From Association to Generation: Text-only Captioning by Unsupervised Cross-modal Mapping

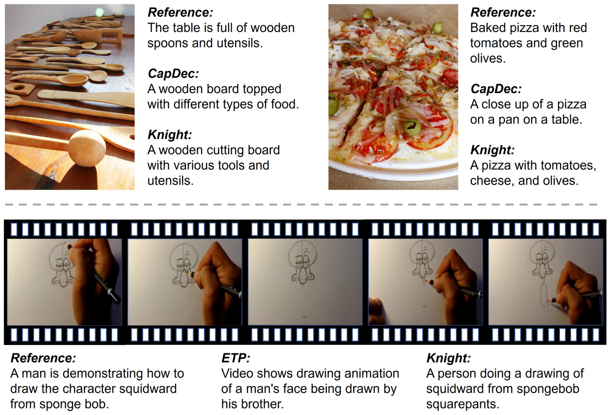
Junyang Wang[1*] , **Ming Yan**[2] , **Yi Zhang**[1] , **Jitao Sang**[1,2†]
[1]School of Computer and Information Technology & Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University
[2]Peng Cheng Lab [3]DAMO Academy, Alibaba Group
{junyangwang, yi.zhang, jtsang}@bjtu.edu.cn, ym119608@alibaba-inc.com

Abstract

With the development of Vision-Language Pre-training Models (VLPMs) represented by CLIP and ALIGN, significant breakthroughs have been achieved for association-based visual tasks such as image classification and image-text retrieval by the zero-shot capability of CLIP without fine-tuning. However, CLIP is hard to apply to generation-based tasks. This is due to the lack of decoder architecture and pre-training tasks for generation. Although previous works have created generation capacity for CLIP through additional language models, a modality gap between the CLIP representations of different modalities and the inability of CLIP to model the offset of this gap, which fails the concept to transfer across modalities. To solve the problem, we try to map images/videos to the language modality and generate captions from the language modality. In this paper, we propose the **K**-**n**earest-ne**igh**bor Cross-modali**t**y Mapping (Knight), a zero-shot method from association to generation. With text-only unsupervised training, Knight achieves state-of-the-art performance in zero-shot methods for image captioning and video captioning.

1 Introduction

In the development of multi-modal learning, two recursive levels arise: (1) multi-modal association; (2) cross-modal generation. The former relies on multi-modal inputs and calculates the association scores for given multi-modal inputs through association expressions. Typical tasks include image classification, image-text retrieval, object detection, etc. The latter is to convert the input from one modality to other modalities, which requires a cross-modal transformation relationship to ensure that the same concept within different modalities can be represented accurately. Typical tasks include image-to-text and text-to-image generation.



Reference:
The table is full of wooden spoons and utensils.

CapDec:
A wooden board topped with different types of food.

Knight:
A wooden cutting board with various tools and utensils.

Reference:
Baked pizza with red tomatoes and green olives.

CapDec:
A close up of a pizza on a pan on a table.

Knight:
A pizza with tomatoes, cheese, and olives.

Reference:
A man is demonstrating how to draw the character squidward from sponge bob.

ETP:
Video shows drawing animation of a man's face being drawn by his brother.

Knight:
A person doing a drawing of squidward from spongebob squarepants.

Figure 1: The examples of *Knight* compared with current state-of-the-art text-only captioning methods.

Currently, Vision-Language Pre-training Models (VLPMs) represented by CLIP [Radford *et al.*, 2021] and ALIGN [Jia *et al.*, 2021] have been successful at the first level by advantage of the multi-modal association-based pre-training task: contrastive learning. With 400 million massive (image, text) training data, CLIP successfully models the association between vision modality and language modality. Benefiting from the diverse web data, CLIP has an extensive perception of open-world knowledge. Without fine-tuning, CLIP achieves an accuracy of 76.2% on ImageNet in the zero-shot setting and rivals or even surpasses the fine-tuning model on datasets with multiple domains [Radford *et al.*, 2021]. Many multi-modal association-based works have benefited from the powerful zero-shot capability of CLIP: not only avoids the high collection cost of supervised data but also simplifies the deployment process.

The great success of CLIP at the association level has sparked the exploration at the generation level. However, since CLIP does not have a decoder architecture and a pre-training task for generation, it is not competent for the generation-based task. Nevertheless, the dominant performance of large-scale language models such as BERT [Devlin *et al.*, 2018] and GPT [Radford *et al.*, 2019] makes it possible to decode from the embedding space of CLIP. It is based on this arising the idea of zero-shot generation based on joint space: CLIP encodes pairs of image and text close enough in embedding space [Su *et al.*, 2022; Wang *et al.*, 2022;

[*] Work done during internship at DAMO Academy, Alibaba Group.

