MultiCapCLIP: Auto-Encoding Prompts for Zero-Shot Multilingual Visual Captioning

Bang Yang[1,2]*, Fenglin Liu[3]*, Xian Wu[4], Yaowei Wang[2], Xu Sun[5]†, and Yuexian Zou[1]†
[1]ADSPLAB, School of ECE, Peking University [2]Peng Cheng Laboratory
[3]University of Oxford [4]Tencent Jarvis Lab [5]School of Computer Science, Peking University
{yangbang, zouyx}@pku.edu.cn; fenglin.liu@eng.ox.ac.uk

Abstract

Supervised visual captioning models typically require a large scale of images or videos paired with descriptions in a specific language (i.e., the vision-caption pairs) for training. However, collecting and labeling large-scale datasets is time-consuming and expensive for many scenarios and languages. Therefore, sufficient labeled pairs are usually not available. To deal with the label shortage problem, we present a simple yet effective zero-shot approach MultiCapCLIP that can generate visual captions for different scenarios and languages without any labeled vision-caption pairs of downstream datasets. In the training stage, MultiCapCLIP only requires text data for input. Then it conducts two main steps: 1) retrieving concept prompts that preserve the corresponding domain knowledge of new scenarios; 2) auto-encoding the prompts to learn writing styles to output captions in a desired language. In the testing stage, MultiCapCLIP instead takes visual data as input directly to retrieve the concept prompts to generate the final visual descriptions. The extensive experiments on image and video captioning across four benchmarks and four languages (i.e., English, Chinese, German, and French) confirm the effectiveness of our approach. Compared with state-of-the-art zero-shot and weakly-supervised methods, our method achieves 4.8% and 21.5% absolute improvements in terms of BLEU@4 and CIDEr metrics. Our code is available at https://github.com/yangbang18/MultiCapCLIP.

1 Introduction

Visual captioning targets to first 1) understand the information of visual inputs, which are typically videos or images, and then 2) produces a corresponding textual sentence describing the visual objects/attributes/relationships. Visual captioning has drawn remarkable attention from natural language processing and computer vision fields due to its wide applications, e.g., cross-modal retrieval (Luo et al., 2022; Cheng et al., 2023b) and helping the visually impaired (Çaylı et al., 2021). Currently, visual captioning models based on the encoder-decoder framework (Huang et al., 2020; Liu et al., 2020; Yang et al., 2021; Zhang et al., 2021; Hu et al., 2022; Lin et al., 2022) have achieved tremendous progress in advancing the state-of-the-art. These models are usually trained with full supervision and rely on large-scale humanly-annotated training data (i.e., vision-caption pairs), which needs expensive labeling work. In particular, when it comes to Non-English caption systems, it is challenging to collect and label sufficient vision-caption pairs in a timely manner, which prevents such encoder-decoder models from rapid deployment in different scenarios and languages.

To deal with the shortage of labeled pairs, we propose the MultiCapCLIP - a prompt-based natural language auto-encoder. As shown in Figure 1, MultiCapCLIP only requires textual input for training, and it can conduct zero-shot multilingual visual captioning, including image and video captioning. Therefore, MultiCapCLIP can deal with the situation where the labeled vision-caption pairs are missing. MultiCapCLIP is particularly suitable for new scenarios and languages, improving the practical value of visual captioning.

To implement MultiCapCLIP, we first adopt a pre-trained vision-language model, i.e., CLIP (Radford et al., 2021), as our backbone. CLIP has shown success in correlating the visual and textual modalities into the same latent space (vision-language embedding space) (Tewel et al., 2022b; Su et al., 2022; Zeng et al., 2023). We observe two critical issues for zero-shot visual captioning: the understanding of domain visual knowledge (e.g., objects, attributes, and relationships) and the generation of descriptive sentences in a specific writing style and language. Therefore, we propose a prompt-based

*Equal contribution.
†Corresponding authors.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 11908–11922

auto-encoder, which introduces the visual concept prompts $\mathcal{P}$ to preserve the corresponding domain knowledge and writing styles of zero-shot visual captioning. During training, given the text-only data, we train the model by reconstructing the caption S in the $S \rightarrow \mathcal{P} \rightarrow S$ auto-encoding pipeline. Since the auto-encoding process reconstructs the same input sentence, the model training needs only unlabeled text data. In the reconstruction process, the model is able to preserve the necessary domain knowledge and the writing styles of visual captioning (Wang et al., 2016; Tschannen et al., 2018). During inference, we can directly take the vision input V as queries to retrieve the domain knowledge preserved in the visual concept prompts and finally rely on the learned writing styles in a specific language in the text decoder to generate visual descriptions in the $V \rightarrow \mathcal{P} \rightarrow S$ pipeline.

Meanwhile, to further bridge the modality gap between the visual and textual data (Liang et al., 2022), we introduce an augmentation method, including input augmentation and feature augmentation, which can boost the robustness of the model and in turn improve the performance of zero-shot visual captioning. The experiments on four benchmark datasets, i.e., MS-COCO (Chen et al., 2015), MSR-VTT (Xu et al., 2016), VATEX (Wang et al., 2019), and Multi30K (Elliott et al., 2016), show that our approach can accurately and data-efficiently generate visual captions in English, Chinese, German, and French.

Overall, our main contributions are as follows:

- We propose a simple yet effective approach MultiCapCLIP that requires no downstream labeled data to make the first attempt for zero-shot multilingual visual captioning.
- MultiCapCLIP first introduces visual concept prompts to preserve the domain knowledge and then auto-encodes them to learn the writing styles of captioning. After text-only training, our approach can shift from text-to-text generation to vision-to-text generation.
- The out-of-domain and in-domain experiments on image and video captioning across different languages show that our approach trained on text-only data significantly outperforms previous zero-shot/weakly-supervised methods trained on unpaired or partial labeled visual and textual data, setting new state-of-the-art zero-shot performance.

2 Approach

In this section, we first give a brief review of CLIP, whose vision-language embedding space lays a foundation for our approach. Next, we introduce the framework of the proposed MultiCapCLIP, followed by two key components: concept prompts and textual augmentations.

