Improving Cross-Modal Alignment with Synthetic Pairs for Text-Only Image Captioning

Zhiyue Liu[1,2*], **Jinyuan Liu**[1], **Fanrong Ma**[1]

[1]School of Computer, Electronics and Information, Guangxi University, Nanning, China
[2]Guangxi Key Laboratory of Multimedia Communications and Network Technology
liuzhy@gxu.edu.cn, {2213394017, 2213301037}@st.gxu.edu.cn

Abstract

Although image captioning models have made significant advancements in recent years, the majority of them heavily depend on high-quality datasets containing paired images and texts which are costly to acquire. Previous works leverage the CLIP's cross-modal association ability for image captioning, relying solely on textual information under unsupervised settings. However, not only does a modality gap exist between CLIP text and image features, but a discrepancy also arises between training and inference due to the unavailability of real-world images, which hinders the cross-modal alignment in text-only captioning. This paper proposes a novel method to address these issues by incorporating synthetic image-text pairs. A pre-trained text-to-image model is deployed to obtain images that correspond to textual data, and the pseudo features of generated images are optimized toward the real ones in the CLIP embedding space. Furthermore, textual information is gathered to represent image features, resulting in the image features with various semantics and the bridged modality gap. To unify training and inference, synthetic image features would serve as the training prefix for the language decoder, while real images are used for inference. Additionally, salient objects in images are detected as assistance to enhance the learning of modality alignment. Experimental results demonstrate that our method obtains the state-of-the-art performance on benchmark datasets.

Introduction

Image captioning is an appealing research task, aiming to produce descriptive textual content for provided images. The major challenge in this task involves comprehending the association between images and texts, which heavily relies on paired datasets (Stefanini et al. 2022). Current researches primarily concentrate on training models using paired data, leading to substantial enhancements (Wu et al. 2022; Luo et al. 2023; Kuo and Kira 2023), while these models generalize poorly to images in the wild (Wu et al. 2018; Agrawal et al. 2019). Despite pre-training on web-scale noisy paired data and subsequent fine-tuning on human-annotated captioning data have shown effectiveness (Radford et al. 2021; Yu et al. 2022a; Alayrac et al. 2022), pre-trained methods still count on paired data to achieve optimal performance.

The generalization ability is crucial in practical applications, especially in the absence of in-domain image-text paired samples. Unsupervised transferability enables the model to work on new tasks or domains without paired training data, making it a valuable feature for various real-world scenarios.

Unsupervised image captioning, as another line of work, has garnered increasing attention for its ability to generate captions without human-annotated data. By capitalizing on the multimodal embedding space of CLIP (Radford et al. 2021), recent methods adopt text-only training paradigms to develop a language model using CLIP text features as the input. During inference, they decode captions from the given image's features extracted by CLIP. In contrast to other unsupervised methods, text-only image captioning utilizes textual data during training without the accessibility to real-world images (Li et al. 2023). Although previous methods have shown competitive results, they heavily depend on the CLIP's cross-domain capability. The presence of the modality gap (Liang et al. 2022; Jiang et al. 2023) poses a challenge in those text-only methods based on CLIP, as it hinders the direct use of visual embeddings as the source to generate captions. Towards this issue, some works propose solutions by noise injection or feature projection (Nukrai, Mokady, and Globerson 2022; Li et al. 2023), assisting image features from CLIP to better align with text features during inference, while the modality gap still exists. Due to the absence of images, only textual information could be utilized for training while real images for inference. This leads to a discrepancy between training and inference. Thus, the learned alignment cannot work efficiently for real image captioning, hindering existing models from performing well. Besides, overreliance on CLIP would indeed obstruct the acquisition of valuable knowledge from other pre-trained models.

This paper proposes a method by exploring synthetic pairs for text-only image captioning, called SynTIC. With multiple frozen pre-trained models, we mitigate the inconsistency of training and inference, while better aligning texts and images. Firstly, a pre-trained text-to-image model is employed to generate the synthetic image corresponding to each training text. Nonetheless, directly training the captioning model on synthetic image-text pairs results in limited performance due to the dissimilarity between synthetic images and natural images. Specifically, images generated from simple descriptive texts are usually short of the depth and details in

real-world images, indicating these generated images lack content richness. The cross-modal alignment acquired from synthetic data may compromise the model's ability to capture the semantics of real images. Therefore, we employ a generation-then-refinement procedure to synthesize image features. The feature of generated images from CLIP acts as initialization. Then, these features are optimized in the CLIP multimodal embedding space. Since CLIP is pre-trained on real-world pairs with the contrastive constraint, the optimization following this constraint helps synthetic image features embed in the CLIP multimodal space (Radford et al. 2021). This would polish pseudo image features to relatively resemble natural image features. We further refine pseudo features with various textual information. A text supporting set is deployed to project synthetic images as text embeddings. By leveraging the projected vectors, the caption decoder effectively assimilates information from the text corpus. To some extent, the limited semantics of synthetic image features are complemented by textual data, enabling our method to create diverse and contextually grounded descriptions for given images. Based on the above steps, the training and inference of SynTIC would follow a unified procedure that takes images as input and then projects image features into a text space for decoding. Moreover, our method introduces object tags detected in images as assistance to ease the learning of semantic alignments between modalities, since projected vectors in the text space may lack parts of semantics in the origin images. Those objects potentially contain essential information that serves as caption contents (Li et al. 2020). In our method, object features would cooperate with image features to guide the decoder for captioning. SynTIC entirely relies on texts and corresponding synthetic images for training, without accessing any real images.

In summary, our contributions are as follows:

- A text-only captioning method, called SynTIC, is proposed, which leverages synthetic data to improve modality alignment, while unifying training and inference.

- A generation-then-refinement procedure, which includes image feature optimization and projection, is developed to align images with texts. Besides, object tags serve as auxiliary points to enhance captioning performance.

