Towards Few-shot Image Captioning with Cycle-based Compositional Semantic Enhancement Framework

1st Peng Zhang*
Department of Computer Science
Durham University
Durham, UK
peng.zhang@durham.ac.uk

2nd Yang Bai*
Department of Computer Science
Durham University
Durham, UK
yang.bai@durham.ac.uk

3rd Jie Su
Department of Computer Science
Newcasle University
Newcasle, UK
J.Su4@newcastle.ac.uk

4th Yan Huang
Institute of Automation
Chinese Academy of Sciences
Beijing, China
yhuang@nlpr.ia.ac.cn

5th Yang Long†
Department of Computer Science
Durham University
Durham, UK
Yang.long@ieee.org

2023 International Joint Conference on Neural Networks (IJCNN) | 978-1-6654-8867-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/IJCNN54540.2023.10191558

***Abstract*—Many efforts paid attention to the multi-modal task, of which image captioning is a classic work. Especially the Clip model improves the performance of image captioning; meantime, its few-shot and zero-shot problems have become a significant research project. In this work, aiming at the image captioning task, we design the new few-shot and zero-shot settings different from popular directions. The direction focuses on the impact of the exited dataset for captioning model ability. According to analysis, we discover the frequency of the word combination can directly influence the performance of the captioning model. Based on this, we define the new few-shot and zero-shot settings. In terms of this, a Cycle-based captioning framework based on data augmentation is proposed to overcome this problem, of which the novelty switcher module is the critical component. Finally, experiments demonstrate that our framework can achieve state-of-the-art performance on both traditional, few-shot and zero-shot settings.**

***Index Terms*—Image captioning, cycle-based, switcher module**

I. Introduction

As a traditional task in deep learning, image captioning aims to describe an image in natural language. Therefore, it generates a sequence of words by designing a model to reflect the relationship between visual and textual information. Recently, many efforts paid attention to tackle this task by applying Recurrent Neural Network models, Graph Neural Networks and Transformer. Especially the Transformer model can extract more key knowledge between an image and caption to achieve state-of-the-art performance. As a significant revolution in deep learning, few-shot and zero-shot learning provide more inspiration. Few-shot learning aims to predict the correct classes when only a few samples are available in the training dataset. Zero-shot learning aims to predict the correct classes which are not observed in the training dataset. [1]. For example, the zero-shot capability was demonstrated in computer vision [1]. Besides, the seminal CLIP [2] image-text transformer model can execute tens of downstream tasks without further training. Impressively, the DALL-E [3] can generate images in terms of unseen descriptions. Although existing algorithms can generate good descriptions on the traditional testing set, the image captioning task needs to be more attentive to the few-shot and zero-shot settings.

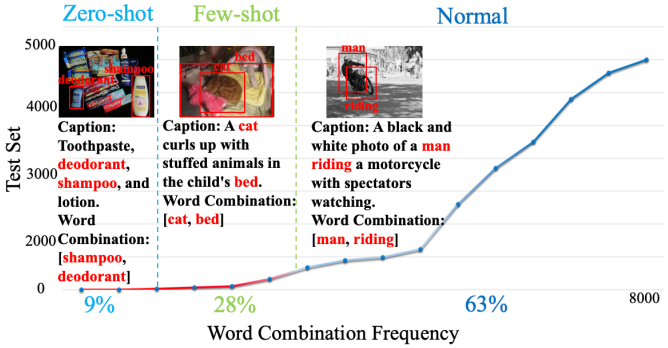

Fig. 1. The normal, few-shot and zero-shot settings on the Test Set based on Word Combination Frequency.

In the visual-textual multi-modal task, the relationship between images and texts is the most important factor. Therefore, several efforts have analyzed the impacts of the word on the model performance. For instance, Yan [4] demonstrated that word frequency affects image-text matching model performance. However, compared with a single word, word combinations are more critical to the meaning and understanding of a sentence. Different word combinations include amounts

* Equal contribution. †Corresponding author.

of semantic information, which means that the frequency of word combinations has more effects on model performance compared with the frequency of a single word. Figure 1 describes the proportions of the normal, few-shot and zero-shot based on word combination frequency on the test set. The few-shot and zero-shot settings have fewer proportions than the normal setting, which are 9%, 28% and 63%. We do experiments to analyze the impact of word combination on the image captioning task in the methodology section to prove our hypothesis. Furthermore, different from traditional few-shot and zero-shot settings, we propose a new direction of few-shot and zero-shot settings based on this hypothesis in the image captioning task.