[†] Corresponding author

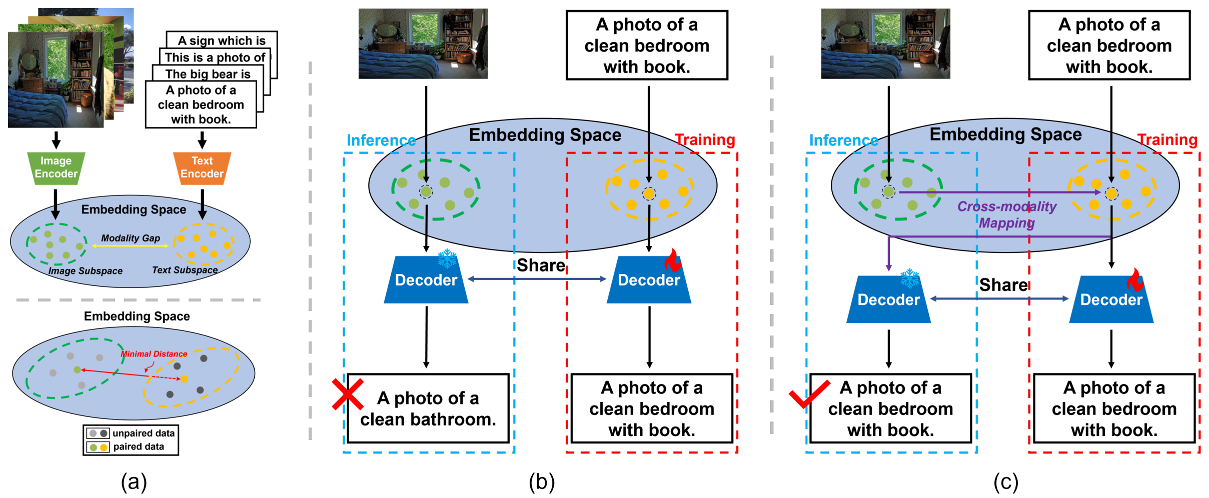


Figure 2: (a) The embedding specificity and cross-modal interaction mode of CLIP. (b) The previous text-only captioning methods based on joint space: in the training phase and the inference phase, the decoder acts in the text and image subspaces, respectively. The modality gap makes the decoder much less effective. (c) Our text-only captioning method *Knight*. Benefiting from cross-modal mapping (in purple arrow), *Knight* alleviates the impact of the modality gap, where the decoder only acts in the text subspaces.

Nukrai *et al.*, 2022]. However, [Liang *et al.*, 2022] has shown that CLIP encodes images and text into two separate subspaces and there is a significant gap between them as shown in Figure 2 (a) top. This means that the decoder is only effective in one modality. When transferring the decoder from one modality to another modality, the modality gap leads to the failure to accurately understand representations as shown in Figure 2 (b).

To eliminate the impact of the modality gap, a major problem is how to establish the transformation relationship between the two modalities. A natural idea is to model the relationship through a large amount of supervised data. However, this requires significant supervised data and training resources [Ramesh *et al.*, 2022]. We argue that this relationship can be established by an unsupervised method through the association capability of CLIP. Based on this, we propose the **K**-**n**earest-ne**igh**bor Cross-modali**t**y Mapping (Knight), a text-only captioning method as shown in Figure 3. First, we collect the captions from the image-text and video-text datasets as the corpus. In the training phase, we first select captions from the corpus for training and use the CLIP similarity to retrieve the k-nearest-neighbor captions that are most similar to the training captions. Then, we use the CLIP features of the training captions to train the decoder by an autoregression loss. In the inference phase, Knight can be applied to both image and video captioning. For image captioning, we retrieve the k-nearest-neighbor captions that are most similar to the inference image. For video captioning, we average the retrieved results for each keyframe to achieve multi-frame input for the video. Knight makes the decoder only act in the text subspace, thus eliminating the effect of the modality gap as shown in Figure 2 (c).

We summarize the contributions as follows:

- We propose a text-only captioning method called Knight based on the unsupervised cross-modal mapping. The method achieves the representation mapping from vision modality to language modality, thus greatly alleviating the impact of the modality gap on cross-modal generation.
- We compare Knight with the other current zero-shot image and video captioning baselines. Experimental results show that Knight achieves state-of-the-art performance. We explore the possibility of employing CLIP association capacity to address the zero-shot generation-based tasks.

2 Background and Related Work

2.1 CLIP

By contrastive learning on a dataset of 400M (image, text) pairs, CLIP [Radford *et al.*, 2021] models the association between the images and text. CLIP is widely used for zero-shot tasks such as classification, retrieval, etc [Radford *et al.*, 2021]. The zero-shot performance is claimed to be close to or even better than fine-tuned models [Radford *et al.*, 2021]. Many works have applied the zero-shot capacity of CLIP to specific application scenarios such as image segmentation [Xu *et al.*, 2021], image generation [Ramesh *et al.*, 2022], and object detection [Zhong *et al.*, 2022].

2.2 Text-only Captioning

Mainstream methods for captioning tasks are mainly divided into two types: (1) extracting visual features typically using a pre-trained network and training a decoder that produces the final captions [Chen and Zitnick, 2014; Chen *et al.*, 2017; Yang *et al.*, 2019; Anderson *et al.*, 2018; Luo *et al.*, 2021]; (2) bridging the gap between vision and language by employing pre-training to create a shared latent space of images and text [Lu *et al.*, 2019; Laina *et al.*, 2019; Tan and Bansal, 2019; Li *et al.*, 2020; Zhou *et al.*, 2020; Zhang *et al.*, 2021; Hu *et al.*, 2022]. With the rise of CLIP, recent captioning methods use CLIP for reducing training time [Mokady

