understanding and learning of comparison features, VLP models are difficult to focus on fine-grained difference features, which cannot be successfully applied directly to change captioning tasks.

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

To solve this problem, PLC [Yao *et al.*, 2022] designs three self-supervised tasks and comparative learning strategies to align visual differences and text descriptions. Their work formally introduces VLP models into the change captioning task and focuses on fine-grained features. Although their work has achieved the alignment of fine-grained feature modes, they have not gone further to find different fine-grained regions. Motivated by these works, we aim to excavate the potential of VLP models for changed fine-grained regions in image change captioning task.

3 Method

As shown in Figure 2, our proposed F&P firstly deploys fine-grained feature learning with the proposed Fine-Grained Semantic Change Perception, and then a multi-layer Transformer is adopted to fuse text and visual features. Finally, the Weakly-Supervised Discriminator is utilised to improve the model's ability of finding the changed regions.

3.1 Model Architecture

Input Representation

As the general vision and language training model, our F&P input are the representations of a changed text caption and a pair of similar images, which form a triplet $[\mathbf{S}, \mathbf{I}^1, \mathbf{I}^2]$. We use word2vec [Mikolov *et al.*, 2013] to encode each word in the caption and express it as:

$$\mathbf{S} = \{[\text{SEC}], [\text{BOS}], w_0, \ldots, w_T, [\text{EOS}]\}, \tag{1}$$

where the token [SEC] is a sentence mark, which is also used to capture the global feature of the caption. [BOS] and [EOS] represent the beginning and end of the caption, respectively. The other two $\mathbf{I}^1$ and $\mathbf{I}^2$ in the triplet are used to represent the features of paired images. We use the pre-trained ResNet101 [He *et al.*, 2016] to extract the grid features of the paired images input and express them as:

$$\begin{aligned} \mathbf{I}^1 &= \{[\text{IMG1}], i_0^1, \ldots, i_M^1\}, \\ \mathbf{I}^2 &= \{[\text{IMG2}], i_0^2, \ldots, i_M^2\}, \end{aligned} \tag{2}$$

where the two tokens [IMG1] and [IMG2] are markers of paired images, which are used to capture the global features of two images at the same time. Then we use a linear layer to make the dimensions of $\mathbf{I}^1$ and $\mathbf{I}^2$ be the same as $\mathbf{S}$. In addition, in order to better process the positional information of sequence modality triplet, we also introduce positional encoding [Vaswani *et al.*, 2017] to each input token.

Caption Generator

As shown in Figure 2, we use a large pre-trained multi-layer Transformer model to encode and align text and image features. The Transformer model consists of multiple layers of self-attention Transformer blocks, which are used for integrating and encoding image and text features. The multi-layer Transformer model takes $[\mathbf{S}, \mathbf{E}^1, \mathbf{E}^2]$ as input and generates the final caption and image representations, where $\mathbf{E}^i$ indicates the enhanced feature of $\mathbf{I}^i$.

3.2 Fine-Grained Semantic Change Perception

The fine-grained features of images are the key factors in change captioning, which often contains the core content of changes that need to be described. Inspired by this, we design a Fine-Grained Semantic Change Perception (FGSCP) module to improve the understanding of detailed semantics and perceive the differences between images. As shown in Figure 2, our FGSCP adopts a two-step manner to achieve the idea mentioned above.

Fine-Grained Feature Learning

At the first step, we introduce the fine-grained categorisation to help the model learn the fine-grained patterns. Specifically, given the input visual features $\mathbf{I}^1$ and $\mathbf{I}^2$, we utilise a self-attention Transformer block [Vaswani *et al.*, 2017], $\mathcal{T}^1_{FG}(\cdot)$ and $\mathcal{T}^2_{FG}(\cdot)$, to produce the fine-grained features. This process can be formulated with:

$$\mathbf{V}^i = \mathcal{T}^i_{FG}(\mathbf{I}^i), \tag{3}$$

where $i \in \{1, 2\}$ indicates the i-th image, $\mathbf{V}^i$ is the i-th fine-grained feature.

Then, to guide this module to learn fine-grained patterns, we introduce a fine-grained categorisation task. Specifically, we feed the class token [IMGi] of fine-grained features $\mathbf{V}^i$ into a liner classifier to produce fine-grained classification results, which can be formulated with:

$$\mathbf{y}_c^i = LC([IMGi]), \tag{4}$$

where $\mathbf{y}_c^i$ indicates the classification result of the i-th image and $LC(\cdot)$ indicates the liner classifier. Then the cross entropy is utilised to train the feature extractor and linear classifier. With the guidance of fine-grained classification, the extracted features are stronger in representing subtle patterns, which is important for image change captioning as mentioned earlier in the paper.