2.1 A Brief Review of CLIP

CLIP uses two independent encoders to process image and text input separately and then bridges the gap between modalities with contrastive learning. The image encoder $\phi_v(\cdot)$ can be a convolutional neural network like ResNet (He et al., 2016) or a vision Transformer like ViT (Dosovitskiy et al., 2021), and it extracts a feature vector for each input image. The text encoder $\phi_t(\cdot)$ is based on Transformer (Vaswani et al., 2017), and it outputs a vector representation of the input text. By training two encoders on 400M image-text data with noisy correspondences under the InfoNCE objective (Oord et al., 2018), CLIP learns a powerful vision-language embedding space that measures image-text similarity well and enables open-vocabulary classification. In this paper, we re-purpose CLIP for zero-shot multilingual visual captioning and always keep $\phi_v(\cdot)$ and $\phi_t(\cdot)$ frozen.

2.2 Overview of MultiCapCLIP

As shown in Figure 1, MultiCapCLIP consists of visual and textual encoders from CLIP and a trainable Multilingual Language Model (MLM). MultiCapCLIP supports English text[1], images or videos as inputs and can produce output in desired language. Specifically, we implement MLM with a stack of Transformer decoder blocks, each of which comprises a masked self-attention layer, a cross-attention layer, and a feed-forward layer. Moreover, we add explicit signals in the embedding layer to indicate which language to be generated.

Let denote the text input as S, the vision input as V, and concept prompts as P. Unlike typical visual captioning models that are trained on a vision-text dataset, MultiCapCLIP relies on a text dataset and follows the $S \rightarrow P \rightarrow S$ auto-encoding pipeline during training. Based on the semantic alignment characteristic of CLIP's feature space, MultiCapCLIP uses the $V \rightarrow P \rightarrow S$ pipeline for visual captioning during inference. We extend MultiCapCLIP to support multilingual text generation by us-

[1]The training corpora for CLIP is mainly in English.

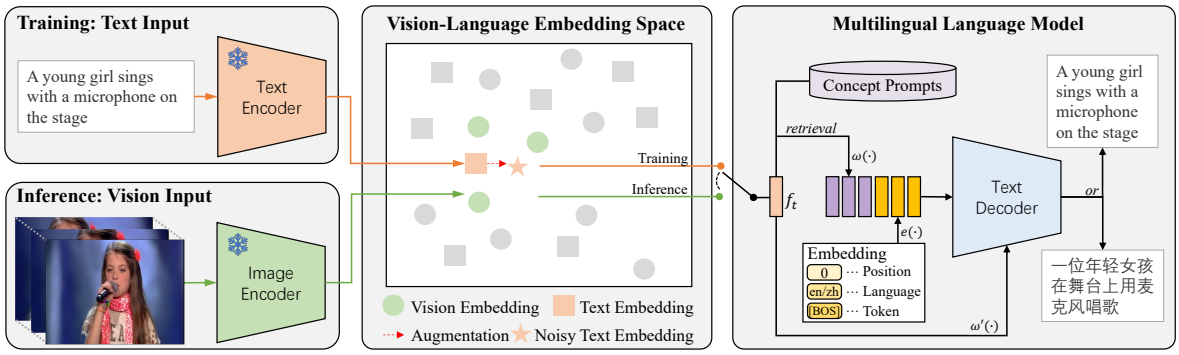

Figure 1: Illustration of MultiCapCLIP. It comprises of a frozen CLIP-like vision-language pre-trained model and a trainable multilingual language model. During training, MultiCapCLIP only requires text data and aims to produce a reconstruction/translation result based on the text feature f_t of the source input and its relevant concept prompts (§ 2.3). During inference, MultiCapCLIP replaces f_t with the feature(s) of an image or a video for captioning.

ing parallel corpora with (S, T) pairs, where T denotes the target text in a desired language. In such a case, MultiCapCLIP follows the $S/V \rightarrow P \rightarrow T$ translation pipeline.

In the following, we will detail how to extract and leverage P in Section 2.3. Then in Section 2.4, we will introduce an augmentation method to improve the training of MultiCapCLIP.

2.3 Decoding with Concept Prompts

A set of visual concepts is a good embodiment of domain visual knowledge because a visual concept (e.g., "a young girl") manifest as the explicit clue in the vision input. Given a pure text dataset, we use the spaCy toolkit[2] to extract noun phrases and reserve the most frequent 1,000 noun phrases as visual concepts, which are first embedded into a prompt template "{concept}"[3] and then fed into the CLIP's text encoder $\phi_t(\cdot)$ to extract l2-normalized concept features $C = \{c_1, \ldots, c_{1000}\}$.

During training, given the text input S, we first encode it into a global feature f_t:

$$f_t = \text{Norm}(\phi_t(S)), \tag{1}$$

where $\text{Norm}(\cdot)$ denotes L2 normalization. Next, we calculate the dot product of f_t and C to measure cosine similarities, based on which we obtain *soft* concept prompts P, a subset of C that includes K concept features most semantically similar to f_t. Assuming that the dimension of vectors outputted by CLIP is d, P is in the shape of $K * d$. To prompt MLM, we prefix embeddings of the target text S with P to obtain the final input embeddings E:

$$E = \text{Concat}(\omega(P), e(S)) \tag{2}$$

where $\omega(\cdot)$ is implemented as a fully connected layer followed by a layer normalization (LN) (Ba et al., 2016), and $e(\cdot)$ denotes the summation of position, language, and token embeddings for each $s_i \in S$, followed by LN. The prompt sequence generated by $\omega(P)$ and token sequence generated by $e(S)$ are concatenated together and sent to the text decoder of MLM to regenerate the input sequence S. Considering that f_t may contain information supplementary to P, we do not discard f_t. We first feed the projected feature $f = \omega'(f_t)$, where $\omega'(\cdot)$ has the same structure as $\omega(\cdot)$ but shares no parameters, into the text decoder of MLM. Then we calculate the cross attention between f and E. We train the model with a cross-entropy loss:

$$\mathcal{L} = -\sum_{i=1}^{|S|} \log p_\theta(s = s_i | S_{<i}, P, f_t), \tag{3}$$

where $p_\theta(\cdot)$ is MLM's predicted distribution over a vocabulary and θ denotes all trainable parameters.