Generally, few-shot and zero-shot methods increase the model generalization to improve the model performance in the kinds of tasks, and data augmentation is the most straightforward direction. In terms of this, we propose a novel Cycle Captioning Framework to improve the model ability on the traditional setting and the new few-shot and zero-shot settings for the image captioning task. In the framework, the proposed Image Generator generates the image with feature-level as new training data to feed into the Caption model; meanwhile, the proposed Word Switcher reasonably exchange words of the caption to augment the training data. Summary the contributions:

- According to the analysis impact of word combination on the image captioning task, the new few-shot and zero-shot settings are proposed in this work. While improving the performance of new settings promotes the extension of a new direction on the image captioning task.
- The proposed Cycle Captioning Framework adequately apply the existing data to improve the model generalization ability on the image captioning task. At the same time, we design a novel Word Switcher to augment the training data.
- The experiments demonstrate that the Cycle Captioning Framework with Word Switcher achieves state-of-the-art performance.

The methodology section describes the details of word combination, Cycle Captioning Framework and Word Switcher. The experiment and ablation study sections analyse the ability of the proposed algorithm on the image captioning task.

II. Related Work

As a traditional task, many efforts were applied to image captioning. Specifically, [5] first proposed the deep learning algorithm to predict the sequence of captions in image captioning. With the development of machine learning techniques, more and more works extracted the semantics relationship based on Attention and Graph neural networks [6]–[8]. Subsequently, the image could be extracted more details through a transformer with self-attention to improve the model performance [9], [10]. On the one hand, plenty of efforts solved the text problems based on improvements to the language model such as LSTMs, Transformer, and CNNs [11], [12]. On the other hand, the generated language of image grounding and non-vision words obtained a better performance by combination with different semantic information [13], [14]. Specifically, as a popular language framework, Transformer has been widely used in image captioning tasks. The CogView constructed a 4-billion-parameter Transformer to achieve a SOTA captioning performance [15].

Despite the captioning model experiencing an improvement, also trained the large-scale vision-language data sets can improve the generating captioning performance. Thus, some image captioning tasks applied the large-scale vision-language data sets in recent years, such as the Visual Genome and MS-COCO. The captioning model utilized millions of image and text pairs from the web to improve the generated language performance [16] [17]. Based on this technique, some methods applied the unsupervised external data through conditioning the model during the training to focus on describing novel objects [18] [19]. The model can execute external object information in the pre-training and inference phases [20]. The model can join an image-language embedding space and the visual detector for the unsupervised methods [21] [22]. Following this direction, the zero-shot language model CLIP was proposed, which acquired a better score in the image captioning task based on 400M image-sentence pairs from the web [2]. Based on powerful CLIP, text-driven image manipulation with GANs and other generative models can be supported by means of CLIP [23] [24]. Unlike existing image captioning few-shot and zero-shot learning directions, we propose new few-shot and zero-shot settings in image captioning. Our framework can augment the image-captions pairs based on an existing data set.

III. Methodology

A. Few-shot and zero-shot settings

Many machine learning tasks apply the MS-COCO dataset as a traditional dataset, such as object detection, text-image generation and text-image matching. The MS-COCO also is the most popular dataset in the image captioning task. Therefore, we analysis the MS-COCO dataset to define our few and zero-shot settings. As a key part, we define the word combination that two objects or two nouns of a caption construct a word combination; for example, in the caption "A woman is drinking water.", we define the '[woman, water]' to be a word combination. Each image includes five captions in the MS-COCO dataset, and we collect about 44,712,680 word combinations.

Based on the word combination, we define the few-shot and zero-shot settings. Firstly, we count the frequency of all word combinations, including the training set and testing set and sort them based on their frequency. Then, applying the SOTA models evaluate the data of high-frequency and low-frequency word combinations, respectively. In this evaluation, the CIDEr and BLEU-4 scores reflect the impacts of high-frequency and low-frequency word combinations on the performance of SOTA models, which is described by Figure 2. It shows that the CIDEr and BLEU-4 scores decline with the decrease of the