- Experimental evaluations on several benchmark datasets demonstrate that SynTIC enhances text-only training and achieves competitive generation performance.

Related Work

Depending on human-annotated image-text data, advanced methods for supervised captioning mainly work on visual encoding, language modeling, and training algorithms to boost captioning performance (Gao et al. 2019; Xing et al. 2021; Li et al. 2022b; Zeng et al. 2022). Due to the lack of supervised datasets in specific scenes, some works pre-train models on the large-scale web-crawled dataset and then fine-tune on a few downstream human-annotated pairs (Hu et al. 2022; Li et al. 2022a). With the rise of pre-trained multimodal methods (e.g., CLIP), they are deployed to improve captioning performance. Such as Dai et al. (2022) attempt

to directly migrate the knowledge from CLIP to image captioning. Wang et al. (2022) realize text generation by the autoregression pre-training on image and text data. Although these methods get improvements from the pre-training stage, their optimal performance still relies on the fine-tuning stage which requires human-annotated data.

Under the unsupervised setting, the model aims to generate image captions without paired data. Unpaired captioning methods use independent image sources and text corpus for training. They mostly conduct adversarial training or detect objects to establish the connection between images and texts (Ben et al. 2021; Meng et al. 2022; Zhu et al. 2023; Yu et al. 2023). Without training, ZeroCap (Tewel et al. 2022) proposes to integrate CLIP and a language model, where CLIP assumes the role of guiding the language model toward a specific visual direction. MAGIC (Su et al. 2022) develops a framework for incorporating visual controls into the generation process, enabling language models to perform zero-shot captioning. With no need for real-world images, text-only captioning methods have emerged to train a decoder solely on textual data, achieving superior unsupervised performance. They could generate captions based on feature alignment in the CLIP embedding space. Considering that the modality gap (Liang et al. 2022) existing between CLIP text and image features, some works (Nukrai, Mokady, and Globerson 2022; Gu, Clark, and Kembhavi 2023) propose the solution by injecting noise to train a model that maps the text embedding closer to its corresponding image embedding. Besides, Li et al. (2023) introduce a training-free mechanism that leverages a text memory set to project visual embeddings into the text embedding space during inference. Nevertheless, previous methods cannot completely narrow the modality gap. Using texts for training while images for inference still brings a discrepancy, which impedes effective text-image alignment. Moreover, the excessive dependence on the cross-modal capability of CLIP would hinder the acquisition of valuable knowledge from other pre-trained models. In contrast, our method synthesizes image features to unify training and inference following a generation-then-refinement procedure. CLIP would cooperate with the pre-trained text-to-image model in SynTIC, where synthetic features would be refined in the CLIP embedding space for better cross-modal alignment. Objects detected in images help further improve performance.

Methodology

Text-only captioning requires training the model to generate captions for given images when textual input is solely available. Unlike fully supervised methods trained on the dataset $D_{\text{pair}} = \{(i_1, t_1), ..., (i_z, t_z)\}$ including pairs between the image i_i and text t_i, text-only training could only use a text corpus $D_t = \{t_1, ..., t_n\}$ as the source. The trained model would perform caption generation on the real-world images $D_v = \{i_1, ..., i_m\}$ during inference. Achieving cross-modal alignment from only textual data poses a major challenge.

Overview

The overall framework of SynTIC is shown in Fig. 1. Our method aims to improve cross-modal alignment by introduc-
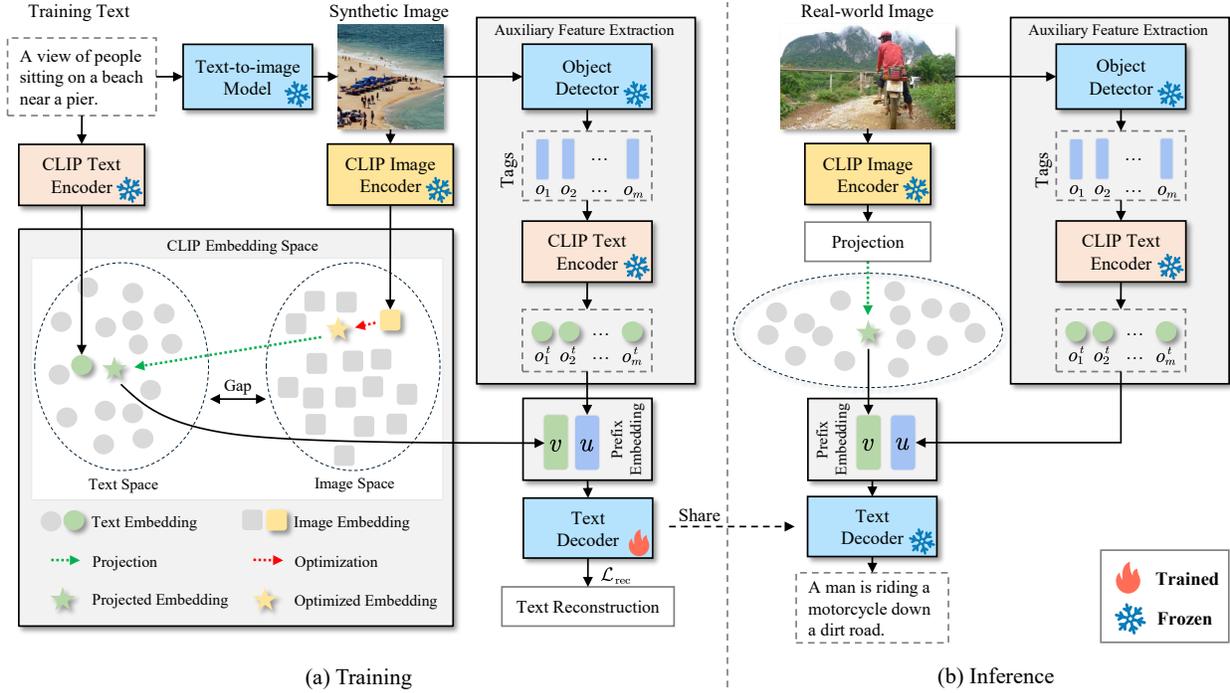
Figure 1: Overview of our proposed method. (a) shows the training procedure. A generated image corresponding to the training text is encoded and then refined in the CLIP embedding space, leading to improved cross-modal alignment. Objects detected in the image guide a caption decoder for text reconstruction. Except for the decoder, other pre-trained models are frozen. (b) shows the inference procedure. A real-world image is encoded and projected in the same way as described in (a), mitigating the discrepancy between training and inference. Then, objects from this natural image help the decoder obtain a caption.