During inference, we process the vision input V in a similar manner, except that we use CLIP's image encoder $\phi_v(\cdot)$ to obtain V's vector representation f_v and obtain relevant concept prompts P based on (averaged) image-concept similarities. Given the previously generated text $S_{<i}$, the prediction of the next token is based on the following probability distribution:

$$p_\theta(s | S_{<i}, P, f_v). \tag{4}$$

2.4 Training with Augmentations

Our MultiCapCLIP's ability of shifting text-to-text generation to vision-to-text generation is built on

[2] https://spacy.io

[3] The simplest prompt template "{concept}" produced better performance than other templates like "a concept of {concept}" in our preliminary experiments.

	MS-COCO	MSR-VTT	VATEX
Vision Type	Image	Video	Video
Training size	113,287	6,513	25,006
Validation size	5,000	497	1,393
Testing size	5,000	2,990	1,500
# Captions	616,767	200,000	278,990
Avg. Length	10.6	9.3	12.3
Target Language	English	English	Chinese

Table 1: Statistics of datasets used in main experiments.

Settings and Languages		Training Data + Prompts (*Text-only data*)	Testing Data
Out-of-Domain	English	MSR-VTT MS-COCO	MS-COCO MSR-VTT
	Chinese	MSR-VTT-CN	VATEX
In-Domain	English	MS-COCO MSR-VTT	MS-COCO MSR-VTT
	Chinese	VATEX	VATEX

Table 2: Training and testing data used for different experimental settings. We adopt two English captioning datasets: MS-COCO (Chen et al., 2015), MSR-VTT (Xu et al., 2016) and two Chinese captioning datasets: MSR-VTT-CN (Wang et al., 2022b), and VATEX (Wang et al., 2019), to conduct the main experiments.

the assumption that the paired vision-text data is well aligned in the vision-language embedding space of CLIP. However, Liang et al. (2022) demonstrated that there exists *modality gap* in CLIP-like models and such gap has a significant impact on model generalization ability. To this end, inspired by *denoising* auto-encoders (Vincent et al., 2008), we propose to train MultiCapCLIP with augmented text features f'_t. Here we consider both the input augmentation (IA) and the feature augmentation (FA). Specifically, IA replaces the source text S with a semantically similar one S' to obtain f'_t:

$$f'_t = \text{Norm}(\phi_t(S')), \tag{5}$$

where $S' \sim \mathbb{X}_S$ and $\mathbb{X}_S = \{S, S'_1, \ldots, S'_{N-1}\}$ denotes the candidate set of S. For simplicity, we use $\phi_t(\cdot)$ to measure similarities among text in the dataset and select the most similar $N-1$ text to construct $\mathbb{X}_S$ for each S. Since we sample text from $\mathbb{X}_S$ with uniform probability, there will be $1/N$ probability that the input text keeps unchanged. As for FA, we follow Li et al. (2021) to add Gaussian noise $n \sim \mathcal{N}(0, \epsilon)$ into text features. Hence, Eq. (5) can be further extended to:

$$f'_t = \text{Norm}(\text{Norm}(\phi_t(S')) + n). \tag{6}$$

During training, we replace f_t in Eq. (1) with f'_t in Eq. (6) to encourage the model to learn more robust latent representations.

3 Main Experiments

In this section, we first introduce the datasets, metrics, settings of the experiments; Then, we provide the out-of-domain and in-domain results of our approach for zero-shot visual captioning.

3.1 Experimental Setups

Datasets. As shown in Table 1, we use three benchmark datasets under CC BY 4.0 licence in this section: MS-COCO (Chen et al., 2015), MSR-VTT (Xu et al., 2016), and VATEX (Wang et al., 2019). We apply the Karpathy and Fei-Fei's (2015) split to MS-COCO and follow the official split of MSR-VTT for English captioning. Besides, VATEX is a multilingual video captioning dataset that contains parallel English-Chinese captions. We use it for Chinese captioning[4]. In Section 4, we will further use the Multi30K dataset (Elliott et al., 2016) for German and French caption generation.

Metrics. Following the common practice in the literature, we report BLEU@4 (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ROUGE-L (Lin, 2004) and CIDEr (Vedantam et al., 2015) for video captioning, and additionally measure SPICE (Anderson et al., 2016) for image captioning. All metrics are computed by Microsoft COCO Evaluation Server[5] (Chen et al., 2015).

Settings As shown in Table 2, we conduct the out-of-domain and in-domain experiments. 1) *Out-of-Domain Experiments* are performed by training the model on the text-only data of A dataset, and then evaluating on the B dataset. 2) *In-Domain Experiments* are conducted by training the model on the text-only data of A dataset, and then evaluating on the A dataset.

[4]The official splits of VATEX is 25,991: 3,000: 6,000. However, some video clips are no longer available, resulting in the splits 25,006: 2,893: 5,792. Besides, VATEX does not provide Chinese captions of the test set (https://eric-xw.github.io/vatex-website). We construct new validation and test sets from the original validation set.

[5]For non-English captioning, we do not report METEOR and SPICE metrics because their implementations consider synonym matching and named entity recognition in English.

Settings	Methods	Pre-trained Backbone	Training Data		MS-COCO (English)					MSR-VTT (English)				VATEX (Chinese)		
			Vision	Text	B@4	M	R-L	C	S	B@4	M	R-L	C	B@4	R-L	C
Weakly-Supervised	UIC† (2019)	Inception + Faster R-CNN	√	√	5.6	12.4	28.7	28.6	8.1	-	-	-	-	-	-	-
	IC-SME† (2019)	ResNet + Faster R-CNN	√	√	6.5	12.9	35.1	22.7	-	-	-	-	-	-	-	-
	R^2M† (2020)	Faster R-CNN	√	√	6.4	13.0	31.3	29.0	**9.1**	-	-	-	-	-	-	-
	TSGAN† (2021)	Faster R-CNN	√	√	6.9	13.0	32.3	28.9	8.3	-	-	-	-	-	-	-
	SGM† (2021)	Inception + Faster R-CNN	√	√	6.3	14.0	34.5	**31.9**	8.6	-	-	-	-	-	-	-
Zero-Shot	ZeroCap (2022b)	CLIP + GPT	✗	✗	2.6	11.5	-	14.6	5.5	2.3	12.9	30.4	5.8	-	-	-
	MAGIC (2022)	CLIP + GPT	✗	√	5.2	12.5	30.7	18.3	5.7	5.5	13.3	35.4	7.4	-	-	-
	EPT (2022a)	CLIP + GPT	✗	✗	-	-	-	-	-	3.0	14.6	27.7	11.3	-	-	-
	ZS-CapCLIP* (2021)	CLIP	✗	√	3.4	13.0	27.6	12.2	6.2	4.0	15.0	31.0	5.0	2.8	25.8	2.0
	MultiCapCLIP (**Ours**)	CLIP	✗	√	**9.7**	**15.8**	**37.6**	30.2	8.9	**13.3**	**19.5**	**43.3**	**15.5**	**8.4**	**31.2**	**6.2**