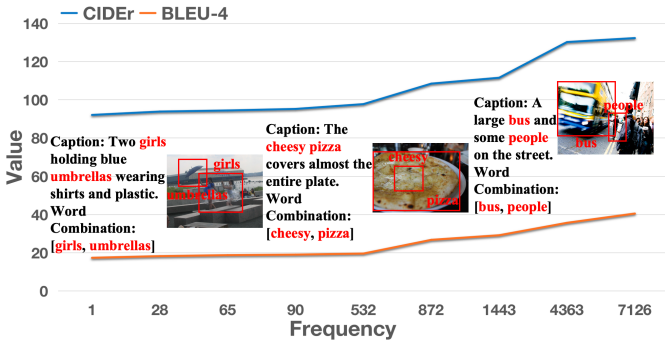

Fig. 2. The developments of CIDEr and BLEU-4 with frequency of word combination

frequency of the word combination, which demonstrates that SOTA models have terrible performance on the data of low-frequency word combinations compared with the data of high-frequency word combinations. Finally, we define the zero-shot test set and the few-shot test set, respectively. The data of low-frequency word combinations of the test set that do not appear in the training set indicates the zero-shot test set. The few-shot test set is the data of low-frequency word combinations of the test set whose appearing frequency in the training set is less than or equal to K.

B. Cycle Captioning Framework

The definition of the few-shot and zero-shot settings show that the amount of data can directly affect the performance of the captioning model. We propose a cycle captioning framework that augments data diversity to overcome the problems in the few-shot and zero-shot settings. Unlike other state-of-the-art captioning models, our framework includes a feature-level image generator and word switcher module in addition to the captioning model. The interaction of the latter two modules enhances the data and thus improves the performance of the captioning model. The details of our framework are described later in the process of cycle captioning framework section.

1) Process of Cycle Caption Framework: Given an image feature $\mathcal{X}$ extracted from an image as input, it is sequentially fed into the Captioning model $\mathcal{G}_c(.)$ and the Feature-Level Image Generator $\mathcal{G}_i(.)$ to generate a sequence of vectors $\widetilde{\mathcal{Y}}$ as a caption:

$$\widetilde{\mathcal{Y}} = \mathcal{G}_c(\mathcal{G}_i(\mathcal{G}_c(\mathcal{X}), \mathcal{X})), \tag{1}$$

In our framework, two types of features are extracted from images $\mathcal{X}$: image feature map X and region features X_R. The sequence of captions $\mathcal{Y}$ is the same as image features, also described by two representations: original captions Y_r and exchanged captions Y_{ex}.

The cycle process of the whole framework is divided into two parts. In the first part, the real caption embeddings Y_r, real image feature map X_r and region features X_R in the training set enter the caption model and feature image generator to obtain the generated caption embeddings $\widetilde{Y}_r$ and image feature map $\widetilde{X}_r$, respectively.

$$\begin{aligned} \widetilde{Y}_r &= \mathcal{G}_c(X_r) \\ \widetilde{X}_r &= \mathcal{G}_i(Y_r, X_R), \end{aligned} \tag{2}$$

Then, the generated caption embeddings $\widetilde{Y}_r$ and image feature map $\widetilde{X}_r$ as new training data are fed into two models to acquire cycle caption embeddings $\widetilde{Y}_f$ and cycle image feature map $\widetilde{X}_f$.

$$\begin{aligned} \widetilde{X}_f &= \mathcal{G}_i(\widetilde{Y}_r, X_R) \\ \widetilde{Y}_f &= \mathcal{G}_c(\widetilde{X}_r), \end{aligned} \tag{3}$$

Through the above steps, we realized the first step of data expansion without changing the training data so that both models could obtain more data for training.

In the second part, the caption embeddings and region features X_R in the training set first input into the word switcher $\mathcal{S}(.)$ to obtain new exchanged caption embeddings Y_{ex} and exchanged region features X_R^{ex}:

$$Y_{ex}, X_R^{ex} = \mathcal{S}(Y_r, X_R), \tag{4}$$

Then these new training data are fed into the image generator $\mathcal{G}_i(.)$ and caption model $\mathcal{G}_c(.)$ to generate the predicted exchanged captions $\widetilde{Y}_{ex}$:

$$\widetilde{Y}_{ex} = \mathcal{G}_c(\mathcal{G}_i(Y_{ex}, X_R^{ex}), Y_{ex}), \tag{5}$$

2) Captioning Model: Our caption model, inspired by Mesh-Memory Transformer [13], is represented by $\mathcal{G}_c(.)$. It is the encoder and decoder structure with stacks of self-attention layers. The encoder module extracts the relationships from the input image, and then the decoder module receives the output of the encoder module to predict each word of a caption. All connections between the image and caption are executed by dot-product attention. The attention operator follows the standard sets of the transformer, namely a set of queries Q, keys K and values V, and according to the weighted sum of value vectors with aggregation between query and key vectors. The operator is shown as:

$$Attention(Q, K, V) = softmax(QK^T/d)V, \tag{6}$$

where Q is a matrix of n_q query vectors, K and V both contain n_k keys and values, all with the same dimensionality, and d is a scaling factor. The encoder layers include the self-attention and position-wise feed-forward with a residual connection and a layer norm $Addnorm$, and then stacks of them define our encoder module:

$$O_{ce} = Addnorm(\mathcal{F}(Attention(W_q X_r, W_k XR, W_v X_r))), \tag{7}$$

where W_q, W_k, W_v indicate the matrices of learnable weights and $\mathcal{F}(.)$ is position-wise feed-forward layer. The O_{ce} represents the output of encoder module.