ing synthetic image-text pairs. We refer to the embeddings encoded by CLIP as text and image features. During training, a pre-trained text-to-image model is utilized to generate related images for the texts within the corpus. Then, pseudo features of these images are optimized in the CLIP embedding space to enhance the alignment between synthetic images and their corresponding texts. We extract the features of all texts to construct a support set, and project the optimized pseudo features into the text embedding space. The objects occurring in images render an anchor point that strengthens cross-modal alignment. A lightweight structure serves as the language decoder, taking both projected features and auxiliary objects as inputs. For inference, real images are directly encoded and projected as the decoder's input.

Pseudo Image Feature Synthesis

The modality gap in the CLIP embedding space prevents the language decoder from directly utilizing image features as input because of its training on text features. The lack of images under text-only captioning causes the discrepancy between training and inference. Toward this issue, we exploit a generation-then-refinement method, which firstly constructs initial pseudo features from synthetic images, and then refine those features for better image-text alignment.

Synthetic Image Generation. Given a text corpus $D_t = \{t_1, ..., t_n\}$ with n texts, we leverage the generative capacity of a pre-trained text-to-image model G to obtain a synthetic

image $s_i = G(t_i)$ for each t_i. But directly training the image captioning model using synthetic image-text pairs results in limited performance. This arises due to the fact that real images are depicted with multi-faceted details, while synthetic images generated from simple text tend to lack diverse contexts. Considering the satisfactory quality of those synthetic images, we adopt their CLIP features as initialization.

Feature Optimization. Since CLIP is pre-trained on web-crawled data, its output features indicate the association of real-world texts and images. Optimizing the pseudo image features in the CLIP multimodal space could lead to better alignment with their paired texts, which may narrow the disparity between the pseudo features and real features. Specifically, the pre-trained CLIP encoder outputs a pseudo image feature $s_i^v = \text{CLIP}_v(s_i)$ for the synthetic image s_i and a text feature $t_j^t = \text{CLIP}_t(t_j)$ for the text t_j, where CLIP_v and CLIP_t denote the image encoder and text encoder, respectively. The similarity between s_i^v and t_j^t is calculated as:

$$f_{\text{sim}}(s_i^v, t_j^t) = \frac{s_i^v}{||s_i^v||} \cdot \frac{t_j^t}{||t_j^t||}. \quad (1)$$

Afterward, the pseudo image feature s_i^v would be optimized by a contrastive loss as follows:

$$\nabla_{s_i^v}\mathcal{L}_{\text{con}} = \nabla_{s_i^v}[-\frac{1}{b}\sum_{i=1}^{b}\log\frac{\exp(f_{\text{sim}}(s_i^v, t_i^t)/\tau)}{\sum_{j=1}^{b}\exp(f_{\text{sim}}(s_i^v, t_j^t)/\tau)}], \quad (2)$$

where b is the mini-batch size and τ is a temperature parameter. Intuitively, Eq. 2 enforces that the pseudo image feature s_i^v exhibits high cosine similarities with its corresponding text features t_i^t within the multimodal feature space of CLIP, while having low similarities with other features. That is, refining s_i^v through the contrastive loss essentially distills the knowledge of real-world image-text correspondence.

Feature Projection. To further enrich the semantics of the optimized pseudo image feature and reduce the modality gap in the CLIP embedding space, SynTIC assembles abundant textual semantics and represents the pseudo image feature in the text embedding space. Since the contrastive loss shapes the cross-modal association, the CLIP text features could be employed to grab image representations (Zhou et al. 2022). Specifically, we reserve the text feature $t_k^t = \text{CLIP}_t(t_k)$ for each t_k^t in the corpus D_t to construct the projection support set $D_t^s = \{t_1^t, ... t_n^t\}$. The pseudo image feature is projected by a weighted combination of all features in this support set. SynTIC utilizes the cross-modal capability of CLIP and expresses the weight depending on the cosine similarity f_{sim} of two modalities. The projected vector of the pseudo image feature s_i^v could be obtained as follows:

$$
\begin{aligned}
v_i = f_{\text{proj}}(s_i^v, D_t^s) &= \sum_{j=1}^n w_j * t_j^t \\
&= \sum_{j=1}^n \frac{\exp(f_{\text{sim}}(s_i^v, t_j^t)/\tau)}{\sum_{k=1}^n \exp(f_{\text{sim}}(s_i^v, t_k^t)/\tau)} * t_j^t,
\end{aligned}
\tag{3}
$$

where w_j denotes the weight of j-th text feature t_j^t. Since v_i is a combination of multiple text features, it contains abundant textual semantics. We utilize this projected vector in the text embedding space as the image feature which serves as a condition to decode a caption.