Difference Enhancement

After extracting fine-grained features, we deploy the second step to enhance the difference between the paired images with the proposed Difference Enhancement module, which can be formulated as:

$$\begin{aligned} \{\mathbf{E}^1, \mathbf{E}^2\} &= DE(\{\mathbf{V}^1, \mathbf{V}^2\}), \\ &= \{\mathbf{V}^1 \oplus \mathcal{T}_c(\mathbf{V}^1, \mathbf{V}^2), \mathbf{V}^2 \oplus \mathcal{T}_c(\mathbf{V}^2, \mathbf{V}^1)\}, \end{aligned} \tag{5}$$

where $\mathbf{E}^i$ indicates the enhanced feature of the i-th image and $DE(\cdot)$ indicates the Difference Enhancement. $\mathcal{T}_c(\cdot)$ refers to the cross-attention Transformer block [Vaswani *et al.*, 2017], which is utilised to model the interaction of paired image features. Finally, we adopt a residual connection $\oplus$ to hold the original fine-grained features.

The core motivation of cross-attention in our Difference Enhancement module lies in that the unchanged regions of the paired image will produce high similarity, while the changed regions will produce low similarity. Therefore, the model could distinguish the changed and unchanged regions more accurately after our Difference Enhancement step.

Preprint

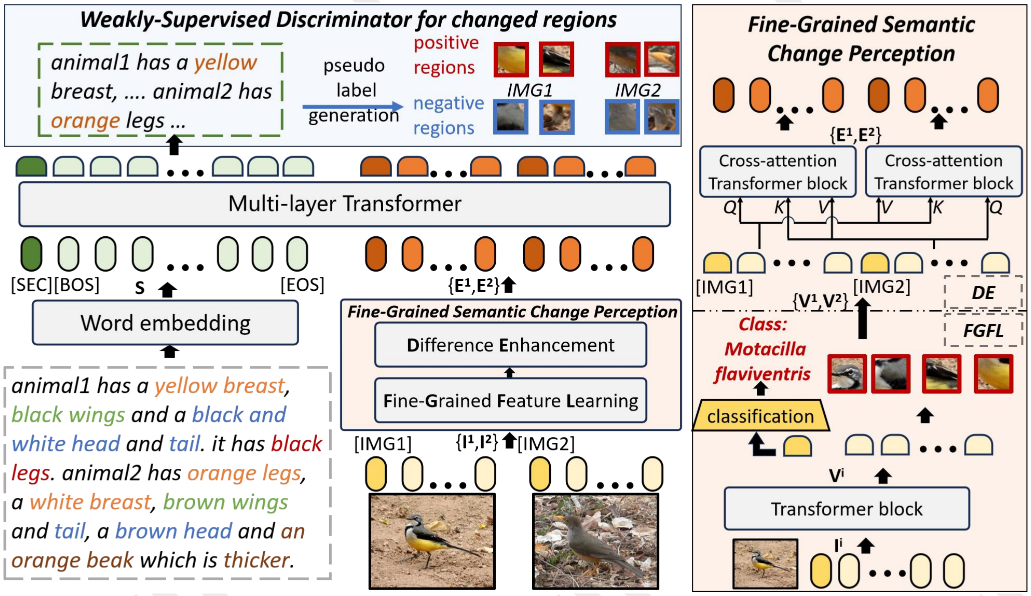

Weakly-Supervised Discriminator for changed regions
animal1 has a yellow breast, ... animal2 has orange legs ...
pseudo label generation
positive regions
negative regions
IMG1
IMG2
Multi-layer Transformer
[SEC][BOS]
S
[EOS]
{E¹,E²}
Word embedding
animal1 has a yellow breast, black wings and a black and white head and tail. it has black legs. animal2 has orange legs, a white breast, brown wings and tail, a brown head and an orange beak which is thicker.
Fine-Grained Semantic Change Perception
Difference Enhancement
Fine-Grained Feature Learning
[IMG1]
{I¹,I²}
[IMG2]
Fine-Grained Semantic Change Perception
{E¹,E²}
Cross-attention Transformer block
Cross-attention Transformer block
Q
K
V
V
K
Q
[IMG1]
[IMG2]
{V¹,V²}
DE
FGFL
Class: Motacilla flaviventris
classification
Vⁱ
Transformer block
Iⁱ

Figure 2: The overall workflow of our proposed F&P model. Our F&P first extracts fine-grained features and perceives the differences through Fine-Grained Semantic Change Perception (FGSCP). Then a multi-layer Transformer is introduced with Weakly-Supervised Discriminator (WSD) to align image and text features and generate change caption. The lower left part shows the main flow of F&P, and the proposed FGSCP and WSD are displayed in orange and blue, respectively.