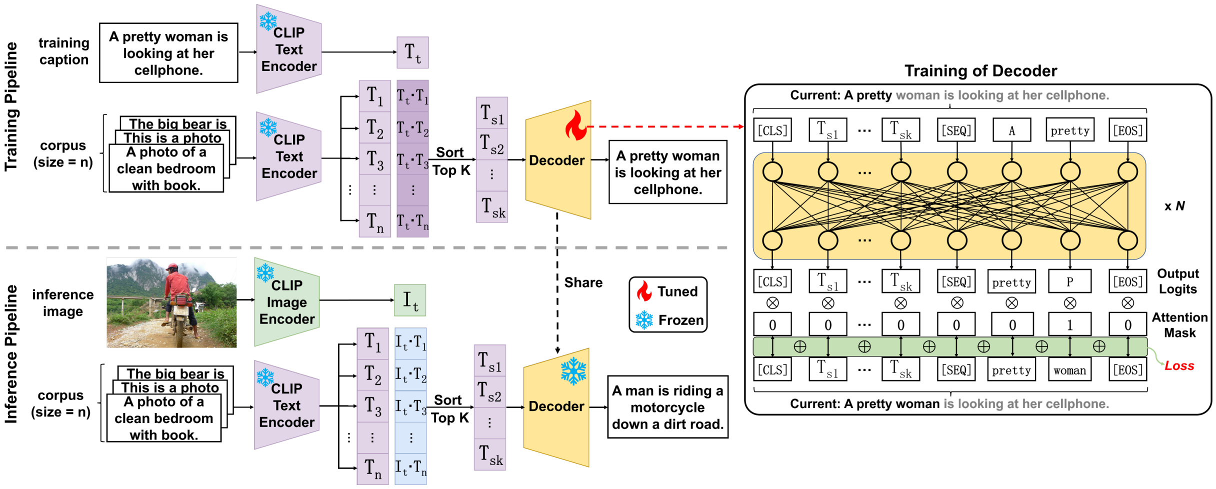


Figure 3: The overview of *Knight*. In the training phase, we first calculate the CLIP similarity to obtain k-nearest-neighbor captions from the corpus that are most similar to the training caption. Then, we feed the representations to the decoder for autoregressive training. The training process requires no image or video participation, but only an unsupervised corpus. In the inference phase, we replace the training caption with the inference image and repeat the above steps.

et al., 2021] and improved captions [Shen *et al.*, 2021; Cornia *et al.*, 2021; Kuo and Kira, 2022].

However, all of the previous works require extensive training and large paired data that are hard to collect. To address this, [Gan *et al.*, 2017] and [Zhao *et al.*, 2020] have suggested style-guided captioning, but also employ training over paired data. OpenAI has conducted remarkable work in learning cross-modal patterns with the CLIP trained by a large number of resources [Radford *et al.*, 2021]. This means that the effective exploitation of the pre-trained knowledge of CLIP can be free from the constraints of supervised data. [Tewel *et al.*, 2022b] attempted to generate a caption with the highest CLIP similarity to a given image using a pre-trained language model. However, the pre-trained language model and CLIP's text encoder use different pretraining data and paradigms, making it difficult for the language model to generate high-quality captions that are relevant to images. Researchers realized that additional language data are needed for aligning the embedding space between the language model and CLIP, hence the rise of text-only captioning methods. [Su *et al.*, 2022] made the language model fit the domain of CLIP by fine-tuning it with unsupervised corpus. However, the above works require the language model to provide candidate words and thus prove to have a severe language prior [Wang *et al.*, 2022]. [Wang *et al.*, 2022; Nukrai *et al.*, 2022] proposed the idea of joint space that argues that CLIP encodes pairs of image and text close enough in embedding space, thus overcoming the problem of language prior by training the decoder in the language modality and transferring it to vision modality in inference phase.

2.3 Addressing the Modality Gap

The previous text-only methods argued that embedded text is relatively close to its corresponding visual embedding [Nukrai *et al.*, 2022]. However, [Liang *et al.*, 2022] demonstrates that images and text are embedded into two subspaces separately and a significant modality gap is between them as shown in Figure 2 (a) top. This gap limits the generation quality of previous methods. Although it is possible to learn the pattern between two modalities with a large amount of data, [Ramesh *et al.*, 2022] demonstrates that the overhead of this process is huge. Nevertheless, we note that the effective cross-modal interaction of CLIP is under the association of paired data as shown in Figure 2 (a) bottom. This makes us realize that although the modality gap makes it hard to achieve cross-modal generation directly, it can be achieved by association indirectly.

3 Method

3.1 Preliminaries

Notations. We first explain the definition of text-only captioning method and the requirements for the training data. The supervised dataset $\mathcal{D}_s = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ consisting of n pairs with images or videos x_i and reference captions $y_i = \{c_i^1, \ldots, c_i^{|y_i|}\}$, where y_i is a set of the captions that describe the x_i from different perspectives, and c_i^j denotes the j_{th} caption of y_i. The unsupervised data include unlabeled image or video dataset $\mathcal{D}_u^I = \{x_1, \ldots, x_i\}$ and text datasets $\mathcal{D}_u^T = \{y_1, \ldots, y_j\}$. Traditional captioning methods use supervised dataset $\mathcal{D}_s$ for training, while text-only captioning methods only assume the availability of unlabeled dataset $\mathcal{D}_u^T$.