Auxiliary Feature Extraction

Salient objects in images could be identified by pre-trained object detectors, and these detected objects are frequently referenced in the paired text. In contrast to the visual object features from over-sampled regions, our method introduces object tags detected in images as auxiliary features to facilitate the process of learning semantic alignments between images and captions. Specifically, textual object information is regarded as a clue for caption decoding. Given a synthetic image s_i, we employ an object detection model to extract objects $O_i = \{o_1, ..., o_m\}$ in the image. Then, the CLIP text encoder could obtain a text feature o_k^t for all objects as:

$$
O_i^t = \{o_1^t, ..., o_m^t\} = \{\text{CLIP}_t(o_1), ..., \text{CLIP}_t(o_m)\}, \quad (4)
$$

where m is the number of objects detected from the image. Since the pseudo image feature would change after our contrastive optimization, we integrate $\{o_1^t, ..., o_m^t\}$ by the multi-head attention mechanism (Vaswani et al. 2017). The auxiliary feature u_i for s_i is obtained as follows:

$$
u_i = \text{MHAttn}(s_i^v, O_i^t, O_i^t), \tag{5}
$$

where $\text{MHAttn}(Q, K, V)$ denotes the multi-head attention for query s_i^v, key O_i^t, and value O_i^t. The presence of textual information related to the objects in u_i serves as an anchor, that assists cross-modal alignment.

Prefix Decoding

The decoder architecture, relying on the Transformer models, has found widespread application in various natural language generation tasks. Thus, following previous works (Li et al. 2023), our method employs a lightweight Transformer-based decoder as the generation model. Both image features and auxiliary object features should be involved in decoding. Considering these features as the prefix embedding, we train the decoder with prefix language modeling. The model would predict the next token conditioned on input features in an autoregressive manner.

Training. Given a caption text $t = \{w_1, ..., w_{|t|}\}$, we obtain the corresponding projected image feature v and auxiliary object feature u as the prefix embedding. Then, the decoder is trained to reconstruct the caption text t conditioned on the prefix, which minimizes the objective as follows:

$$
\mathcal{L}_{\text{rec}} = -\frac{1}{|t|} \sum_{i=1}^{|t|} \log P_\theta(w_i | w_{<i}, v, u), \tag{6}
$$

where $|t|$ denotes the sequence length of t, and θ is the trainable parameter of the decoder. The caption decoder is trainable, while implemented pre-trained models are frozen.

Inference. Our method could unify training and inference by utilizing synthetic paired data. For a provided real-world image, we project it as the image feature using the textual support set, and simultaneously pick up the image's object tags that serve as the auxiliary feature. These image and object features are prefixes for the trained decoder to generate caption tokens step by step.

Experimentation

Experimental Settings

Datasets. Experiments are conducted on three benchmark datasets: MSCOCO (Chen et al. 2015), Flickr30k (Young et al. 2014), and SS1M (Feng et al. 2019). We follow Karpathy (Karpathy and Fei-Fei 2015) to split the MSCOCO and Flickr30k datasets into the training, validation, and test sets. Only text annotations from the training set are used for text-only training. SS1M is a web-crawled corpus, designed for MSCOCO captions. It relies on the names of object classes in MSCOCO as keywords to gather 2,322,628 unique image descriptions from Shutterstock. We use this corpus and additionally filter out sentences with more than fifteen words, resulting in 978,662 descriptions. More experimental settings and results are given in the supplementary material[1].

Evaluation Metrics. Following common settings (Tewel et al. 2022; Li et al. 2023), there are several metrics considered to evaluate the generated caption, including BLEU-4 (**B@4**) (Papineni et al. 2002), METEOR (**M**) (Banerjee and Lavie 2005), ROUGE-L (**R-L**) (Lin 2004), CIDEr (**C**) (Vedantam, Lawrence Zitnick, and Parikh 2015), and SPICE (**S**) (Anderson et al. 2016).

---

[1]https://arxiv.org/abs/2312.08865

Method	Data		MSCOCO					Flickr30K				
	I	T	B@4(↑)	M(↑)	R-L (↑)	C(↑)	S(↑)	B@4(↑)	M(↑)	R-L (↑)	C(↑)	S(↑)
BUTD	✓	✓	36.2	27.0	56.4	113.5	20.3	27.3	21.7	-	56.6	16.0
ClipCap	✓	✓	33.5	27.5	-	113.1	21.1	-	-	-	-	-
Barraco et al.	✓	✓	36.0	27.8	56.5	114.9	20.8	-	-	-	-	-
Feng et al.	✓	✓	18.6	17.9	43.1	54.9	11.1	-	-	-	-	-
Laina et al.	✓	✓	19.3	20.2	45.0	61.8	12.9	-	-	-	-	-
ESPER-Style	✓	✓	21.9	21.9	-	78.2	-	-	-	-	-	-
CLIPRe		✓	4.9	11.4	29.0	13.6	5.3	5.2	11.6	27.6	10.0	5.7
ZeroCap		✓	7.0	15.4	31.8	34.5	9.2	5.4	11.8	27.3	16.8	6.2
MAGIC		✓	12.9	17.4	39.9	49.3	11.3	6.4	13.1	31.6	20.4	7.1
CapDec		✓	26.4	25.1	51.8	91.8	11.9	17.7	20.0	43.9	39.1	9.9
CLOSE		✓	28.6	25.2	-	95.4	18.1	-	-	-	-	-
DeCap		✓	24.7	25.0	-	91.2	18.7	21.2	21.8	-	**56.7**	15.2
SynTIC (Ours)		✓	**29.9**	**25.8**	**53.2**	**101.1**	**19.3**	**22.3**	**22.4**	**47.3**	56.6	**16.6**

Table 1: In-domain image captioning results on MSCOCO and Flickr30K. I and T denote training with image data and text data, respectively. ↑ means higher is better. The best results are in bold.