3.3 Weakly-Supervised Discriminator

When generating change captions, the network does not need to pay attention to all fine-grained features but only needs to locate the changed regions, which have a significant impact on the paired images. For this purpose, we design a Weakly-Supervised Discriminator (WSD) for changed regions. We obtain the attention weights of each token of the last layer of the multi-layer Transformer to represent the important regions. For example, when the model in Figure 1 generates the word "orange", it will obtain weights on all tokens of $\mathbf{I}^1$. We define the regions with high weights as 1 and assume that these regions are changed regions.

To achieve this, we need to first generate a caption through the network. Therefore, we initially use an incompletely trained network to generate intermediate captions. After that, based on the feedback information from the captions, we determine the changed regions and set a threshold to set the most influential regions to 1 and the less influential regions to 0. In this way, we construct positive and negative regions as shown in Figure 2. To this end, we obtain a series of pseudo labels $\mathbf{C}_k$ for the k-th triplet sample, denoted as:

$$\begin{aligned} \mathbf{C}_k &= \{c_0^{k,1}, ...c_M^{k,1}, c_0^{k,2}, ...c_M^{k,2}\}, \\ c_j^{k,i} &= \mathbb{B}(a_j^{k,i}, t), \end{aligned} \tag{6}$$

where $\mathbb{B}$ indicates the binarization operation, which sets the pixels as 1 when their values are greater than the threshold t, otherwise, as 0. $a_j^{k,i} = \sum_{t=1}^{T} a_{j,t}^{k,i}$ indicates the sum of attention weight of the t-th token of intermediate caption on the j-th token of the i-th image. Afterwards, we design a discriminator on the output results of I^1 and I^2 based on the pseudo labels. It is worth noting that although the intermediate result captions may not be satisfying, the generation of pseudo labels is not a direct result, but rather a 0,1 vector generated after designing the threshold, which serves as a guiding role to assist the network in paying attention to areas with significant influence.

3.4 Training and Inference

In our proposed F&P, we design a three-step training paradigm to improve the captioning model step-by-step. Firstly, following previous work [Yao *et al.*, 2022], we conduct three well-known pre-training tasks to initialise the network, which are Masked Language Modelling (MLM), Masked Visual Contrastive Learning (MVCL), and Fine-grained Difference Aligning (FDA). By doing this, the model can learn the alignment of image and text. We set this to the baseline model.

Secondly, we apply our proposed FGSCP and WSD to improve the model's perception ability of subtle changes and sensitivity of localising changed regions, which is highly important for image change captioning. For the FGSCP, as mentioned earlier, we utilise the cross-entropy to train the fine-grained feature extractor and liner classifier, which is formulated with:

$$L_{FG} = CE(\mathbf{y}_c, \mathbf{y*}), \tag{7}$$

where $\mathbf{y}_c$ denotes the classification result, CE denotes the cross entropy loss and $\mathbf{y*}$ denotes the ground-truth class label provided in the change captioning dataset. For the Birds-to-Words [Forbes *et al.*, 2019] dataset, we perform fine-grained classification for the birds in the image. Then, the WSD is applied with the following loss function to train the whole network:

$$L_{WSD} = \begin{cases} -log(y_d^m), & c_m = 1 \\ -log(1 - y_d^m), & c_m = 0 \end{cases}$$

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

where the y_d^m denotes the discrimination result of the m-th region.

Finally, we apply the image change captioning task to make model obtain the ability to tell the perceived changes. During this stage, the whole network is trained with:

$$L_{cap}(\theta) = -\sum_{t=1}^{T} \log(p_\theta(y_t|y_{1:t-1}, I_1, I_2)), \tag{8}$$

where T is the length of sentence, p_θ is the probability of the output word at step t, and y_t denotes the caption result.

In the inference stage, the network can only see the paired images. Therefore, the network input includes all visual features I^1, I^2, and special tokens [SEC] of the text. The network generates a sentence "word by word", where [BOS] represents the beginning of the sentence and [EOS] represents the end of the prediction. During the prediction of the t-th word, the network can only see the first $t-1$ predicted words as input.