Then, the decoder collects outputs from the encoder module and the self-attention mask module S_{mask} to obtain a generated caption $\widetilde{\mathcal{Y}}$, which is described by:

$$\widetilde{\mathcal{Y}} = Addnorm(\mathcal{F}(Attention(O_{ce}, S_{mask}(\mathcal{Y})))), \tag{8}$$

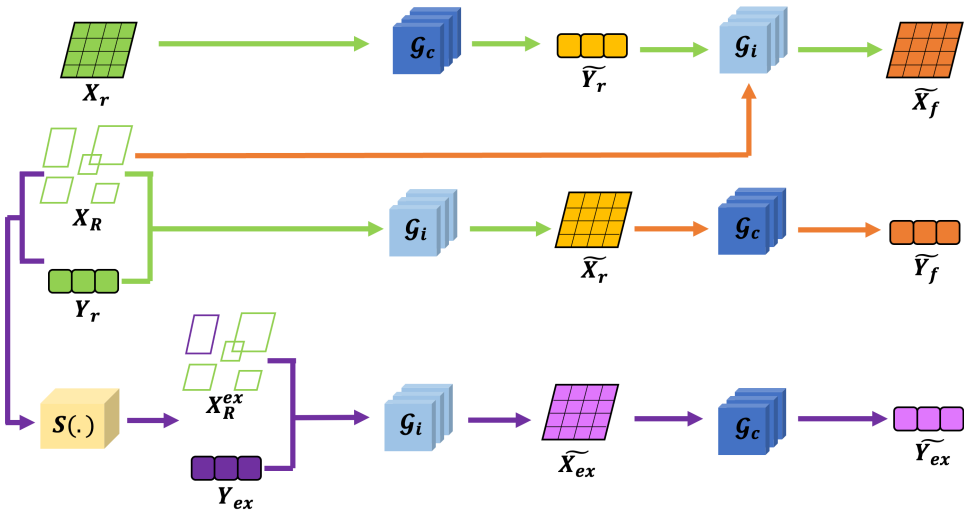

Fig. 3. The structure of Cycle Captioning framework. The green line is the training process using training data and the orange line indicates the training process using predicted data. The purple line represents the switch module and the training process using exchanged data.

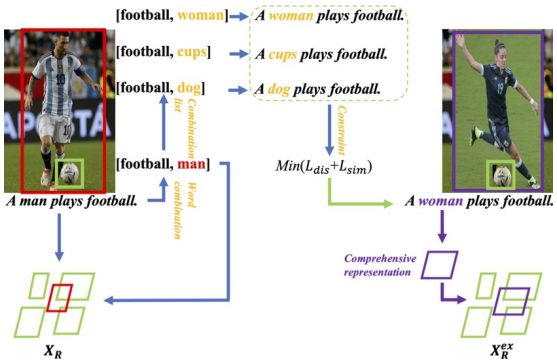

Fig. 4. The details of the switcher module. The red word is the exchanged word and purple is the new word.

3) Feature-Level Image Generator: The image generator synthesis an entire image in the most multi-modal task, which is hard to optimize. However, the proposed feature-level image generator only generates the image features from captions. The $\mathcal{G}_i(.)$ represents the feature-level image generator (FL image-G) in this work, whose structure is similar to the captioning model based on the transformer. The main difference is that captions $\mathcal{Y}$ combined with the image region feature X_R are the inputs to generate the image feature map $\widetilde{X}_r$.

In this generator, the image region feature provides extra information to improve the accuracy of the synthesised image feature map. The switcher module executes the image region feature to generate the new exchanged image region feature as weak ground truth to train the FL image-G. The reason is that the switcher module can create the exchanged captions but cannot generate the exchanged image feature map, which means that when we apply the exchanged captions to train the FL image-G, there is no ground truth of the image feature map to supervise. But we can directly exchange the region proposal feature corresponding to the exchanged object word to obtain weak ground truth.