Method	MSCOCO→Flickr30K					Flickr30K→MSCOCO				
	B@4(↑)	M(↑)	R-L (↑)	C(↑)	S(↑)	B@4(↑)	M(↑)	R-L (↑)	C(↑)	S(↑)
CLIPRe	4.4	9.6	27.2	5.9	4.8	3.0	9.9	22.8	8.5	3.9
MAGIC	6.2	12.2	31.3	17.5	5.9	5.2	12.5	30.7	18.3	5.7
CapDec	17.3	18.6	42.7	35.7	7.2	9.2	16.3	36.7	27.3	10.4
DeCap	16.3	17.9	-	35.7	11.1	12.1	18.0	-	44.4	10.9
DeCap-TT	17.7	**20.2**	-	42.0	13.8	19.7	20.9	-	63.1	13.9
SynTIC	17.9	18.6	42.7	38.4	11.9	14.6	19.4	40.9	47.0	11.9
SynTIC-TT	**19.4**	**20.2**	**44.8**	**43.2**	**13.9**	**20.6**	**21.3**	**46.6**	**64.4**	**14.3**

Table 2: Cross-domain image captioning results, where X→Y means source domain → target domain.

Method	Dataset	B@4	M	R-L	C	S
ZeroCap	-	2.6	11.5	-	14.6	5.5
CLIPRe	CC3M	4.6	13.3	-	25.6	9.2
DeCap	CC3M	8.8	16.0	-	42.1	10.9
DeCap	SS1M	8.9	17.5	-	50.6	13.1
SynTIC	SS1M	**13.3**	**17.6**	**36.7**	**55.6**	**13.3**

Table 3: Zero-shot image captioning results on MSCOCO.

Implementations. For synthetic image generation, we utilize Stable Diffusion v1-5 (Rombach et al. 2022) as the text-to-image model, which leverages 20 sampling steps to generate a 512×512 image for each input text. A pre-trained CLIP VIT-B/32 model is used as the feature extractor. The Adam optimizer (Kingma and Ba 2015) is employed to optimize parameters. For better alignment, we optimize pseudo image features using Eq. 2 with a temperature τ of 1/100 and a learning rate of 1e-5. We use a transformer decoder structure with 4 layers and 4 attention heads as the caption generator. The projection temperatures τ in Eq. 3 for MSCOCO, Flickr30k, and SS1M are set to 1/100, 1/80, and 1/100, respectively. A pre-trained object detection model, DETR (Carion et al. 2020), is used to extract the objects.

Baselines. Supervised and unsupervised methods are set as baselines in the experiments. The fully supervised methods, such as ClipCap (Mokady, Hertz, and Bermano 2021) and BUTD (Anderson et al. 2018), are considered for in-domain captioning. The unsupervised methods, containing CLIPRe (Su et al. 2022), MAGIC (Su et al. 2022), ZeroCap (Tewel et al. 2022), CapDec (Nukrai, Mokady, and Globerson 2022), CLOSE (Gu, Clark, and Kembhavi 2023), and DeCap (Li et al. 2023), are the main comparison baselines crossing different captioning scenes.

In-domain Unpaired Image Captioning

Unpaired captioning experiments are first conducted to evaluate performance. For the supervised baselines, BUTD uses object detection networks to extract visual features for caption generation. ClipCap and the method of (Barraco et al. 2022) utilize CLIP as the visual encoder. Regarding the unsupervised methods, we take CLIPRe, ZeroCap, MAGIC, CapDec, CLOSE, and DeCap into consideration. Moreover, ESPER-Style (Yu et al. 2022b), Feng et al. (2019), and Laina et al. (2019) achieve unpaired image captioning by taking images and captions from MSCOCO as unpaired data.

Results. Table 1 illuminates the in-domain captioning results on the MSCOCO and Flickr30K datasets. Compared

Method	MSCOCO					MSCOCO→Flickr30K				
	B@4(↑)	**M**(↑)	**R-L** (↑)	**C**(↑)	**S**(↑)	**B@4**(↑)	**M**(↑)	**R-L** (↑)	**C**(↑)	**S**(↑)
Baseline	25.1	23.8	49.4	88.5	17.7	15.9	17.5	41.0	34.8	11.3
+Feature Optimization (FO)	26.2	24.0	50.2	89.8	17.8	16.3	18.1	42.0	35.5	11.3
+Feature Projection (FP)	28.5	25.8	52.7	98.2	19.3	16.9	18.0	42.2	36.2	11.3
+Auxiliary Feature (AF)	25.5	24.2	49.6	91.2	18.2	16.9	17.6	41.6	34.7	11.2
+FO & AF w/o FP	26.3	24.4	50.8	92.1	18.2	16.5	17.8	41.0	35.4	11.2
+FP & FO w/o AF	29.1	25.7	52.9	100.0	19.4	17.1	18.4	42.5	36.8	11.5
+FP & AF w/o FO	29.1	**26.0**	53.1	100.3	**19.6**	16.9	18.3	42.4	37.1	11.7
SynTIC	**29.9**	25.8	**53.2**	**101.1**	19.3	**17.9**	**18.6**	42.7	**38.4**	**11.9**

Table 4: Ablation study results on MSCOCO under in-domain and cross-domain settings.
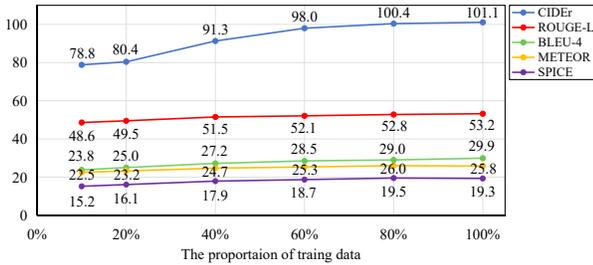

Figure 2: The performance of SynTIC on MSCOCO with different proportions of training data.

with other unsupervised methods, SynTIC exhibits significant performance improvement on most evaluation metrics. showing our method's effectiveness. The unpaired captioning methods could achieve competitive results since they take both images and texts within datasets into consideration. In contrast to CLIPRe, ZeroCap, and MAGIC, text-only methods, including CapDec, CLOSE, DeCap, and SynTIC, obtain better performance by mitigating the modality gap in CLIP. It indicates that the feature gap matters among CLIP-based captioning methods, since this gap degrades performance. Competing with two strong baselines, CLOSE and DeCap, our method increases all metrics on MSCOCO, while also demonstrating performance gains on Flickr30k. The noise injection in CLOSE cannot completely narrow the modality gap. DeCap is hindered by inconsistencies between training and inference, as it only projects image features during inference. Note that SynTIC even outperforms the supervised method on Flickr30k with respect to both METEOR and SPICE.