CLIP is a VLPM with dual-encoder architecture. It consists of two independent encoders for vision and language modalities. Similarities between vision and language representations on large-scale image-text pairs are used to pre-train CLIP, bridging the gap between vision-language semantics in the embedding space of CLIP. The similarity is calculated as

$$\begin{gathered} I = f_{\mathrm{V}}(x) \\ T = f_{\mathrm{L}}(y) \\ \mathrm{Sim}(I, T) = \cos < I, T >= \frac{I}{|I|} \cdot \frac{T}{|T|} \end{gathered} \tag{1}$$

Method	Flickr30k						MS-COCO						**Training Params**
	B@1	B@4	M	R-L	CIDEr	SPICE	B@1	B@4	M	R-L	CIDEr	SPICE	
training-free													
CLIPRe	38.5	5.2	11.6	27.6	10.0	5.7	39.5	4.9	11.4	29.0	13.6	5.3	-
ZeroCap	44.7	5.4	11.8	27.3	16.8	6.2	49.8	7.0	15.4	31.8	34.5	9.2	-
SMs	-	-	-	-	-	-	-	6.9	15.0	34.1	44.5	10.1	-
text-only training													
MAGIC	44.5	6.4	13.1	31.6	20.4	7.1	56.8	12.9	17.4	39.9	49.3	11.3	345M
CLMs	58.3	16.8	16.2	39.6	22.5	9.8	59.3	15.0	18.7	41.8	55.7	10.9	345M
CapDec	55.5	17.7	20.0	43.9	39.1	9.9	69.2	26.4	25.1	51.8	91.8	11.9	919M
Knight (Ours)	**64.0**	**22.6**	**24.0**	**48.0**	**56.3**	**16.3**	**71.7**	**27.8**	**26.4**	**52.3**	**98.9**	**19.6**	771M

Table 1: Image captioning results of different methods on Flickr30k and MS-COCO, where the B@1, B@4, M, and R-L represent BLEU@1, BLEU@4, METEOR, and Rouge-L respectively.

Method	MS-COCO ⟹ Flickr30k						Flickr30k ⟹ MS-COCO					
	B@1	B@4	M	R-L	CIDEr	SPICE	B@1	B@4	M	R-L	CIDEr	SPICE
CLIPRe	38.7	4.4	9.6	27.2	5.9	4.2	31.1	3.0	9.9	22.8	8.5	3.9
MAGIC	46.4	6.2	12.2	31.3	17.5	5.9	41.4	5.2	12.5	30.7	18.3	5.7
CLMs	49.2	10.1	12.5	33.8	12.7	5.7	47.6	7.7	14.9	35.9	38.5	8.2
CapDec	60.2	17.3	18.6	42.7	35.7	7.2	43.3	9.2	16.3	36.7	27.3	10.4
Knight (Ours)	**66.0**	**21.1**	**22.0**	**46.3**	**48.9**	**14.2**	**62.1**	**19.0**	**22.8**	**45.8**	**64.4**	**15.1**

Table 2: Cross-Domain Evaluation. X ⟹ Y means source domain ⟹ target domain.

where f_V and f_L are the image encoder and text encoder of CLIP respectively.

Unsupervised Language Modelling is a method for self-supervised training on the unsupervised corpus. It learns the relationship of sentence context by the next-token prediction. The pre-training loss is the Maximum Likelihood Estimation (MLE) that calculated as

$$\mathcal{L}_{\text{MLE}} = -\frac{1}{|T|}\sum_{i=1}^{|T|}\log M_\theta(T_i|T_1T_2\ldots T_{i-1}) \tag{2}$$

where θ denotes the parameter that needs to be optimized for the model M.

3.2 Knight

We consider a training caption $y^{'}$ from $\mathcal{D}_u^T$. The T_t and $T_{1\sim n}$ calculated as

$$\begin{gathered} T_t = f_{\text{L}}(y^{'}) \\ T_{1\sim n} = \{f_{\text{L}}(y_1),\ldots,f_{\text{L}}(y_n)\} = \{T_1,\ldots,T_n\} \end{gathered} \tag{3}$$

where n is the size of training corpus. Then, we get the CLIP similarities S between T_t and $T_{1\sim n}$ calculated as

$$S = \{\text{Sim}(T_t,T_1),\ldots,\text{Sim}(T_t,T_n)\} \tag{4}$$

Then we sort the S from large to small and get the $T^{'}$ = { T_{s1}, …, T_{sk} }, where T_{si} is the i-th largest of the sorted result and k is a hyperparameter. Finally, we use the $T^{'}$ for autoregression training as following equation

$$\mathcal{L}_{\text{MLE}} = -\frac{1}{|T'|}\sum_{i=1}^{|T^{'}|}\log M_\theta(T_i|T_{s1}\ldots T_{sk},T_1\ldots T_{i-1}) \tag{5}$$

Image Captioning

After training, we fix the parameters of the decoder. In the inference phase, we replace the input of the training caption with the inference image x and get the I_t = $f_{\text{V}}(x)$. We get the S as the following equation

$$S = \{\text{Sim}(\boldsymbol{I_t},T_1),\ldots,\text{Sim}(\boldsymbol{I_t},T_n)\} \tag{6}$$

Finally, we follow the training process to obtain $T^{'}$ and feed it into the decoder to get the generated captions as the equation (5).

Video Captioning

The above process of image captioning can be understood as a process of filtering and regrouping the information from k-nearest-neighbor captions. This process is similar to the generation of video caption: the decoder establishes the connection between the preceding and following frames and reasons out the combined information of these frames. Therefore, we argue that Knight can be transferred to video captioning.