4) Switcher Module: The main novel part in our framework, the word switcher module plays a key role. It is represented by $\mathcal{S}(.)$. We follow a principle every time we change words: we only exchange one noun in a caption. However, the exchanged word is not random because some new captions constructed by newly exchanged words are not reasonable, which means that these data can affect our model performance. Therefore, we follow two steps to select the exchanged word. The first step is choosing the newly exchanged word from our word combination. For example, in the caption "A man plays football.", the word combination is '[football, man]'. We first decide to exchange the word man, and we will select the newly exchanged word from the combination list containing the word 'football', such as '[football, woman]', '[football, cups]' and '[football, dog]' etc. Then, these newly exchanged words construct different new captions: "A woman plays football.", "A cups plays football." and "A dog plays football.". The second step is to compute the similarity and distance between these new sentences and original sentence to select the final exchanged sentence:

$$L_{dis} = ||Y_o - \dot{Y}_{ex}||_2,$$
$$L_{sim} = \frac{Y_o \cdot \dot{Y}_{ex}}{||Y_o||||\dot{Y}_{ex}||}, \tag{9}$$

where Y_o and $\dot{Y}_{ex}$ denote the original sentence and the new sentence candidates, L_{dis} and L_{sim} are Euclidean distance and Cosine similarity.

We can obtain the weak exchanged image region feature when we acquire the final newly exchanged caption. We collect the representations of each objects in the whole dataset. Then, we obtain the comprehensive representations through the Equation 10:

$$r_c = (\sum_{n=0}^{N} r_n)/N, \tag{10}$$

TABLE I
THE COMPARISON WITH SOTA ON NEW FEW-SHOT AND ZERO-SHOT SETTING. B@1, B@4, M, R AND C INDICATE BLEU-1 [25], BLEU-4 [25], METEOR [26], ROUGE [27] AND CIDER [28].

Method	***Few-shot Setting***					***Zeo-shot Setting***				
	B@1	B@4	M	R	C	B@1	B@4	M	R	C
Clip-VL [29]	72.80	19.30	27.38	54.13	97.11	72.27	17.06	26.84	55.30	92.75
Ours	***74.41***	***24.26***	***27.76***	***58.02***	***113***	***73.30***	***22.88***	***29.43***	***58.54***	***110.68***

TABLE II
THE COMPARISON WITH SOTA ON TRADITIONAL SETTING OF MS-COCO.

Method	***Metrics***				
	B@1	B@4	M	R	C
SCST [30]	-	34.2	26.7	55.7	114
Up-Down [31]	79.8	36.3	27.7	56.9	120.1
RFNet [32]	79.1	36.5	27.7	57.3	121.9
GCN-LSTM [33]	80.5	38.2	28.5	58.3	127.6
ORT [34]	80.5	38.6	28.7	58.4	128.3
AoANet [35]	80.2	38.9	29.2	58.8	129.3
M^2 Transformer [13]	80.8	39.1	29.2	58.6	131.2
Clip-VL [29]	-	40.2	31.1	-	134.2
Ours	80.8	***40.6***	***31.6***	***59.3***	***134.6***

where r_c and r_n are the comprehensive representation and each representation of object. For example, if the exchanged word is 'man' and the new word is 'woman', we can acquire their representations from the image region features by class probability. We directly apply the comprehensive representation of 'woman' to replace 'man', which obtains the new exchanged image region feature X_R^{ex} based on this principle. Figure 4 describes details.

IV. EXPERIMENTS

In this section, two evaluation settings demonstrate our model performance. First, our model and SOTA models are evaluated in a traditional setting. Then, our few-shot and zero-shot setting, as the second setting, evaluate our model and SOTA models.

A. Datasets

The MS-COCO is applied to evaluate our model performance. The dataset includes more than 120000 images, and 5 different captions annotate each image. Most image captioning tasks widely follow Karpathy's split setup [36], where 110000 images are applied for training, 5000 for validation and the rest for testing.

Regarding the Methodology section, the zero-shot and few-shot settings are set up based on our word combination principle. Hence, the zero-shot setting splits conventional images for training and validation. We select partial images from the standard test setting for zero-shot testing based on our zero-shot principle. The training and validation set of our few-shot setting also follows the common training and validation setting, and we set K-shot ($K = 2$) to choose testing images.

B. Experiments settings

We follow the standard evaluation protocol to apply the typical image captioning metrics to show our model performance: BLEU [25], METEOR [26], ROUGE [27] and CIDEr [28].