Cross-domain Image Captioning

Cross-domain image captioning tests the generalization ability. It requires training the model on a source dataset and directly testing it on a target dataset. SynTIC is compared with various unsupervised methods, including CLIPRe, MAGIC, CapDec, DeCap, and DeCap-TT, where TT represents a feature projection variation using target domain captions.

Results. As shown in Table 2, our methods consistently outperform baselines on various metrics under both cross-

domain settings. Specifically, SynTIC demonstrates significant superiority over CLIPRe and MAGIC. Compared with CapDec, our method beats it in 8 out of 10 metrics, while having similar results on METEOR and ROUGE-L. We attribute CapDec's impressive cross-domain captioning performance to its utilization of the pre-trained language model, GPT2-Base (Radford et al. 2019). The knowledge derived from GPT2 boosts CapDec through the fine-tuning on text data, especially for the ample captions in MSCOCO. SynTIC deploys a lightweight decoder instead of large language models, which is more computation-friendly while achieving superior results. In contrast to DeCap, SynTIC synthesizes image data to unify training and inference, showing excellent performance. Regardless of the availability of data in the target domain, SynTIC and its variation achieve the best results on all metrics.

Zero-shot Image Captioning

Both in-domain and cross-domain image captioning rely on human-annotated text data with rich prior knowledge and specific descriptive styles (e.g., texts in MSCOCO). Thus, it cannot comprehensively demonstrate the model's zero-shot capability. We utilize a web-crawled text corpus, SS1M, for training and verify the zero-shot captioning performance on MSCOCO, with ZeroCap, CLIPRe, and DeCap as baselines.

Results. The zero-shot performance is shown in Table 3. ZeroCap relies solely on pre-trained models without fine-tuning and obtains the lowest caption quality. In contrast, CLIPRe shows improvement, as it uses CLIP to retrieve texts from CC3M (Sharma et al. 2018). Our method outperforms ZeroCap and CLIPRe significantly across all evaluation metrics. Compared with DeCap, which also uses SS1M, SynTIC greatly improves BLEU-4 by 4.4 and CIDEr by 5, while obtaining competitive performance on other metrics. This demonstrates our method could realize better cross-modal alignment with the help of synthetic images and detected objects. The effectiveness of our method is not contingent on the text data with specific descriptive styles.

Ablation Study

Effect of Components. An ablation study is performed on MSCOCO under both in-domain and cross-domain settings, as presented in Table 4. To evaluate different compo-

Reference:
A man cutting a slice of cake with the knife.
DeCap:
a person cuts a cake with a large spoon.
SynTIC:
a person cutting a piece of cake with a knife.
(Objects: person, knife, cake, clock, dining table)
(a)

Reference:
A cat in a kitchen on top of a refrigerator.
DeCap:
a cat is standing up in the corner of a bathroom.
SynTIC:
a cat that is sitting on top of a refrigerator.
(Objects: refrigerator, cat)
(b)

Reference:
A stop sign and street signs next to a brick building.
DeCap:
a street sign is in front of a stop sign.
SynTIC:
a red stop sign sitting in front of a brick building.
(Objects: stop sign)
(c)

Reference:
A little boy sitting on a suitcase on the floor.
DeCap:
a young child sitting in front of a laptop.
SynTIC:
a little boy sitting in front of a suitcase.
(Objects: suitcase, person)
(d)

Reference:
A park with a walking track is pictured with mountains in the distance.
DeCap:
a park with a couple of people are sitting on the ground .
SynTIC:
a view of a park with mountains in the background.
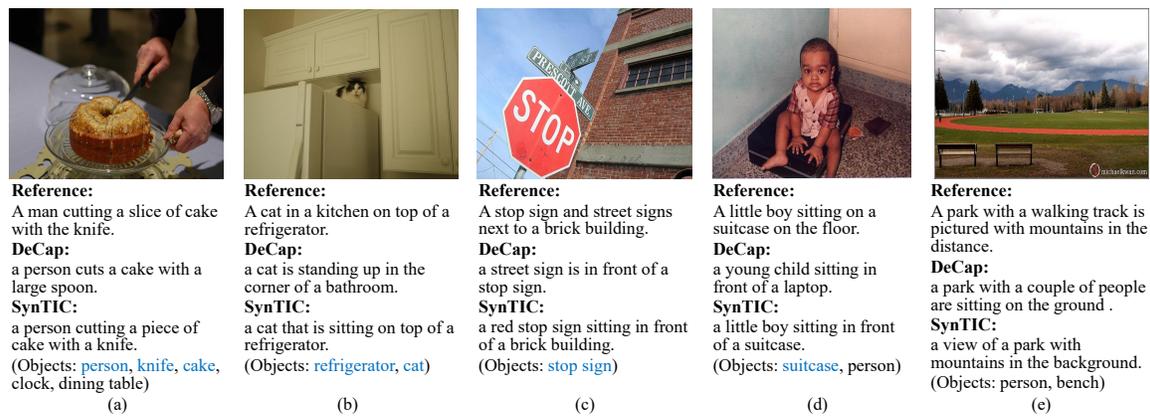(Objects: person, bench)
(e)

Figure 3: Examples generated from DeCap and SynTIC compared with reference captions on MSCOCO.