Table 3: **Out-of-domain visual captioning results**. B@4, M, R-L, C, and S are short for BLEU@4, METEOR, ROUGE-L, CIDEr, and SPICE, respectively. †: Training with a corpus with more than 2.3M sentences. * denotes our implementations. All previous works can not deal with multilingual zero-shot captioning. Our approach achieves the best results on most metrics across three datasets.

Baselines Since previous works can not generate zero-shot multilingual visual captions directly, we implement a zero-shot CLIP-based model: *ZS-CapCLIP*, which is trained on text-only data with the same architecture as our MultiCapCLIP but without our proposed concept prompts and text augmentations. To observe the gap between zero-shot and fully-supervised methods, We also implement *CapCLIP* trained on vision-caption pairs.

Implementations. Following the previous works in zero-shot captioning (Tewel et al., 2022b; Su et al., 2022; Zeng et al., 2023), we adopt the CLIP (ViT-B/16 variant) (Radford et al., 2021) as our image encoder and text encoder, and adopt a randomly initialized Transformer-BASE (Vaswani et al., 2017) as our language decoder. We adopt the same vocabulary as BERT / multilingual BERT (Devlin et al., 2019) for English / non-English captioning. We use the Jieba toolkit[6] to segment Chinese sentences. We select the hyperparameter K from values $\{4, 8, 16, 32\}$, N from values $\{5, 10, 20\}$ and ϵ from values $\{0.01, 0.1, 1.0\}$ according to the CIDEr performance on the validation split, and set $K = 16$, $N = 5$, $\epsilon = 0.01$ for all datasets and settings except that $\epsilon = 0.1$ for in-domain experiments on MS-COCO. During training, we apply label smoothing (Szegedy et al., 2016) of 0.1, use batches of 32 samples and AdamW (Loshchilov and Hutter, 2019) with L2 weight decay of 0.01 to train models for 10 epochs. We set the learning rate fixed to 1e-4 with 10% warm-up iterations when training on text-only data. During inference, we use beam search with a beam size of 3 to generate captions.

[6] https://github.com/fxsjy/jieba

3.2 Out-of-Domain Results

In this section, we evaluate the zero-shot multilingual captioning performance of our approach under out-of-domain settings. We can notice from Table 3 that our zero-shot model MultiCapCLIP achieves competitive performance on three datasets across English and Chinese. Although SGM (Honda et al., 2021) and R^2M (Guo et al., 2020) perform better than our model on CIDEr and SPICE metrics on MS-COCO, they require the large-scale image datasets for training and use a larger training corpus (2.3M sentences) than ours (130K sentences). While the previous methods do not target non-English caption generation, our MultiCapCLIP gains obvious relative improvements against the CapCLIP on VATEX Chinese captioning. The out-of-domain results show that our approach is able to generate multilingual visual captions without any labeled vision-caption data, which could have the potential to promote the application of visual captioning for low-resource language applications.

3.3 In-Domain Results

For comparisons, we further consider state-of-the-art fully-supervised and large-scale pre-trained models and models under the *unpaired* setting, i.e., both vision and text data of the target dataset are utilized for training independently, leaving their pairing annotations unused. As shown in Table 4, our approach significantly outperforms previous unpaired/zero-shot competitors by up to 4.8% BLEU@4, 3.9% ROUGE-L, and 21.5% CIDEr scores in MS-COCO English captioning. When it comes to MSR-VTT English captioning and VATEX Chinese captioning, our MultiCapCLIP surpasses ZS-CapCLIP by a large margin under the CIDEr metric, e.g., an absolute improvement of

Settings	Methods	MS-COCO (English)					MSR-VTT (English)				VATEX (Chinese)		
		B@4	M	R-L	C	S	B@4	M	R-L	C	B@4	R-L	C
Fully-Supervised	VATEX (2019)	-	-	-	-	-	-	-	-	-	23.4	46.0	39.4
	MAD+SAP (2020)	37.0	28.1	57.2	117.3	21.3	41.3	28.3	61.4	48.5	-	-	-
	$\text{Oscar}_{\text{base}}$ (2020)	36.5	30.3	-	123.7	23.1	-	-	-	-	-	-	-
	OpenBook (2021)	-	-	-	-	-	42.8	29.3	61.7	52.9	-	-	-
	ClipCap (2021)	33.5	27.5	-	113.1	21.1	-	-	-	-	28.3*	49.5*	51.3*
	CapCLIP* (2021)	32.3	27.7	55.4	109.5	20.7	42.9	29.8	62.3	54.5	29.7	49.8	51.0
	CaMEL (2022)	39.1	29.4	58.5	125.7	22.2	-	-	-	-	-	-	-
	$\text{LEMON}_{\text{base}}$ (2022)	40.3	30.2	-	133.3	23.3	-	-	-	-	-	-	-
	SwinBERT (2022)	-	-	-	-	-	45.4	30.6	64.1	55.9	-	-	-
	CLIP-DCD (2022a)	-	-	-	-	-	48.2	31.3	64.8	58.7	-	-	-
	MV-GPT (2022)	-	-	-	-	-	48.9	38.7	64.0	60.0	-	-	-
	GIT (2022a)	44.1	31.5	-	144.8	24.7	53.8	32.9	67.7	73.9	-	-	-
Weakly-Supervised	UIC (2019)	18.6	17.9	43.1	54.9	11.1	-	-	-	-	-	-	-
	IC-SME (2019)	19.3	20.2	45.0	61.8	12.9	-	-	-	-	-	-	-
	Graph-Align (2019)	21.5	20.9	47.2	69.5	15.0	-	-	-	-	-	-	-
	IGGAN (2020)	21.9	21.1	46.5	64.0	14.5	-	-	-	-	-	-	-
	TSGAN (2021)	18.9	18.2	43.3	55.2	11.3	-	-	-	-	-	-	-
	USGAE (2022c)	17.1	19.1	43.8	55.1	12.8	-	-	-	-	-	-	-
	SCS (2022)	22.8	21.4	47.7	74.7	15.1	-	-	-	-	-	-	-
Zero-Shot	MAGIC (2022)	12.9	17.4	39.9	49.3	11.3	-	-	-	-	-	-	-
	ZS-CapCLIP* (2021)	6.1	15.8	33.0	27.3	9.3	8.6	19.8	37.3	11.1	21.2	45.0	31.8
	MultiCapCLIP (**Ours**)	**27.6**	**25.2**	**51.6**	**96.2**	**18.5**	**22.0**	**24.4**	**50.2**	**33.6**	**22.8**	**46.0**	**38.2**