In terms of our framework, an object in the caption is selected randomly to be exchanged with another different object constructing a new caption and then generating a new image feature, which means that the exchanged object of the caption should correspond to the object of images. Hence, we need to obtain image regions in our framework besides the feature map. To acquire image regions, we execute Faster R-CNN [37] with ResNet-101 [38] fine-tuned on the Visual Genome [39] to obtain a 2048-dimensional feature for each region. For caption representation, we linearly project words of one-hot vectors to the input dimensionality of the model d. Then, the positional encoder [10] represents word positions added into the sequence to acquire two embeddings. In our framework, the dimensionality d of each layer is set to 512, the number of memory vectors is 10, and the number of heads is 6. We follow the most common training strategy in image captioning tasks, which is divided into two stages. The first stage is training our captioning model and image generator with a batch size of 256 and learning rate scheduling strategy with a warmup to 100 epochs. Then, two models are optimized with the Adam optimizer, and the beam size is set to 5. The second stage is that the captioning model is fine-tuned with CIDEr-D optimization with a fixed learning rate of 3×10^{-4}.

C. Comparison with state of the art

In this part, a comparison between the performance of several recent SOTA proposals and our image captioning framework in both settings demonstrates that our framework can achieve SOTA performance. The compared models include SCST [30] and Up-down [31], which applied attention to the grid of features and regions, respectively. Then, the RFNet [32] applies a recurrent fusion network to merge CNN features, and GCN-LSTM [33] executes a Graph CNN to obtain pairwise relationships between image regions. Further, our framework compares with AoANet [35], ORT [34] and M^2 Transformer [13], which apply Transformer for encoding image regions. Finally, we compare with the Clip-VL model [29], which uses the pre-trained Clip model to extract the image region feature.

Our framework and aforementioned SOTA models evaluate the traditional test split. Table II reports the comparison performances, applying the caption model and fine-tuning optimization on the CIDEr score. According to observation

TABLE III
ABLATION STUDY FOR EFFECTS OF FEATURE-LEVEL IMAGE GENERATOR AND SWITCHER MODULE ON THE DIFFERENT SETTINGS. FL IMAGE-G REPRESENTS FEATURE IMAGE GENERATOR AND SCM INDICATES THE SINGLE CAPTIONING MODEL. CYCLE MEANS OUR ENTIRE CYCLE-BASED CAPTIONING FRAMEWORK.

Component			***Tradtional Setting***				***Few-shot Setting***				***Zero-shot Setting***			
Method	FL Image-G	Switcher Module	B@4	M	R	C	B@4	M	R	C	B@4	M	R	C
Baseline	×	×	32.96	28.49	57.69	104.22	22.51	27.54	57.52	110.48	22.19	27.48	57.09	98.73
SCM	✓	×	33.62	29.31	58.04	110.63	23.08	27.73	***58.03***	111.50	22.27	27.63	58.01	110.02
Cycle	✓	✓	***34.93***	***29.84***	***58.20***	***114.49***	***24.26***	***27.76***	58.02	***113***	***22.88***	***29.43***	***58.54***	***110.68***

TABLE IV
THE COMPARISON WITH SOTA ON SINGLE CAPTIONING MODEL.

Method	***Metric***				
	BLEU-1	BLEU-4	METEOR	ROUGE	CIDEr
Clip-VL	75.30	33.39	27.69	56.09	111.5
Ours	***75.51***	***34.93***	***29.84***	***58.20***	***114.49***

from Table II, our framework achieves the best performance on BLEU-1, BLEu-4, METEOR, ROUGE and CIDEr. Our framework especially increases the SOTA on ROUGE by 0.7.

Because the Clip-VL is the best performance, we compare the testing results with it on our few-shot setting and zero-shot test setting, which are represented by Table I. In particular, we mainly report the performances of the few-shot setting with $K = 2$. As it can be observed from Table I, the performances of all metrics are worse than the traditional test setting, which also proves that the frequency of the word combination can impact the model performance. However, Table I indicates that our framework surpasses SOTA approach in terms of BLEU-1, BLEU-4, METEOR and ROUGE being the best performer.

To further prove our framework performance, Figure 5 proposes qualitative results and visualization. In all SOTA approaches, the Clip-VL model is the best performer. Hence, our framework compares with it. On average, our framework can generate more accurate and reasonable captions to describe the corresponding images. In addition, our framework also describes more details and object relationships for images.

In addition, we compare the performance of the single captioning model between our framework and the SOTA model, which is shown by Table IV. Because most SOTA image captioning models fine-tune the captioning model with an reinforced strategy to improve performance, but a single captioning model is the most significant part, which directly represents the actual ability of captioning generation for each SOTA approach. Table IV reports the results, showing that our framework is the best performer on all evaluation metrics and reflects our framework's superiority. Although our captioning model includes an image generator, it is a crucial augmentation part of our captioning model, and our framework's total number of layers is fewer than other methods.