nents within SynTIC, we establish a basic baseline method (i.e., Baseline) by using generated pseudo image features directly as inputs for training while real-world images for inference, without feature optimization (FO), feature projection (FP), and auxiliary features (AF). Due to the difference between synthetic and real images, Baseline is inferior to SynTIC but already approaching DeCap. The method of optimizing pseudo image features (i.e., +Feature Optimization) brings performance gains over Baseline, indicating that pseudo image features refined by the contrast loss enhance cross-modal alignment. The method of projecting pseudo image features into the CLIP text space (i.e., +Feature Projection) shows great improvements. It demonstrates that gathering text data to represent image features enriches the semantics and effectively bridges the modality gap. Note that this method beats DeCap, which also shows the usefulness of synthetic pairs. When directly attaching auxiliary object features to Baseline (i.e., +Auxiliary Feature), the performance has limited improvements. The reason may be that objects are detected from images, and thus their features are redundant with the image features to some extent. However, when projected features are used as the decoder input (i.e., +FP & AF w/o FO), the auxiliary features work efficiently and achieve optimal performance in METEOR and SPICE. It indicates that auxiliary features could attach essential information to projected features, benefiting caption generation and cross-modal alignment. Besides, the combination of other components (e.g., +FO & AF w/o FP and +FP & FO w/o AF) also contributes to performance gains. SynTIC achieves the best performance by utilizing the above three components, demonstrating the component's effectiveness.

Effect of the Training Data Size. To explore the effect of training data size on SynTIC, we sample data with different proportions from MSCOCO for training. The results are shown in Fig. 2. The performance of SynTIC declines as the training data decreases, but the performance does not drop catastrophically even when there is only 10% of the training data available. It demonstrates that SynTIC remains competitive even in scenarios with less text data. Note that, regarding all metrics, our method consistently outperforms MAGIC, CapDec, and DeCap with 10%, 60%, and 60% of

training data, respectively. In summary, increasing the size of training data enhances SynTIC, and our method still performs well with a sharp decrease of training data.

Quantitative Evaluation

We compare the captions generated by SynTIC with those from DeCap under the in-domain setting on the MSCOCO dataset, as shown in Fig. 3. The detected objects are listed, and the used ones are highlighted in blue. SynTIC generates more accurate and contextually grounded captions by leveraging synthetic pairs and detected objects. In detail, our captions contain "knife" in case (a), "refrigerator" in case (b), and "suitcase" in case (d), whereas DeCap generates incorrect words like "spoon" in case (a), "bathroom" in case (b), and "laptop" in case (d). The reason may be that DeCap uses only projected features as the decoder input, causing the lack of some image details, while SynTIC achieves better cross-modal alignment. Besides, DeCap's description in case (c) confuses the positional relationship of objects. In case (e), without useful objects, our method obtains correct background information: "mountains", while DeCap misses that part and generates nonessential details about "a couple of people". This indicates that the auxiliary feature does not restrict our generated descriptions to recognized objects.

Conclusion

In this paper, we explore synthetic pairs for text-only image captioning with unified training and inference. Considering that the generated images from simple text are distinct from natural images, a generation-then-refinement procedure is developed to adjust image representation. We first obtain pseudo features of the generated images, and optimize them with contrastive constraint, enforcing pseudo features toward the real ones. Then, textual information from the training corpus is assembled to project pseudo features. This process results in a semantically rich image representation in the text embedding space, contributing to improved cross-modal alignment. Moreover, objects detected in images aid the caption decoder in capturing essential content. Experimental results on several benchmark datasets demonstrate that SynTIC outperforms the state-of-the-art methods.

References

Agrawal, H.; Desai, K.; Wang, Y.; Chen, X.; Jain, R.; Johnson, M.; Batra, D.; Parikh, D.; Lee, S.; and Anderson, P. 2019. Nocaps: Novel object captioning at scale. In *ICCV*, 8948–8957.

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: A visual language model for few-shot learning. In *NeurIPS*, 23716–23736.

Anderson, P.; Fernando, B.; Johnson, M.; and Gould, S. 2016. SPICE: Semantic propositional image caption evaluation. In *ECCV*, 382–398.

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 6077–6086.

Banerjee, S.; and Lavie, A. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 65–72.

Barraco, M.; Cornia, M.; Cascianelli, S.; Baraldi, L.; and Cucchiara, R. 2022. The unreasonable effectiveness of CLIP features for image captioning: an experimental analysis. In *CVPR*, 4662–4670.

Ben, H.; Pan, Y.; Li, Y.; Yao, T.; Hong, R.; Wang, M.; and Mei, T. 2021. Unpaired image captioning with semantic-constrained self-learning. *IEEE Transactions on Multimedia*, 24: 904–916.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *ECCV*, 213–229.

Chen, X.; Fang, H.; Lin, T.-Y.; Vedantam, R.; Gupta, S.; Dollár, P.; and Zitnick, C. L. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Dai, W.; Hou, L.; Shang, L.; Jiang, X.; Liu, Q.; and Fung, P. 2022. Enabling Multimodal Generation on CLIP via Vision-Language Knowledge Distillation. In *Findings of ACL*, 2383–2395.

Feng, Y.; Ma, L.; Liu, W.; and Luo, J. 2019. Unsupervised image captioning. In *CVPR*, 4125–4134.

Gao, J.; Wang, S.; Wang, S.; Ma, S.; and Gao, W. 2019. Self-critical n-step training for image captioning. In *CVPR*, 6300–6308.

Gu, S.; Clark, C.; and Kembhavi, A. 2023. I Can't Believe There's No Images! Learning Visual Tasks Using only Language Supervision. In *ICCV*, 2672–2683.

Hu, X.; Gan, Z.; Wang, J.; Yang, Z.; Liu, Z.; Lu, Y.; and Wang, L. 2022. Scaling up vision-language pre-training for image captioning. In *CVPR*, 17980–17989.