Table 4: **In-domain visual captioning results**. Fully-supervised and large-scale pre-trained models are included for comparisons. * denotes our implementations. Our MultiCapCLIP significantly outperforms previous zero-shot/weakly-supervised models, but still suffers from performance gaps compared with fully-supervised counterparts.

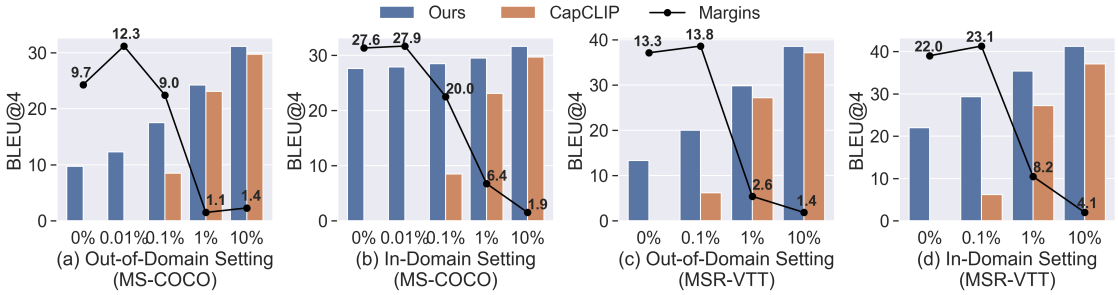

Figure 2: Results of out-of-domain and in-domain experiments with respect to different ratios of training data. The margins in different ratios are shown with the polyline. Our method consistently surpasses CapCLIP, and the fewer the vision-caption pairs, the larger the margins.

22.5% on MSR-VTT. These results prove the effectiveness of MultiCapCLIP in zero-shot multilingual visual captioning. Nevertheless, there still exists performance gaps between MultiCapCLIP trained on text-only data and existing state-of-the-art fully-supervised models trained on full vision-text data.

4 Analysis

In this section, we conduct several analyses to better understand our approach.

4.1 Semi-Supervised Visual Captioning

To further prove the effectiveness of our approach, we fine-tune MultiCapCLIP with partial labeled vision-caption data of downstream datasets. To this end, in Figure 2, we evaluate the performance of MultiCapCLIP with respect to the increasing amount of labeled data. Specifically, we randomly sample a small portion of training images/videos and use the resulting vision-caption pairs for fine-tuning. We repeat this process by three times and report the average performance. For a fair comparison, we also train CapCLIP (Section 3.1) with the same amount of pairs. As we can see in Figure 2, for both in-domain or a out-of-domain corpus, MultiCapCLIP consistently outperforms CapCLIP with different ratios of training data. It is worth noting that the fewer the labeled vision-caption pairs, the larger the margins. In detail, under the extremely low label setting, e.g., 0.1% of paired data on MSR-VTT (only 6 videos), our approach under the in-domain setting significantly surpasses the

Setting	Component			K	Concept Type	Out-of-Domain Setting					In-Domain Setting				
	CP	IA	FA			B@4	M	R-L	C	S	B@4	M	R-L	C	S
Base Model				-	-	3.4	13.0	27.6	12.2	6.2	6.1	15.8	33.0	27.3	9.3
(a)	√			16	Noun	5.7	15.2	32.4	18.1	8.6	7.7	17.6	35.9	40.6	12.3
(b)		√		-	-	7.4	15.3	34.1	23.8	8.5	15.1	20.5	42.2	57.6	14.5
(c)			√	-	-	5.5	14.3	32.1	15.3	7.3	26.1	25.2	51.2	91.5	18.3
(d)		√	√	-	-	7.4	15.7	34.4	23.9	**9.2**	26.6	25.2	51.3	92.5	18.4
(e)	√	√	√	4	Noun	8.2	15.1	35.9	27.7	8.1	27.7	25.2	51.9	94.6	18.4
(f)	√	√	√	8	Noun	8.1	15.6	35.3	29.3	8.5	27.5	25.1	51.6	95.0	18.3
(g) Full Model	√	√	√	16	Noun	**9.7**	15.8	**37.6**	**30.2**	8.9	27.6	**25.2**	51.6	**96.2**	**18.5**
(h)	√	√	√	32	Noun	9.1	**16.2**	37.1	30.1	9.1	**28.4**	**25.2**	**51.9**	95.7	**18.5**
(i)	√	√	√	16	Verb	7.0	15.0	34.1	21.1	7.1	27.8	**25.2**	**51.9**	93.2	18.3
(j)	√	√	√	16	Noun + Verb	9.2	15.7	37.0	28.4	8.6	27.1	25.1	51.4	94.3	18.5

Table 5: Quantitative analysis of the proposed MultiCapCLIP, which includes visual concept prompts (CP), input augmentation (IA), and feature augmentation (FA). We conduct the ablation study on the MS-COCO dataset under out-of-domain and in-domain settings. The full model denotes our proposed MultiCapCLIP.