V. ABLATION STUDY

The quantitative and Qualitative results evident that our framework achieves the best performance compared with other SOTA models. Furthermore, this ablation study section reports the effects of the feature-level image generator and switcher module on task performance. To directly analyze the effects of each component, the following experiments are executed based on the single captioning model without the reinforced fine-tuning strategy.

A. The Effect of Feature-Level Image Generator

A part of the cycle-captioning model, Table III proves that the feature-level image generator produces essential effects for the whole framework. Compared with the baseline, the entire framework obtains a noticeable improvement in all settings when applying the feature-level image generator. Significantly, the ROUGE increase by approximately 0.51 points compared with the baseline and acquires the best performance on the few-shot setting. Impressively, the CIDEr improves by about 11.29 points in the zero-shot setting.

In our cycle framework, the feature-level image generator based on transformer executes the image region feature to generate the image feature map. Besides, we also applied the traditional GAN to model it without the image region feature. The Table VI reports the differences between two methods. Besides the BLEU-4 of the traditional setting, the results of GAN are worse than the transformer with region feature. However, the performance of GAN is improved compared with the baseline on all settings, which further proves that our cycle framework can obtain an enhancement for the image captioning task. The GAN-based image generator only applies the caption to generate the image feature map without any other data to supervise the model further. But the transformer-based generator supervises the model by using the region feature and the caption.

B. The Effect of Switcher Module

Switcher module is the most novelty component in the whole cycle framework, which can exchange the word of a caption to augment the new training data. Therefore, we try different methods to improve the performance of the framework. All methods are applied based on the cycle framework. Table V, the Transformer based indicates no switcher module, and Random represents that the switcher module randomly exchanges a word in a caption. Both Word nearly and Combination methods follow the word combination. The first one means that we fix the first word of the word combination and exchange its neighbour; for example, in a caption "A man

TABLE V
THE COMPARISONS BETWEEN DIFFERENT SWITCHER METHODS.

Method	*Component*			*Traditional Setting*		*Few-shot Setting*		*Zero-stho Setting*	
	Switch	Constraint	Combination list	B@4	C	B@4	C	B@4	C
Without Switch	×	×	×	33.62	110.63	23.08	110.50	22.27	110.02
Random	✓	×	×	33.41	104.38	22.04	104.01	20.66	100.68
Word nearly	✓	✓	×	32.69	104.51	21.09	104.35	22.13	104.23
Combination	✓	✓	✓	***34.93***	***114.49***	***24.26***	***113***	***22.88***	***110.68***

Fig. 5. The comparison of visualization with SOTA.

plays football.", the word combination is '[man, football]', we will exchange 'man' neighbour 'plays' to other word based on constrain. The second one is that we exchange the second word of the word combination and select a new word from the combination list of the first word; for instance, in a caption "A man plays football.", the word combination is '[man, football]', and we exchange the 'football'. If the combination list may include '[man, tennis]' and '[man, baseball]', we select 'tennis' or 'baseball' to construct a new caption based on the constraint. Table V indicates that the unreasonable switcher method can destroy the ability of the whole framework, such as the Random method and the Word nearly method. These methods generate the new training data as noise to attack the model. Although the Word nearly method executes the word combination and constraint to generate the new weakly reasonable caption, it still exits the instability when exchanging a near word. Finally, we apply the interaction between the combination list and constraint to generate the new caption as much as stable. Table V demonstrates that our switcher method supports the framework to achieve the best performance, especially on the few-shot and zero-shot settings.

TABLE VI
THE COMPARISONS BETWEEN DIFFERENT IMAGE GENERATORS.

	Traditional Setting		*Few-shot Setting*		*Zero-shot Setting*	
Method	B@4	C	B@4	C	B@4	C
Baseline	32.96	104.22	22.51	110.48	22.19	98.73
GAN based	***33.97***	109.74	22.81	110.48	22.11	107.60
Transformer based	33.62	***110.63***	***23.08***	***110.50***	***22.27***	***110.02***

VI. CONCLUSION

In this work, we define the new few-shot and zero-shot settings based on the principle of the word combination. Meanwhile, a cycle-based captioning framework is proposed to solve this task. Firstly, the word combination is designed through the popular dataset. Then, the experiments demonstrate that the word combination frequency can impact the captioning performance of the model, proving that the proposed few-shot and zero-shot settings are reasonable existing. Finally, the cycle-based captioning framework augments the data with a feature-level image generator and the novelty switcher module to achieve state-of-the-art performance on traditional, few-shot and zero-shot settings. Although the cycle-based captioning framework acquires the best ability, the algorithm of the switcher module can still be improved. In the future, we can apply reinforcement learning to design the switcher module, and the reward, as the feedback, can weakly supervise the feature-level image generator.