Jiang, Q.; Chen, C.; Zhao, H.; Chen, L.; Ping, Q.; Tran, S. D.; Xu, Y.; Zeng, B.; and Chilimbi, T. 2023. Understanding and constructing latent modality structures in multimodal representation learning. In *CVPR*, 7661–7671.

Karpathy, A.; and Fei-Fei, L. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 3128–3137.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*.

Kuo, C.-W.; and Kira, Z. 2023. HAAV: Hierarchical Aggregation of Augmented Views for Image Captioning. In *CVPR*, 11039–11049.

Laina, I.; Rupprecht, C.; and Navab, N. 2019. Towards unsupervised image captioning with shared multimodal embeddings. In *ICCV*, 7414–7424.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022a. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *ICML*, 12888–12900.

Li, J.; Mao, Z.; Fang, S.; and Li, H. 2022b. ER-SAN: Enhanced-Adaptive Relation Self-Attention Network for Image Captioning. In *IJCAI*, 1081–1087.

Li, W.; Zhu, L.; Wen, L.; and Yang, Y. 2023. DeCap: Decoding CLIP Latents for Zero-Shot Captioning via Text-Only Training. In *ICLR*, 1–18.

Li, X.; Yin, X.; Li, C.; Zhang, P.; Hu, X.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, 121–137.

Liang, V. W.; Zhang, Y.; Kwon, Y.; Yeung, S.; and Zou, J. Y. 2022. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *NeurIPS*, 17612–17625.

Lin, C.-Y. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

Luo, J.; Li, Y.; Pan, Y.; Yao, T.; Feng, J.; Chao, H.; and Mei, T. 2023. Semantic-conditional diffusion networks for image captioning. In *CVPR*, 23359–23368.

Meng, Z.; Yang, D.; Cao, X.; Shah, A.; and Lim, S.-N. 2022. Object-centric unsupervised image captioning. In *ECCV*, 219–235.

Mokady, R.; Hertz, A.; and Bermano, A. H. 2021. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*.

Nukrai, D.; Mokady, R.; and Globerson, A. 2022. Text-Only Training for Image Captioning using Noise-Injected CLIP. In *Findings of EMNLP*, 4055–4063.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: A method for automatic evaluation of machine translation. In *ACL*, 311–318.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, 8748–8763. PMLR.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *CVPR*, 10684–10695.

Sharma, P.; Ding, N.; Goodman, S.; and Soricut, R. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2556–2565.

Stefanini, M.; Cornia, M.; Baraldi, L.; Cascianelli, S.; Fiameni, G.; and Cucchiara, R. 2022. From show to tell: A survey on deep learning-based image captioning. *IEEE transactions on pattern analysis and machine intelligence*, 45(1): 539–559.

Su, Y.; Lan, T.; Liu, Y.; Liu, F.; Yogatama, D.; Wang, Y.; Kong, L.; and Collier, N. 2022. Language models can see: Plugging visual controls in text generation. *arXiv preprint arXiv:2205.02655*.

Tewel, Y.; Shalev, Y.; Schwartz, I.; and Wolf, L. 2022. ZeroCap: Zero-shot image-to-text generation for visual-semantic arithmetic. In *CVPR*, 17918–17928.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *NeurIPS*, 1–11.

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. CIDEr: Consensus-Based Image Description Evaluation. In *CVPR*, 4566–4575.

Wang, Z.; Yu, J.; Yu, A. W.; Dai, Z.; Tsvetkov, Y.; and Cao, Y. 2022. SimVLM: Simple Visual Language Model Pretraining with Weak Supervision. In *ICLR*, 1–17.

Wu, M.; Zhang, X.; Sun, X.; Zhou, Y.; Chen, C.; Gu, J.; Sun, X.; and Ji, R. 2022. Difnet: Boosting visual information flow for image captioning. In *CVPR*, 18020–18029.

Wu, Y.; Zhu, L.; Jiang, L.; and Yang, Y. 2018. Decoupled novel object captioner. In *ACM Multimedia*, 1029–1037.

Xing, Y.; Shi, Z.; Meng, Z.; Lakemeyer, G.; Ma, Y.; and Wattenhofer, R. 2021. KM-BART: Knowledge Enhanced Multimodal BART for Visual Commonsense Generation. In *ACL-IJCNLP*, 525–535.

Young, P.; Lai, A.; Hodosh, M.; and Hockenmaier, J. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2: 67–78.

Yu, J.; Li, H.; Hao, Y.; Zhu, B.; Xu, T.; and He, X. 2023. CgT-GAN: CLIP-Guided Text GAN for Image Captioning. In *ACM Multimedia*, 2252–2263.

Yu, J.; Wang, Z.; Vasudevan, V.; Yeung, L.; Seyedhosseini, M.; and Wu, Y. 2022a. CoCa: Contrastive Captioners are Image-Text Foundation Models. *Transactions on Machine Learning Research*, Aug 2022.

Yu, Y.; Chung, J.; Yun, H.; Hessel, J.; Park, J.; Lu, X.; Ammanabrolu, P.; Zellers, R.; Bras, R. L.; Kim, G.; et al. 2022b. Multimodal knowledge alignment with reinforcement learning. *arXiv preprint arXiv:2205.12630*.

Zeng, P.; Zhang, H.; Song, J.; and Gao, L. 2022. S^2 transformer for image captioning. In *IJCAI*, 1608–1614.

Zhou, Y.; Li, C.; Chen, C.; Gao, J.; and Xu, J. 2022. Lafite2: Few-shot text-to-image generation. *arXiv preprint arXiv:2210.14124*.

Zhu, P.; Wang, X.; Zhu, L.; Sun, Z.; Zheng, W.-S.; Wang, Y.; and Chen, C. 2023. Prompt-Based Learning for Unpaired Image Captioning. *IEEE Transactions on Multimedia*, 1–15.