Compared with the image, video is a set of multiple frames. We first extract the keyframes x = { x_1, …, x_m } of the inference video x, where m is the number of keyframes. Then, we get the I_t^i = $f_{\text{V}}(x_i)$, where x_i is the i-th keyframe. For each keyframe, we get the S_i as the following equation

$$S_i = \{\text{Sim}(\boldsymbol{I_t^i},T_1),\ldots,\text{Sim}(\boldsymbol{I_t^i},T_n)\} \tag{7}$$

We sort each S_i and select the k representations with the greatest similarity thus obtaining $T_i^{'}$. Finally, we get the $T^{'}$ for video as the following equation

$$T^{'} = \{T_{s1},\ldots,T_{sm}\} = \{mean(T_1^{'}),\ldots,mean(T_m^{'})\} \tag{8}$$

Finally, we feed $T^{'}$ into the decoder to generate the captions as the equation (5).

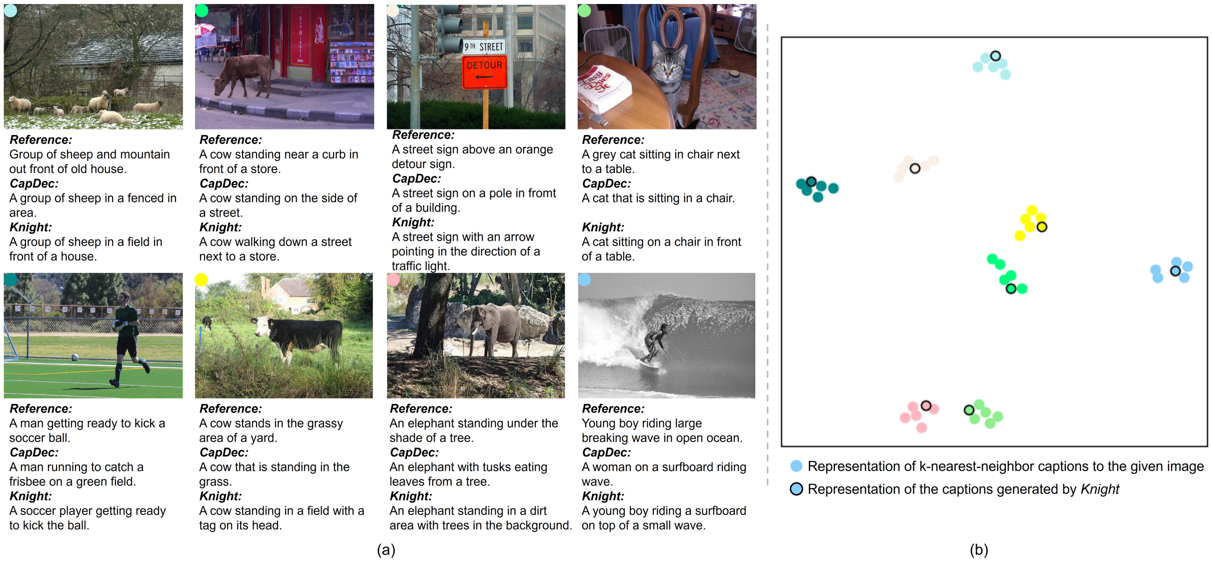


Figure 4: (a): Examples of image captioning generated by *Knight* compared with reference captions. (b): The t-SNE results for the examples in (a). We set k to 5. We take the caption generated by the model back to CLIP and encode it to get the embedding representation of that caption. The solid-colored dots represent the embedding representations of the 5 nearest captions to the given image, and the black-circle dots represent the embedding representations of the captions generated by the model.

4 Experiment

4.1 Experimental Settings

Evaluation Benchmarks. For the image captioning task, we conduct experiments on two widely used benchmarks: Flickr30k [Plummer *et al.*, 2015] and MS-COCO [Lin *et al.*, 2014; Chen *et al.*, 2015]. And we set up the training, validation, and test splits according to the protocols provided by [Karpathy and Fei-Fei, 2015] for both datasets. For the video captioning task, we choose two video datasets: MSR-VTT [Xu *et al.*, 2016] and MSVD [Wu *et al.*, 2017]. We use the captions in the training set as the training corpus and evaluate the methods on the test set, as in other text-only captioning works.

Implementation Details. For CLIP, we choose the Resnet50x64 architecture which encodes each image as a 1024-dimension vector. For the decoder, we choose the large vision of GPT-2 [Radford *et al.*, 2019] with a 1280-dimension embedding space. To align CLIP and decoder on the representation layer, we use a 3-layer MLP that transforms the representation of CLIP into 1280 dimensions. We optimize the decoder with the Adam optimizer [Kingma and Ba, 2014] and a learning rate of 1e-6. The training process is less than 6 hours with 1 Tesla A100 GPU. In the inference phase, we use beam search as same as other methods. For the acquiring of video keyframes, we choose the isometric sampling.

Baselines. For image captioning, we include several zero-shot methods as our baselines. First, we compare with a training-free method, called CLIPRe [Su *et al.*, 2022]. Given an image, it retrieves the most related caption from the training corpus based on the image-text similarity as measured by CLIP. Then, we compare two training-free methods with the language model ZeroCap [Tewel *et al.*, 2022b] and SMs [Zeng *et al.*, 2022]. Finally, we compare with text-only methods MAGIC [Su *et al.*, 2022], CLMs [Wang *et al.*, 2022], and current state-of-the-art method CapDec [Nukrai *et al.*, 2022]. For video captioning, we compare it with the state-of-the-art method EPT [Tewel *et al.*, 2022a]. EPT is a zero-shot video captioning method based on evolving pseudo-tokens. And we also adapt the baselines of image captioning to video captioning by referring to [Tewel *et al.*, 2022a]. These methods use the average features of video keyframes as input.