4 Experiments

4.1 Datasets and Metrics

We perform our main evaluation on two commonly used datasets, Birds-to-Words dataset [Forbes *et al.*, 2019] and CLEVR-Change [Park *et al.*, 2019] to verify the effectiveness of our method. In addition, we also compare our method with other methods on two additional datasets, Spot-the-Diff [Jhamtani and Berg-Kirkpatrick, 2018] and Image-Editing-Request [Tan *et al.*, 2019] to verify the generality of our method. **Birds-to-Words** [Forbes *et al.*, 2019] consists of 41k sentences that describe fine-grained changes between photographs of birds. **CLEVR-Change** [Park *et al.*, 2019] is a large-scale synthetic dataset with moderate viewpoint change. It has 79,606 image pairs and 493,735 captions. **Spot-the-Diff** [Jhamtani and Berg-Kirkpatrick, 2018] includes 13,192 aligned image pairs from surveillance cameras without fine-grained change but only obvious content changes. **Image-Editing-Request** [Tan *et al.*, 2019] includes 3,939 aligned image pairs with 5,695 editing instructions produced by image editing.

We evaluate the model performance using four most popular automatic language metrics: CIDEr [Vedantam *et al.*, 2015], BLEU-4 [Papineni *et al.*, 2002], METEOR [Banerjee and Lavie, 2005] and ROUGE-L [Lin and Hovy, 2003].

4.2 Implementation Detail

To compare our F&P with most other methods in a fair way, we perform our main evaluation with ResNet101 [He *et al.*, 2016] to extract paired image visual features and flatten it to the shape of (49, 2048). All the hidden size is 512. In addition, we adjust the input to CLIP features [Guo *et al.*, 2022], to compare with large model methods, and to extend our method to other less fine-grained datasets. For Transformer blocks, the attention head is set to 8, and layer number is set to 3 for multi-layer Transformer, 2 for fine-grained feature learning, 2 for different enhancement. To ensure stable and progressively refined pseudo label selection, we apply fixed thresholds to attention weights in each iteration (0.04 in the first and 0.06 in the second). These threshold values are determined based on experimental performance. We first perform the training process based on PLC [Yao *et al.*, 2022]. After that, we train the Fine-Grained Semantic Change Perception to improve the understanding of detail semantics. In the FGFL step, we perform fine-grained classification tasks for birds in Birds-to-Words dataset [Forbes *et al.*, 2019], and classify the change types for CLEVR-Change [Park *et al.*, 2019] dataset. Then we generate the preliminary results by first fine-tuning the change captioning task with 1 iteration and generating pseudo labels for WSD. Afterwards, we execute the second fine-tuning phase. In the fine-tuning stage, the learning rate is set as 3e-5. Early-stop is applied on the main metric to avoid overfitting.

Model	BLEU-4	METEOR	CIDEr	ROUGE-L
Neural Naturalist 2019	22.0	-	25.0	43.0
Relational Speaker 2019	21.5	22.4	5.8	43.4
DUDA 2019	23.9	21.9	4.6	44.3
L2C 2021	31.3		15.1	45.3
L2C(+CUB) 2021	31.8	-	16.3	45.6
PLC 2022	31.0	23.4	25.3	49.1
Ours	**33.4**	**25.7**	**28.1**	**52.2**

Table 1: Comparison with the state of the arts on Birds-to-Words.

Model	BLEU-4	METEOR	CIDEr	ROUGE-L
DUDA 2019	47.3	33.9	112.0	-
VAM+ 2020	51.3	37.8	115.8	70.4
VARD-T 2023a	55.4	40.1	126.4	73.8
DIRL+CCR 2024	54.6	38.1	123.6	71.9
NCT 2023b	55.1	40.2	124.1	73.8
SCORER+CBR 2023c	56.3	41.2	126.8	74.5
PLC 2022	51.2	36.2	128.9	71.7
Ours	**56.4**	**41.6**	**129.0**	**74.7**
CLIP4IDC 2022	56.9	38.4	150.7	76.4
VIR-VLFM 2023	58.2	42.6	153.4	78.9
Ours(w/ CLIP)	**58.9**	**42.7**	**153.6**	**79.1**

Table 2: Comparison with the state of the arts on CLEVR-Change.

4.3 Comparisons with State-of-the-art Methods

Results on Birds-to-Words Dataset

Birds-to-Words dataset contains a significant amount of fine-grained changes, which can effectively demonstrate the effectiveness of our method in fine-grained changes. To validate the superiority of our proposed method F&P, we compare our method with other well-known change captioning methods. It can be observed that our F&P achieves the best performance.

As shown in Table 1, methods in the upper part of the table are not pre-trained specifically for Birds-to-Words Dataset,

Model	BLEU-4	METEOR	CIDEr	ROUGE-L
DUDA 2019	8.1	11.8	32.5	29.1
VAM+ 2020	11.1	12.9	42.5	33.2
DUDA+Aux 2021	8.1	12.5	34.5	29.9
CLIP4IDC 2022	11.6	14.2	47.4	35.0
DIRL+CCR 2024	10.3	13.8	40.9	32.8
VIR-VLFM 2023	12.2	15.3	48.9	36.2
SCORER+CBR 2023c	10.2	12.2	38.9	-
Ours	**12.4**	**15.3**	**49.1**	**36.7**

Table 3: Comparison with the state of the arts on Spot-the-Diff.