ACKNOWLEDGMENT

This project is supported by International Exchanges 2022 IEC\NSFC\223523 and Securing the Energy/Transport Interface EP/X037401/1.

REFERENCES

[1] Yang Long, Li Liu, Fumin Shen, Ling Shao, and Xuelong Li. Zero-shot learning using synthesised unseen visual data with diffusion regularisation. *IEEE transactions on pattern analysis and machine intelligence*, 40(10):2498–2512, 2017.

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

[3] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.

[4] Yan Huang, Yang Long, and Liang Wang. Few-shot image and sentence matching via gated visual-semantic embedding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8489–8496, 2019.

[5] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv preprint arXiv:1412.6632*, 2014.

[6] Yang Bai, Junyan Wang, Yang Long, Bingzhang Hu, Yang Song, Maurice Pagnucco, and Yu Guan. Discriminative latent semantic graph for video captioning. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3556–3564, 2021.

[7] Junyan Wang, Yang Bai, Yang Long, Bingzhang Hu, Zhenhua Chai, Yu Guan, and Xiaolin Wei. Query twice: Dual mixture attention meta learning for video summarization. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4023–4031, 2020.

[8] Yang Bai, Desen Zhou, Songyang Zhang, Jian Wang, Errui Ding, Yu Guan, Yang Long, and Jingdong Wang. Action quality assessment with temporal parsing transformer. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV*, pages 422–438. Springer, 2022.

[9] Idan Schwartz, Alexander G Schwing, and Tamir Hazan. A simple baseline for audio-visual scene-aware dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12548–12558, 2019.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[11] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015.

[12] Yunpeng Luo, Jiayi Ji, Xiaoshuai Sun, Liujuan Cao, Yongjian Wu, Feiyue Huang, Chia-Wen Lin, and Rongrong Ji. Dual-level collaborative transformer for image captioning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 2286–2293, 2021.

[13] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10578–10587, 2020.

[14] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Neural baby talk. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7219–7228, 2018.

[15] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, 34:19822–19835, 2021.

[16] 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.

[17] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Making visual representations matter in vision-language models. *CVPR 2021*, 2021.

[18] Lisa Anne Hendricks, Subhashini Venugopalan, Marcus Rohrbach, Raymond Mooney, Kate Saenko, and Trevor Darrell. Deep compositional captioning: Describing novel object categories without paired training data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–10, 2016.

[19] Subhashini Venugopalan, Lisa Anne Hendricks, Marcus Rohrbach, Raymond Mooney, Trevor Darrell, and Kate Saenko. Captioning images with diverse objects. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5753–5761, 2017.

[20] Qianyu Feng, Yu Wu, Hehe Fan, Chenggang Yan, Mingliang Xu, and Yi Yang. Cascaded revision network for novel object captioning. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(10):3413–3421, 2020.

[21] Yang Feng, Lin Ma, Wei Liu, and Jiebo Luo. Unsupervised image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4125–4134, 2019.

[22] Iro Laina, Christian Rupprecht, and Nassir Navab. Towards unsupervised image captioning with shared multimodal embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7414–7424, 2019.

[23] Hila Chefer, Sagie Benaim, Roni Paiss, and Lior Wolf. Image-based clip-guided essence transfer. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIII*, pages 695–711. Springer, 2022.

[24] Aditya Sanghi, Hang Chu, Joseph G Lambourne, Ye Wang, Chin-Yi Cheng, Marco Fumero, and Kamal Rahimi Malekshan. Clip-forge: Towards zero-shot text-to-shape generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18603–18613, 2022.

[25] 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.

[26] 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.

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

[28] 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.

[29] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How much can clip benefit vision-and-language tasks? *arXiv preprint arXiv:2107.06383*, 2021.

[30] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7008–7024, 2017.

[31] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018.

[32] Wenhao Jiang, Lin Ma, Yu-Gang Jiang, Wei Liu, and Tong Zhang. Recurrent fusion network for image captioning. In *Proceedings of the European conference on computer vision (ECCV)*, pages 499–515, 2018.

[33] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *Proceedings of the European conference on computer vision (ECCV)*, pages 684–699, 2018.

[34] Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. Image captioning: Transforming objects into words. *Advances in neural information processing systems*, 32, 2019.

[35] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4634–4643, 2019.

[36] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.

[37] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[38] 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.

[39] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.