Evaluation Metrics. Following the common practice in the literature, we perform an evaluation using BLEU-1 (B@1), BLEU-4 (B@4) [Papineni *et al.*, 2002], METEOR (M) [Denkowski and Lavie, 2014], ROUGE-L (R-L) [Lin and Och, 2004], CIDEr [Vedantam *et al.*, 2015], and SPICE [Anderson *et al.*, 2016].

4.2 Performance Comparison

Image Captioning

The results of image captioning are shown in Table 1. We see that Knight achieves the best performance in all 12 evaluation metrics. It is worth noting that compared to the current state-of-the-art method CapDec, Knight achieves significant performance improvement using fewer training parameters. This is because CapDec is based on joint space and alleviates the impact of the modality gap by noise interference. However, it is not reasonable to model the modality gap by random Gaussian noise because the offset direction of the modality gap is directional, while random Gaussian noise is non-directional. Compared with CapDec, Knight utilizes the similarity mapping that is best suited for CLIP. The direction of this mapping is consistent with the offset direction of the modality gap.

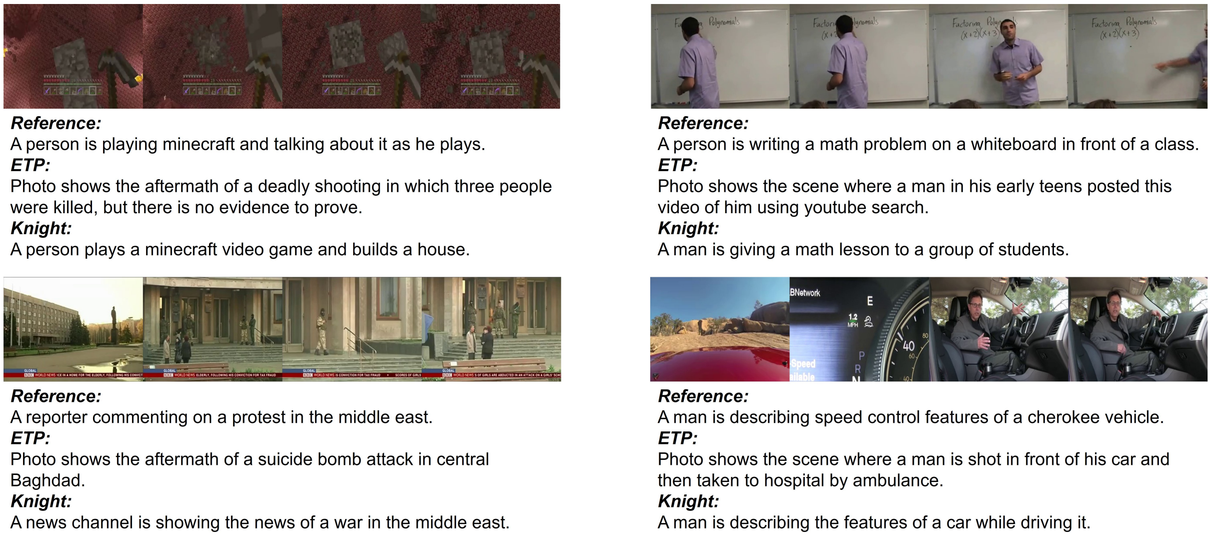
To test the generalization ability of Knight, we conduct a cross-domain experiment. Specifically, we apply the text-

Method	MSR-VTT						MSVD					
	B@1	B@4	M	R-L	CIDEr	SPICE	B@1	B@4	M	R-L	CIDEr	SPICE
full-supervised training												
VNS-GRU	-	45.3	29.9	64.7	53.0	-	-	66.5	42.1	79.7	121.5	-
SemSynAN	-	46.4	30.4	46.7	51.9	-	-	64.4	41.9	79.5	111.5	-
unsupervised training												
ZeroCap*†	-	2.3	12.9	30.4	5.8	-	-	2.9	16.3	35.4	9.6	-
MAGIC*	22.3	5.5	13.3	35.4	7.4	4.2	24.7	6.6	16.1	40.1	14.0	2.9
CLMs*	25.7	6.2	17.8	15.7	10.1	6.5	26.9	7.0	16.4	44.3	20.0	3.1
CapDec*	30.2	8.9	23.7	17.2	11.5	5.9	33.1	7.9	23.3	25.2	34.5	3.2
ETP†	-	3.0	14.6	22.7	11.3	-	-	3.0	17.8	31.4	17.4	-
Knight (Ours)	**72.6**	**25.4**	**28.0**	**50.7**	**31.9**	**8.5**	**73.1**	**37.7**	**36.1**	**66.0**	**63.8**	**5.0**

*The method is adapted from zero-shot image captioning to zero-shot video captioning.

†The method is training-free.

Table 3: Video captioning results of different methods on MSR-VTT and MSVD.



Reference:
A person is playing minecraft and talking about it as he plays.
ETP:
Photo shows the aftermath of a deadly shooting in which three people were killed, but there is no evidence to prove.
Knight:
A person plays a minecraft video game and builds a house.