CapCLIP by 23.1% absolute BLEU@4 score. It further proves the effectiveness of our approach, which can relax the reliance on the vision-caption annotations. We can make use of available unpaired text-only data as a solid basis for multilingual visual captioning tasks.

4.2 Quantitative Analysis

In this section, we analyze the contributions of each component in our approach.

Ablation Study We conduct the ablation study on the out-of-domain and in-domain settings using MS-COCO dataset (Chen et al., 2015). As shown in Table 5, each component in our proposed approach can boost the performance over all metrics, verifying our arguments and the effectiveness of our approach. In particular, setting (a) shows that the introduced prompts can improve the base model with absolute gains up to 5.9% and 13.3% CIDEr scores under out-of-domain and in-domain settings, respectively. Settings (b,c) show that either input augmentation (IA) or feature augmentation (FA) respectively boost performance, indicating the importance of bridging the modality gap between the visual and textual data and in turn, boosting the robustness of the model and improving the performance of zero-shot visual captioning. Moreover, by comparing the results of (b) and (c), we observe that FA brings more improvements under the in-domain setting whereas IA is better under the out-of-domain setting. This indicates that structure noises are more suitable to bridge the modality gap between vision and text data from the same domain. From another perspective, we need a more complex feature adaptation method for out-of-domain transfer. Since the IA and FA can improve the performance from different perspectives, as shown in setting (d), combining them can lead to the most prominent improvement across all metrics. Moreover, compared with (d), our full model in the setting (g) can still gain improvements under most metrics, especially the CIDEr metric, showing that concept prompts benefit visual captioning by generating more accurate details.

Effect of K As shown in Table 5 (e-h), when we set the number of prompts K to 16, the model substantially performs the best. For other K values, when $K < 16$, the performance is improved with an increasing K due to more adequate guidance signals to the model. However, when $K > 16$, we can observe saturated or impaired captioning performance, possibly because retrieving more prompts do not include additional useful clues and introduce irrelevant noises to the model.

Concept Type Other than prompting the model with noun phrases (Section 2.3), we also consider the effect of verbs. As shown in Table 5, setting (g) surpasses settings (i) and (j) at most cases, i.e., using verb-based prompts degrades performance. We speculate the reason is that the vision-language model we used (i.e., CLIP) can recognize salient objects more accurately than human actions.

4.3 Robustness Analysis: Extensions to More Non-English Languages

We adopt the Multi30K dataset (Elliott et al., 2016) to further evaluate in-domain performance on German and French image captioning. As shown in

Figure 3: Captioning comparisons between ground-truths (GT), CapCLIP trained with full in-domain vision-text pairs, and our zero-shot MultiCapCLIP trained with out-of-domain corpora. We emphasize **accurate** and ***wrong*** keywords and highlight reasonable and *noisy* concepts used for prompting. Our approach can generate plausible visual descriptions in English and Chinese without the need of vision-caption pairs.

Setting	German			French		
	B@4	R-L	C	B@4	R-L	C
Supervised	20.0	45.7	55.8	7.1	28.0	54.0
ZS-Base Model	3.8	27.7	10.7	2.6	19.4	20.4
ZS-Full Model	**13.3**	**38.3**	**36.7**	**5.2**	**23.9**	**40.5**

Table 6: In-domain performance on German and French image captioning. ZS is short for "Zero-Shot".

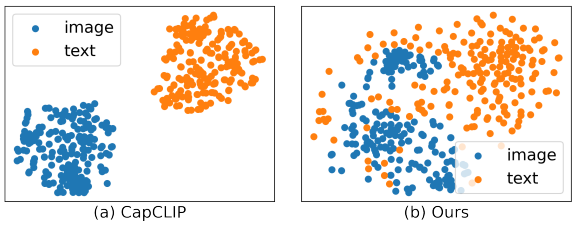

Figure 4: T-SNE visualization (van der Maaten and Hinton, 2008) of image and text embeddings produced by (a) ZS-CapCLIP and (b) our MultiCapCLIP during training. We plot the scatter diagrams for 200 image-caption pairs. Our approach can effectively bridge the gap between the vision and text modalities.

Table 6, our full model again outperforms the base model by a large margin, proving the effectiveness of concept prompts and text augmentations.

4.4 Qualitative Analysis

In this section, we give some visualization results and examples to better understand our approach.

Visualization To verify the effect of our method on representation learning, we use t-SNE (van der Maaten and Hinton, 2008) to visualize the features produced by ZS-CapCLIP and our MultiCapCLIP in Figure 4, which shows that our approach can bridge the modality gap between visual and textual inputs during training and obtain a blended distribution, leading to a more robust shift from text-to-text generation to vision-to-text generation.

Examples In Figure 3, we compare our model trained with out-of-domain corpora with CapCLIP trained on full in-domain supervision. As we can see, our model can generate accurate keywords, e.g., "sand" in (a), "tire" in (c), and "helmet" in (d), which can be attributed to the useful clues of concept prompts. However, there exist noises in the retrieved concepts in some cases, e.g., "a punching bag" in (b), misleading the model to produce wrong details. Besides, in (e), we can observe how the training corpus affect the writing style of the model: the corpus of a video caption dataset (VATEX) makes the model focus on the temporal evolution of events, resulting in a speculative description "the catcher catches the ball". Overall, our approach can be a solid basis for zero-shot multilingual visual captioning. It requires no vision-caption pairs but generates plausible visual descriptions.

5 Related Works

The related works are introduced from zero-shot learning and visual captioning.

Zero-shot Learning Adapting models to novel tasks with limited labeled data is an important research topic toward general intelligence (Griffiths et al., 2019). Contrastive pre-training is an effective technique to achieve this goal and has revolutionized multimodal research (Hou et al., 2021; Gan et al., 2022; Jin et al., 2022; Cheng et al., 2023a). Specifically for the vision-language field, models such as CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) learn a shared multimodal embedding space from large-scale noisy image-text pairs, leading to an impressive zero-shot performance on tasks like image classification and vision-text retrieval (Zhang et al., 2022; Luo et al., 2022). Nevertheless, employing CLIP-like models in low-data vision-grounded text generation (i.e., visual captioning) remains challenging.