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

Model	BLEU-4	METEOR	CIDEr	ROUGE-L
Relational Speaker 2019	6.7	12.8	26.4	37.5
DUDA 2019	6.5	12.4	22.8	37.3
BiDiff 2022	6.9	14.6	27.7	38.5
SCORER+CBR 2023c	10.0	15.0	33.4	39.6
NCT 2023b	8.1	15.0	34.2	38.8
DIRL+CCR 2024	10.9	15.0	41.0	34.1
VIXEN-C 2024	8.6	15.4	38.1	42.5
VARD-T 2023a	10.0	14.8	35.7	39.0
Ours	**11.0**	**15.4**	**48.4**	**43.1**

Table 4: Comparison with the state of the arts on Image-Editing-Request.

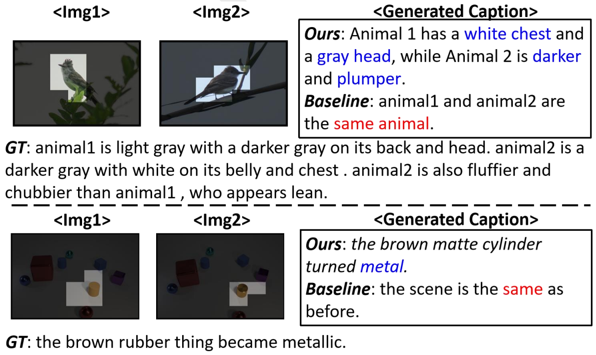
while the lower part of the table introduces the birds pre-training methods. Compared to the previous class of methods, it can be seen that our method is 1.6 points higher on BLEU-4 and 6.6 points higher on ROUGE-L than the L2C(+CUB) [Yan *et al.*, 2021] method and 3.1 points higher on CIDEr than the Natural Naturalist method. Compared with the latter method, our method is 2.4, 2.3, 2.8, and 3.1 points higher than PLC [Yao *et al.*, 2022] method on BLEU-4, METEOR, CIDEr, and ROUGE-L. This is because compared to the latter, our F&P method focuses on fine-grained features and selects discriminative regions from them, truly treating the change captioning task as a "finding the changed regions", "perceiving semantic change" and "generating descriptions" pattern. Compared to other novel methods, we additionally use a better finding and perceiving method and introduce the pre-training understanding tasks. Overall, our method achieves new state-of-the-art performance on the Birds-to-Words dataset.

Results on CLEVR-Change Dataset

The CLEVR-Change dataset is a widely used benchmark for image change captioning. We compare our F&P model with state-of-the-art encoder-decoder and vision-language pre-training methods, as shown in Table 2. The upper section reports results with traditional inputs, while the lower section shows results with vision-language pre-training. Compared to SCORER+CBR [Tu *et al.*, 2023c], F&P achieves gains of 0.1 BLEU-4, 0.4 METEOR, 2.2 CIDEr, and 0.2 ROUGE-L. Although the improvement is less pronounced than on Birds-to-Words, this is due to CLEVR-Change focusing on subtle changes (e.g., color, size), which limits the benefit of our fine-grained feature extraction. Nonetheless, F&P outperforms prior methods on most metrics.

Generality Results on Other Two Datasets

To verify the generality of our method, we evaluate F&P on Spot-the-Diff [Jhamtani and Berg-Kirkpatrick, 2018] and Image-Editing-Request [Tan *et al.*, 2019], where changes are less fine-grained. We replace the FGFL module with CLIP-based features [Guo *et al.*, 2022], retaining the DE module to form a general FGSCP. As shown in Table 3 and Table 4, F&P outperforms VIR-VLFM [Lu *et al.*, 2023] by 0.2 BLEU-4, 0.2 CIDEr, and 0.5 ROUGE-L on Spot-the-Diff, and surpasses VARD-T [Tu *et al.*, 2023a] by 0.2 BLEU-4, 0.6 METEOR, 0.7 CIDEr, and 1.1 ROUGE-L on Image-Editing-Request. It is worth noting that most of the changes in these two datasets are significant content changes, with few fine-grained changes that need to be clearly pointed out. The results demonstrate that our method still performs well on datasets with few fine-grained changes.



Ours: Animal 1 has a white chest and a gray head, while Animal 2 is darker and plumper.
Baseline: animal1 and animal2 are the same animal.

GT: animal1 is light gray with a darker gray on its back and head. animal2 is a darker gray with white on its belly and chest . animal2 is also fluffier and chubbier than animal1 , who appears lean.