Reference:
A person is writing a math problem on a whiteboard in front of a class.
ETP:
Photo shows the scene where a man in his early teens posted this video of him using youtube search.
Knight:
A man is giving a math lesson to a group of students.

Reference:
A reporter commenting on a protest in the middle east.
ETP:
Photo shows the aftermath of a suicide bomb attack in central Baghdad.
Knight:
A news channel is showing the news of a war in the middle east.

Reference:
A man is describing speed control features of a cherokee vehicle.
ETP:
Photo shows the scene where a man is shot in front of his car and then taken to hospital by ambulance.
Knight:
A man is describing the features of a car while driving it.

Figure 5: Examples of video captioning generated by *Knight* with reference captions.

only training on the corpus of the source domain (e.g., MS-COCO) to perform inference on the test set of the target domain (e.g., Flickr30k). The results of the cross-domain experiment are shown in Table 2. From the results, we can see that Knight has a significant advantage in generalization ability.

We show some examples generated by Knight compared with CapDec in Figure 4 (a). We see that the caption generated by Knight corresponds well to the content in the image, both in the foreground and in the background. We understand the reason for the high-quality generation of Knight by visualizing the results in Figure 4 (b). We reduce the embedding space representations of the k-nearest-neighbor captions to a 2-dimensional representation by t-SNE (solid-colored dots), while bringing the generated caption back to the embedding space to obtain the corresponding representation (black-circle dots). We can see that the k-nearest-neighbor captions of different images are significantly distinguished on the embedding space, which indicates that Knight has a good discriminatory ability for the different concepts.

Video Captioning

The results of video captioning are shown in Table 3. From the results of the methods adapted from image captioning, we see that these methods do not apply to video captioning. This is because the input to these methods is a representation of only a single image or text. When multiple keyframes from a video are used as input, these methods need to fuse the representations of multiple frames into a single representation. This not only causes the loss of features, but also fails to model the relationship between different keyframes, and thus fails to inference about the behavior of the whole video.

From the results of ETP, we see that although it is a method designed for video captioning, it still does not work well. This is for the same reason that ZeroCap in image captioning fails to generate high-quality captions: the CLIP used for matching computation cannot be aligned with the language model used for a generation without text training. Finally, although there are still gaps in Knight compared to the full-supervised method, Knight achieves the best in all 12 evaluation met-

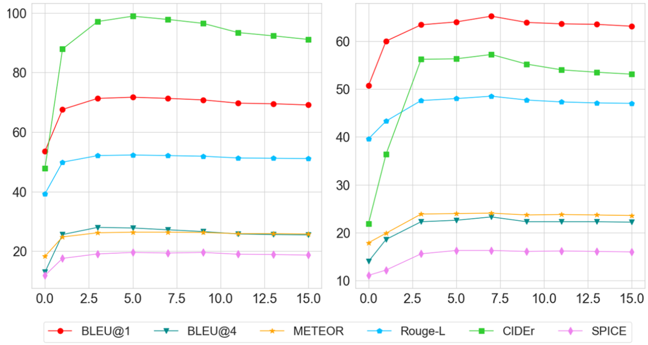


Figure 6: The image captioning results of *Knight* under different k value settings on MS-COCO (left) and Flickr30K (right), where the horizontal coordinate is the value of k and the vertical coordinate is the value of each metric.

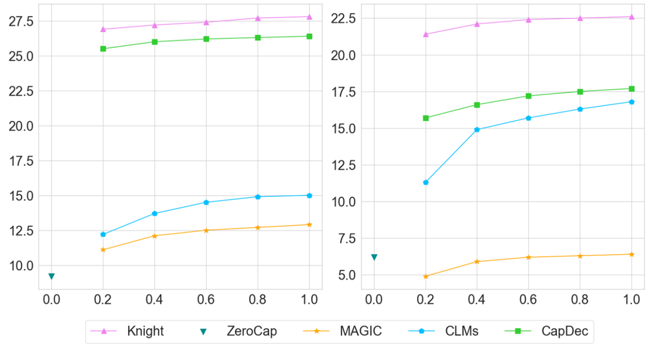


Figure 7: The results of B@4 of *Knight* with different proportions of training corpus compared with other methods on MS-COCO (left) and Flickr (right), where the horizontal coordinate is the proportions of training corpus and the vertical coordinate is the value of B@4.

rics in the unsupervised method and is far ahead of the other methods.

In the example of Figure 1, although Knight does not use the image of *Squidward* for training, it successfully maps to *Squidward* by unsupervised cross-modal mapping of CLIP. In Figure 5, we show more examples. It is clear from the results that Knight has a strong capacity to recognize the concepts in open world. For example, the game name *Minecraft* in the first example, *math* in the second example, and the *Middle East* in the third example.

4.3 Discussions

A. Are the results of Knight due to the captions in the training corpus that are similar to the inference images or videos?

We compare CLIPRe (Table 1) to analyze the issue. Given an image, CLIPRe retrieves the most related caption from the training corpus based on the image-text similarity as measured by CLIP. From the result, we see that a better result cannot be achieved by only retrieving from the corpus, although the Flickr30K and MS-COCO are large enough. This indicates that our method is not a retrieval method but a generated method.