Visual Captioning As a key vision-language task, visual captioning has achieved tremendous progress under the encoder-decoder framework (Xu et al., 2015) and the "pre-training and fine-tuning" paradigm. Yet, typical visual captioning methods require curated datasets of numerous images or videos paired with descriptions in a specific language, which are costly to collect. To this end, some weakly-supervised approaches are proposed (Feng et al., 2019; Guo et al., 2020; Honda et al., 2021; Ben et al., 2022). These methods require disjoint vision and text data for training and rely on a pre-trained object detector like Faster R-CNN (Ren et al., 2015) to construct weak supervision signals. However, the detectors they use are limited to a pre-defined set of categories. Recently, several works integrate CLIP with large language models (LLMs) like GPT (Radford et al., 2019; Brown et al., 2020) for zero-shot visual captioning (Tewel et al., 2022b; Su et al., 2022; Liu et al., 2022; Zeng et al., 2023). Although effective, these methods suffer from over-parameterization of large LLMs. We instead train a lightweight decoder from scratch. Besides, some concurrent works address zero-shot visual captioning by training CLIP with text-only data (Nukrai et al., 2022; Gu et al., 2022; Li et al., 2023; Yang et al., 2023). What differentiates our work from them is that we consider visual concept prompts that perverse domain visual knowledge.

6 Conclusions

We have presented a data-efficient method dubbed MultiCapCLIP to re-purpose CLIP-like vision-language pre-trained models for zero-shot multilingual visual captioning. Our approach reduces the reliance on labeled vision-caption pairs of downstream datasets by auto-encoding concept prompts on text-only data. Extensive experiments on four datasets and four languages confirm the effectiveness of our approach, which can be a solid basis for visual captioning in low-data regimes and low-resource languages.

Limitations

Although the proposed MultiCapCLIP can generate multilingual zero-shot visual captions without any labeled vision-caption training pairs. We still need the independent set of text for training/translating, which may still be difficult to collect for some low-resource languages. This might be alleviated in the future with techniques such as knowledge distillation from publicly-available pre-trained models, e.g., BERT (Devlin et al., 2019). Besides, our approach uses CLIP to measure text-text similarities for retrieving concept prompts and conducting input augmentation during training. Considering that CLIP is optimized by image-text global contrast (Radford et al., 2021) and intra-modal retrieval of such a model is not as well as its cross-modal retrieval (Jia et al., 2021), an improvement direction of our approach is using a vision-language pre-trained model that measures intra-modal and inter-modal semantic similarities well (Yang et al., 2022b).

Ethics Statement

We conduct the experiments on public datasets, which are exclusively about natural images, videos, and captions. These datasets have been carefully pre-processed for the academic study purpose, and therefore do not contain any information that names or uniquely identifies individual people or offensive content. It is noteworthy that our approach inherits the drawback of the pre-trained backbone, i.e., CLIP, which has demonstrated that improper class design used for prompting may raise unwanted biases (Radford et al., 2021). Therefore, careful examination is needed before employing our approach in real-world scenarios to avoid prejudices.

Acknowledgements

This paper was partially supported by NSFC (No: 62176008) and Shenzhen Science & Technology Research Program (No: GXWD20201231165807007-20200814115301001).

References

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

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

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Manuele Barraco, Matteo Stefanini, Marcella Cornia, Silvia Cascianelli, Lorenzo Baraldi, and Rita Cucchiara. 2022. Camel: Mean teacher learning for image captioning. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 4087–4094. IEEE.

Huixia Ben, Yingwei Pan, Yehao Li, Ting Yao, Richang Hong, Meng Wang, and Tao Mei. 2022. Unpaired image captioning with semantic-constrained self-learning. *IEEE Transactions on Multimedia*, 24:904–916.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Shan Cao, Gaoyun An, Zhenxing Zheng, and Qiuqi Ruan. 2020. Interactions guided generative adversarial network for unsupervised image captioning. *Neurocomputing*, 417:419–431.

Özkan Çaylı, Burak Makav, Volkan Kılıç, and Aytuğ Onan. 2021. Mobile application based automatic caption generation for visually impaired. In *Intelligent and Fuzzy Techniques: Smart and Innovative Solutions: Proceedings of the INFUS 2020 Conference, Istanbul, Turkey, July 21-23, 2020*, pages 1532–1539. Springer.

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

Xuxin Cheng, Bowen Cao, Qichen Ye, Zhihong Zhu, Hongxiang Li, and Yuexian Zou. 2023a. Ml-lmcl: Mutual learning and large-margin contrastive learning for improving asr robustness in spoken language understanding. In *Findings of the 61st Annual Meeting of the Association for Computational Linguistics*.

Xuxin Cheng, Zhihong Zhu, Hongxiang Li, Yaowei Li, and Yuexian Zou. 2023b. Ssvmr: Saliency-based self-training for video-music retrieval. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30K: Multilingual English-German image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74, Berlin, Germany. Association for Computational Linguistics.

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

Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu, Jianfeng Gao, et al. 2022. Vision-language pre-training: Basics, recent advances, and future trends. *Foundations and Trends® in Computer Graphics and Vision*, 14(3–4):163–352.

Thomas L Griffiths, Frederick Callaway, Michael B Chang, Erin Grant, Paul M Krueger, and Falk Lieder. 2019. Doing more with less: meta-reasoning and meta-learning in humans and machines. *Current Opinion in Behavioral Sciences*, 29:24–30.

Jiuxiang Gu, Shafiq Joty, Jianfei Cai, Handong Zhao, Xu Yang, and Gang Wang. 2019. Unpaired image captioning via scene graph alignments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10323–10332.

Sophia Gu, Christopher Clark, and Aniruddha Kembhavi. 2022. I can't believe there's no images! learning visual tasks using only language data. *arXiv preprint arXiv:2211.09778*.

Dan Guo, Yang Wang, Peipei Song, and Meng Wang. 2020. Recurrent relational memory network for unsupervised image captioning. In *Proceedings of the*

Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence, pages 920–926.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.

Ukyo Honda, Yoshitaka Ushiku, Atsushi Hashimoto, Taro Watanabe, and Yuji Matsumoto. 2021. Removing word-level spurious alignment between images and pseudo-captions in unsupervised image captioning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3692–3702, Online. Association for Computational Linguistics.