Ours: *the brown matte cylinder turned metal.*
Baseline: the scene is the same as before.

GT: the brown rubber thing became metallic.

Figure 3: Examples of generated cases from Birds-to-Words(first block) and CLEVR-Change(second block) by F&P and baseline. The wrong phrases and fine-grained information are highlighted in red and blue respectively.

4.4 Ablation Study

To verify the effectiveness of our F&P, we conduct ablation studies on Birds-to-Words [Forbes *et al.*, 2019] and CLEVR-Change [Park *et al.*, 2019]. The results are shown in Table 5.

Fine-Grained Semantic Change Perception

We divide FGSCP into two parts: Fine-Grained Feature Learning (FGFL) and Difference Enhancement (DE) to evaluate its performance. In order to evaluate the effectiveness of FGFL, we set up a control group to extract image features only with Transformer blocks with the same layers number, instead of initialising FGFL with the pre-training of fine-grained classification. As shown in Table 5, when we perform FGFL, it can promote the improvements 1.3/0.7 of BLEU-4, 1.3/0.6 of METEOR, 1.6/0.7 of CIDEr, 0.9/0.8 of ROUGE-L, on Birds-to-Words/CLEVR-Change. It can be seen that the FGFL shows better performance on the fine-grained Birds-to-Words dataset, which shows that the FGFL effectively assists the network in the extraction of fine-grained words. More visual details are expanded in Section 4.5.

For DE, our F&P can promote the improvements 0.5/1.8 of BLEU-4, 0.4/1.0 of METEOR, 1.1/1.7 of CIDEr, 0.4/2.2 of ROUGE-L, on Birds-to-Words/CLEVR-Change. The experiment shows that DE shows an effective improvement in both two datasets while the performance on the CLEVR-Change dataset is better. The reason is that the content of paired images of CLEVR-Change dataset changes less, and most of them are limited to local changes, so the DE step for calculating regional similarity brings greater improvement. When it comes to Birds-to-Words dataset, the difference between paired images is too large, and it is difficult to judge the change only by the local content similarity. This also confirms our previous statement.

Weakly-Supervised Discriminator

In order to evaluate the effectiveness of WSD, we set the control group not to perform the weakly-supervised training step, and only use one-stage training to directly generate the final caption. The results are shown in Table 5. When we execute

Model	Birds-to-Words				CLEVR-Change			
	BLEU-4	METEOR	CIDEr	ROUGE-L	BLEU-4	METEOR	CIDEr	ROUGE-L
Baseline	30.4	23.1	25.3	49.5	48.5	34.7	125.8	69.3
+FGFL	31.2	23.4	25.8	51.3	50.0	37.6	126.2	70.6
+DE	31.2	23.7	25.4	50.9	49.7	40.2	126.8	70.9
+WSD	32.0	24.1	25.9	50.7	53.5	40.6	127.1	71.4
+FGFL+DE	31.7	23.9	26.3	51.1	55.7	41.1	128.2	73.8
+FGFL+WSD	32.9	25.3	27.0	51.8	54.6	40.6	127.3	72.5
+DE+WSD	32.1	24.4	26.5	51.3	55.7	41.0	128.3	73.9
+FGFL+DE+WSD (F&P)	**33.4**	**25.7**	**28.1**	**52.2**	**56.4**	**41.6**	**129.0**	**74.7**

Table 5: Ablation studies on Birds-to-Words and CLEVR-Change. FGFL, DE, WSD are short for Fine-Grained Feature Learning, Difference Enhancement and Weakly-Supervised Discriminator.

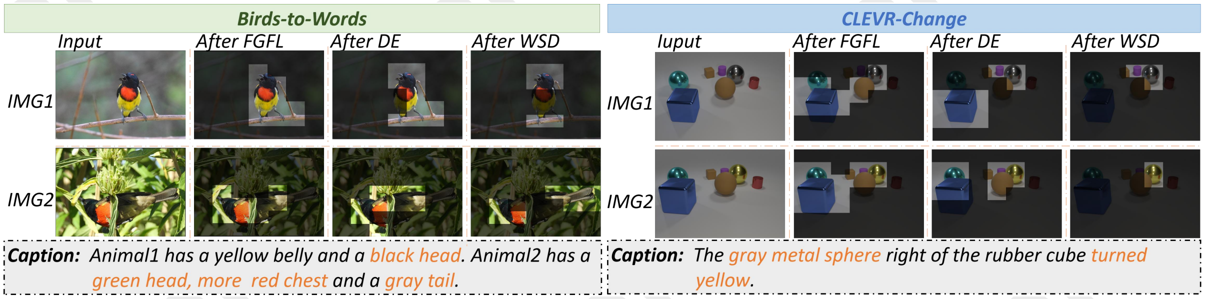

Birds-to-Words
Input
After FGFL
After DE
After WSD
IMG1
IMG2
Caption: Animal1 has a yellow belly and a black head. Animal2 has a green head, more red chest and a gray tail.
CLEVR-Change
Iuput
After FGFL
After DE
After WSD
IMG1
IMG2
Caption: The gray metal sphere right of the rubber cube turned yellow.