B. Does irrelevant content in the k-nearest-neighbor captions of an image impact the generation quality?

From the result in Figure 4 (b), we can see that although the generated caption is strongly associated with the k-nearest-neighbor captions of the image, there is still a distinction between the two. This indicates that the decoder is not overly dependent on a caption and generates according to its feature, but is a reorganization of the information in the k-nearest-neighbor captions. Thus even if there is irrelevant information in one of the k-nearest-neighbor captions, the decoder does not rely on this information completely.

C. What effect does the choice of k have on the performance of Knight?

We set different values of k to explore this issue. The results on image captioning are shown in Figure 6 and the results on video captioning are shown in appendix. It is worth noting that k equals 0 means that the representation of the inference image is used to generate directly. First, the performance at k of 0 is not good. This confirms the existence of the modality gap. Then, when k is 1, i.e., the nearest-neighbor mapping, the performance is still not optimal. Finally, as k increases, the performance of Knight no longer changes significantly after rising to a certain level. Although the value of k increases, there is a limited number of captions associated with inference images in the corpus. An excessively large k maybe introduce the noise.

D. Does Knight show a significant performance decline when the size of the training corpus is reduced?

We set different sizes of the training corpus of Knight to explore this issue. The results of B@4 are shown in Figure 7 and the results of other metrics are shown in the appendix. We observe that the performance of Knight does decline when the size of the training corpus is reduced. However, it is worth noting that Knight does not suffer from a catastrophic performance drop even with just 10% of the training corpus. Compared to other methods, Knight still has a significant improvement.

5 Conclusion

In this paper, we explore how to achieve generation by utilizing the zero-shot capability of CLIP. To alleviate the negative impact of the modality gap on the text-only captioning method, we propose the association-to-generation method, Knight. Unlike previous works, in the inference phase, we represent a given image as k-nearest-neighbor captions by computing the similarity between the inference image or video and the captions in the corpus. This unifies the decoding range in the training and inference phases and makes the generated captions independent of the modality gap. Experimental results show that Knight achieves state-of-the-art performances on both text-only image and video captioning.

Acknowledgments

This work is supported by the Beijing Natural Science Foundation (No.JQ20023), National Natural Science Foundation of China (Grant No. 61832002, 62172094), and CCF-Zhipu AI Large Model Fund.

References

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

[Anderson *et al.*, 2018] 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.

[Chen and Zitnick, 2014] Xinlei Chen and C Lawrence Zitnick. Learning a recurrent visual representation for image caption generation. *arXiv preprint arXiv:1411.5654*, 2014.

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

[Chen *et al.*, 2017] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5659–5667, 2017.

[Cornia *et al.*, 2021] Marcella Cornia, Lorenzo Baraldi, Giuseppe Fiameni, and Rita Cucchiara. Universal captioner: long-tail vision-and-language model training through content-style separation. *arXiv preprint arXiv:2111.12727*, 2021.

[Denkowski and Lavie, 2014] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380, 2014.

[Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[Gan *et al.*, 2017] Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng. Stylenet: Generating attractive visual captions with styles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3137–3146, 2017.

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

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

[Karpathy and Fei-Fei, 2015] 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.

[Kingma and Ba, 2014] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[Kuo and Kira, 2022] Chia-Wen Kuo and Zsolt Kira. Beyond a pre-trained object detector: Cross-modal textual and visual context for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17969–17979, 2022.

[Laina *et al.*, 2019] 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.

[Li *et al.*, 2020] 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 *European Conference on Computer Vision*, pages 121–137. Springer, 2020.

[Liang *et al.*, 2022] Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *arXiv preprint arXiv:2203.02053*, 2022.

[Lin and Och, 2004] Chin-Yew Lin and Franz Josef Och. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 605–612, 2004.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

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

[Luo *et al.*, 2021] 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.

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

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

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

[Plummer *et al.*, 2015] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015.

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

[Radford *et al.*, 2021] 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.

[Ramesh *et al.*, 2022] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[Shen *et al.*, 2021] 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.

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

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

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

[Tewel *et al.*, 2022b] Yoad Tewel, Yoav Shalev, Idan Schwartz, and Lior Wolf. 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, 2022.

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

[Wang *et al.*, 2022] Junyang Wang, Yi Zhang, Ming Yan, Ji Zhang, and Jitao Sang. Zero-shot image captioning by anchor-augmented vision-language space alignment. *arXiv preprint arXiv:2211.07275*, 2022.

[Wu *et al.*, 2017] Zuxuan Wu, Ting Yao, Yanwei Fu, and Yu-Gang Jiang. Deep learning for video classification and captioning. In *Frontiers of multimedia research*, pages 3–29. 2017.

[Xu *et al.*, 2016] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016.

[Xu *et al.*, 2021] Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. A simple baseline for zero-shot semantic segmentation with pre-trained vision-language model. *arXiv preprint arXiv:2112.14757*, 2021.

[Yang *et al.*, 2019] Xu Yang, Hanwang Zhang, and Jianfei Cai. Learning to collocate neural modules for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4250–4260, 2019.

[Zeng *et al.*, 2022] Andy Zeng, Adrian Wong, Stefan Welker, Krzysztof Choromanski, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, et al. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598*, 2022.

[Zhang *et al.*, 2021] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2021.

[Zhao *et al.*, 2020] Wentian Zhao, Xinxiao Wu, and Xiaoxun Zhang. Memcap: Memorizing style knowledge for image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12984–12992, 2020.

[Zhong *et al.*, 2022] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16793–16803, 2022.

[Zhou *et al.*, 2020] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13041–13049, 2020.