Ji Hou, Benjamin Graham, Matthias Nießner, and Saining Xie. 2021. Exploring data-efficient 3d scene understanding with contrastive scene contexts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15587–15597.

Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. 2022. 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.

Yiqing Huang, Jiansheng Chen, Wanli Ouyang, Weitao Wan, and Youze Xue. 2020. Image captioning with end-to-end attribute detection and subsequent attributes prediction. *IEEE Transactions on Image processing*, 29:4013–4026.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR.

Peng Jin, Jinfa Huang, Fenglin Liu, Xian Wu, Shen Ge, Guoli Song, David Clifton, and Jie Chen. 2022. Expectation-maximization contrastive learning for compact video-and-language representations. *Advances in Neural Information Processing Systems*, 35:30291–30306.

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

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

Pan Li, Da Li, Wei Li, Shaogang Gong, Yanwei Fu, and Timothy M Hospedales. 2021. A simple feature augmentation for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8886–8895.

Wei Li, Linchao Zhu, Longyin Wen, and Yi Yang. 2023. Decap: Decoding CLIP latents for zero-shot captioning via text-only training. In *International Conference on Learning Representations*.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer.

Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. 2022. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *Advances in Neural Information Processing Systems*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Kevin Lin, Linjie Li, Chung-Ching Lin, Faisal Ahmed, Zhe Gan, Zicheng Liu, Yumao Lu, and Lijuan Wang. 2022. Swinbert: End-to-end transformers with sparse attention for video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17949–17958.

Fenglin Liu, Xuancheng Ren, Xian Wu, Shen Ge, Wei Fan, Yuexian Zou, and Xu Sun. 2020. Prophet attention: Predicting attention with future attention. *Advances in Neural Information Processing Systems*, 33:1865–1876.

Fenglin Liu, Xian Wu, Chenyu You, Shen Ge, Yuexian Zou, and Xu Sun. 2022. Aligning source visual and target language domains for unpaired video captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9255–9268.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. 2022. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304.

Ron Mokady, Amir Hertz, and Amit H Bermano. 2021. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*.

David Nukrai, Ron Mokady, and Amir Globerson. 2022. Text-only training for image captioning using noise-injected clip. *arXiv preprint arXiv:2211.00575*.

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

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

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

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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

Paul Hongsuck Seo, Arsha Nagrani, Anurag Arnab, and Cordelia Schmid. 2022. End-to-end generative pretraining for multimodal video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17959–17968.

Yixuan Su, Tian Lan, Yahui Liu, Fangyu Liu, Dani Yogatama, Yan Wang, Lingpeng Kong, and Nigel Collier. 2022. Language models can see: Plugging visual controls in text generation. *arXiv preprint arXiv:2205.02655*.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Yoad Tewel, Yoav Shalev, Roy Nadler, Idan Schwartz, and Lior Wolf. 2022a. Zero-shot video captioning with evolving pseudo-tokens. *arXiv preprint arXiv:2207.11100*.

Yoad Tewel, Yoav Shalev, Idan Schwartz, and Lior Wolf. 2022b. Zerocap: Zero-shot image-to-text generation for visual-semantic arithmetic. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17918–17928.

Michael Tschannen, Olivier Bachem, and Mario Lucic. 2018. Recent advances in autoencoder-based representation learning. *arXiv preprint arXiv:1812.05069*.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 4566–4575.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103.

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. 2022a. GIT: A generative image-to-text transformer for vision and language. *Transactions on Machine Learning Research*.

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. 2019. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the International Conference on Computer Vision*, pages 4581–4591.

Yabing Wang, Jianfeng Dong, Tianxiang Liang, Minsong Zhang, Rui Cai, and Xun Wang. 2022b. Cross-lingual cross-modal retrieval with noise-robust learning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 422–433.

Yasi Wang, Hongxun Yao, and Sicheng Zhao. 2016. Auto-encoder based dimensionality reduction. *Neurocomputing*, 184:232–242.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR.

Bang Yang, Fenglin Liu, Yuexian Zou, Xian Wu, Yaowei Wang, and David A Clifton. 2023. Zeronlg: Aligning and autoencoding domains for zero-shot multimodal and multilingual natural language generation. *arXiv preprint arXiv:2303.06458*.

Bang Yang, Tong Zhang, and Yuexian Zou. 2022a. Clip meets video captioning: Concept-aware representation learning does matter. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 368–381. Springer.

Bang Yang, Yuexian Zou, Fenglin Liu, and Can Zhang. 2021. Non-autoregressive coarse-to-fine video captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3119–3127.

Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. 2022b. Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15671–15680.

Xu Yang, Hanwang Zhang, and Jianfei Cai. 2022c. Auto-encoding and distilling scene graphs for image captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2313–2327.

Andy Zeng, Maria Attarian, brian ichter, Krzysztof Marcin Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael S Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, and Pete Florence. 2023. Socratic models: Composing zero-shot multimodal reasoning with language. In *International Conference on Learning Representations*.

Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. 2022. Tip-adapter: Training-free adaption of CLIP for few-shot classification. In *European Conference on Computer Vision*, volume 13695, pages 493–510. Springer.

Ziqi Zhang, Zhongang Qi, Chunfeng Yuan, Ying Shan, Bing Li, Ying Deng, and Weiming Hu. 2021. Open-book video captioning with retrieve-copy-generate network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9837–9846.

Yucheng Zhou, Wei Tao, and Wenqiang Zhang. 2021. Triple sequence generative adversarial nets for unsupervised image captioning. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7598–7602. IEEE.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Please see Section Limitations

☑ A2. Did you discuss any potential risks of your work?
Please see Section Ethics Statement

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Please see Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Please see Section 3.1

☑ B1. Did you cite the creators of artifacts you used?
Please see Section 3.1

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Please see Section 3.1

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Please see Section 3.1

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
We carry out a detailed anonymization process. We manually examine the data of widely adopted benchmark datasets. If there exists information that names individual people, we replace it with expressions like "he", "she", and "a person".

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Please see Table 1

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Please see Table 1

C ☑ Did you run computational experiments?

Please see Sections 3 and 4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Please see Section 3.1

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Please see Section 3.1

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Please see Section 4.1

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Please see Sections 2.3 and 3.1

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.