Figure 4: Visualisation of the process of generating captions from Birds-to-Words and CLEVR-Change. The fine-grain information and the discriminative parts are highlighted in orange, respectively.

WSD training, it can promote the improvements 1.7/0.7 of BLEU-4, 1.8/0.5 of METEOR, 1.8/0.8 of CIDEr, 1.1/0.9 of ROUGE-L, on Birds-to-Words/CLEVR-Change, compared with our FGFL+DE method. Similarly, it can be seen that our WSD shows better performance on more complex Birds-to-Words dataset, which indicates that our WSD can effectively locate discriminative regions among many features to guide the generation of captions.

4.5 Qualitative Results and Visualisation

To intuitively evaluate our method, we conduct the following qualitative analysis and visualisation. We visualise the generation results and the discriminative regions of our method F&P and baseline. After that, we visualise the focus areas of the three modules in our F&P.

Figure 3 shows qualitative results generated by our F&P and the baseline method. We visualise the attention weights of the input regions for the generated important words. As shown in Figure 3, our proposed method F&P is able to generate more accurate and descriptive captions. It is worth noting that our method can generate better results for texture changes on the CLEVR-Change dataset. It is because texture changes are closely related to fine-grained features, and our method is sensitive to this kind of change.

Finally, to better understand the effectiveness of our find and perceive step, we visually generate the attention weights over tokens of our FGFL, DE, and WSD step as shown in Figure 4. Our FGFL perceives the fine-grained features of each paired image, DE initially perceives the differences, and WSD ultimately locates the changed regions that need to be described. For the Birds-to-Words dataset, it is worth noting that during the perception process, some background areas that are different but unrelated to the task will be noticed (such as the plants in Figure 4). These background regions may be related to the living habits of this type of bird, thus helping the perception process. but they are not related to our task, and our WSD can find the changed regions that need to be described in these perceived discriminative regions. More visualisation results can be found in the appendix.

5 Conclusions

In this paper, we propose a novel change captioning learning paradigm, Find and Perceive (F&P) to pinpoint and understand the fine-grained change. Inspired by tell change pattern of human-being, our F&P consists of two main components, Fine-Grained Semantic Change Perception (FGSCP) for perceiving fine-grained change semantics and content and the Weakly-Supervised Discriminator (WSD) of changed regions for localising the important regions. For the FGSCP, we treat it as a two-step manner, firstly introducing the fine-grained categorization and then enhancing the interaction of the two paired images. For the WSD, we achieve the localization of changed regions with a weakly-supervised approach, which utilizes the response of the generation captions to the changed regions to guide our F&P. Finally, we conduct extensive experiments on two commonly used datasets Birds-to-Words and CLEVR-Change to verify the effectiveness of F&P, and additional experiments on two datasets Spot-the-Diff and Image-Editing-Request to verify the generality. Experimental results show that our proposed method outperforms existing change captioning methods and achieves new state-of-the-art performance.

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China Under Grants No. 62176253.

References

[Addepalli *et al.*, 2024] Sravanti Addepalli, Ashish Ramayee Asokan, Lakshay Sharma, and R Venkatesh Babu. Leveraging vision-language models for improving domain generalization in image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23922–23932, 2024.

[Banerjee and Lavie, 2005] Satanjeev Banerjee and Alon Lavie. 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, 2005.

[Black *et al.*, 2024] Alexander Black, Jing Shi, Yifei Fan, Tu Bui, and John Collomosse. Vixen: Visual text comparison network for image difference captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 846–854, 2024.

[Chen *et al.*, 2020] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020.

[Dessì *et al.*, 2023] Roberto Dessì, Michele Bevilacqua, Eleonora Gualdoni, Nathanaël Carraz Rakotonirina, Francesca Franzon, and Marco Baroni. Cross-domain image captioning with discriminative finetuning. *ArXiv*, abs/2304.01662, 2023.

[Driess *et al.*, 2023] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.

[Farhadi *et al.*, 2010] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a story: Generating sentences from images. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV 11*, pages 15–29. Springer, 2010.

[Forbes *et al.*, 2019] Maxwell Forbes, Christine Kaeser-Chen, Piyush Sharma, and Serge Belongie. Neural naturalist: generating fine-grained image comparisons. *arXiv preprint arXiv:1909.04101*, 2019.

[Guo *et al.*, 2022] Zixin Guo, Tzu-Jui Julius Wang, and Jorma Laaksonen. Clip4idc: Clip for image difference captioning. *arXiv preprint arXiv:2206.00629*, 2022.

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

[He *et al.*, 2019] Sen He, Hamed R Tavakoli, Ali Borji, and Nicolas Pugeault. Human attention in image captioning: Dataset and analysis. In *ICCV*, 2019.

[Hirota *et al.*, 2023] Yusuke Hirota, Yuta Nakashima, and Noa García. Model-agnostic gender debiased image captioning. *ArXiv*, abs/2304.03693, 2023.

[Hosseinzadeh and Wang, 2021] Mehrdad Hosseinzadeh and Yang Wang. 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, 2021.

[Hu *et al.*, 2022] Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. Scaling up vision-language pre-training for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17980–17989, 2022.

[Jhamtani and Berg-Kirkpatrick, 2018] Harsh Jhamtani and Taylor Berg-Kirkpatrick. Learning to describe differences between pairs of similar images. *arXiv preprint arXiv:1808.10584*, 2018.

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

[Li *et al.*, 2020a] Linjie Li, Zhe Gan, and Jingjing Liu. A closer look at the robustness of vision-and-language pre-trained models. *arXiv preprint arXiv:2012.08673*, 2020.

[Li *et al.*, 2020b] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. *arXiv preprint arXiv:2012.15409*, 2020.

[Li *et al.*, 2020c] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 121–137. Springer, 2020.

[Lin and Hovy, 2003] Chin-Yew Lin and Eduard Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 human language technology conference of the North American chapter of the association for computational linguistics*, pages 150–157, 2003.

[Lu *et al.*, 2019] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

[Lu *et al.*, 2023] Xiaonan Lu, Jianlong Yuan, Ruigang Niu, Yuan Hu, and Fan Wang. Viewpoint integration and registration with vision language foundation model for image

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

change understanding. *arXiv preprint arXiv:2309.08585*, 2023.

[Mikolov *et al.*, 2013] Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations*, 2013.

[Papineni *et al.*, 2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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, 2002.

[Park *et al.*, 2019] Dong Huk Park, Trevor Darrell, and Anna Rohrbach. Robust change captioning. In *ICCV*, 2019.

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

[Shi *et al.*, 2020] Xiangxi Shi, Xu Yang, Jiuxiang Gu, Shafiq Joty, and Jianfei Cai. Finding it at another side: A viewpoint-adapted matching encoder for change captioning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 574–590. Springer, 2020.

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

[Tan and Bansal, 2019] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

[Tan *et al.*, 2019] Hao Tan, Franck Dernoncourt, Zhe Lin, Trung Bui, and Mohit Bansal. Expressing visual relationships via language. *arXiv preprint arXiv:1906.07689*, 2019.

[Tu *et al.*, 2023a] Yunbin Tu, L. Li, Li Su, Jun Du, Kelvin Lu, and Qin Huang. Viewpoint-adaptive representation disentanglement network for change captioning. *IEEE Transactions on Image Processing*, 32:2620–2635, 2023.

[Tu *et al.*, 2023b] Yunbin Tu, Liang Li, Li Su, Kelvin Lu, and Qin Huang. Neighborhood contrastive transformer for change captioning. *ArXiv*, abs/2303.03171, 2023.

[Tu *et al.*, 2023c] Yunbin Tu, Liang Li, Li Su, Zheng-Jun Zha, Chenggang Yan, and Qingming Huang. Self-supervised cross-view representation reconstruction for change captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2805–2815, 2023.

[Tu *et al.*, 2024] Yunbin Tu, Liang Li, Li Su, Chenggang Yan, and Qingming Huang. Distractors-immune representation learning with cross-modal contrastive regularization for change captioning. *arXiv preprint arXiv:2407.11683*, 2024.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[Vedantam *et al.*, 2015] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4566–4575, 2015.

[Vinyals *et al.*, 2015] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

[Yan *et al.*, 2021] An Yan, Xin Eric Wang, Tsu-Jui Fu, and William Yang Wang. L2c: Describing visual differences needs semantic understanding of individuals. *arXiv preprint arXiv:2102.01860*, 2021.

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

[Zeng *et al.*, 2023] Zequn Zeng, Hao Zhang, Zhengjue Wang, Ruiying Lu, Dongsheng Wang, and Bo Chen. Conzic: Controllable zero-shot image captioning by sampling-based polishing. *ArXiv*, abs/2303.02437, 2023.