electronics

MDPI

Article

Semantic-Enhanced Cross-Modal Fusion for Improved Unsupervised Image Captioning

Nan Xiang [1,2,3,*], Ling Chen [1], Leiyan Liang [1], Xingdi Rao [1] and Zehao Gong [1]

[1] Liangjiang International College, Chongqing University of Technology, Chongqing 401135, China; lingchen@stu.cqut.edu.cn (L.C.); 51201910121@2020.cqut.edu.cn (L.L.); 18581064310@163.com (X.R.); gzh_study@163.com (Z.G.)
[2] College of Computer Science, Chongqing University, Chongqing 400044, China
[3] Chongqing Jialing Special Equipment Co., Ltd., Chongqing 400032, China
* Correspondence: xiangnan@cqut.edu.cn

Abstract: Unsupervised image captioning often grapples with challenges such as image–text mismatches and modality gaps, resulting in suboptimal captions. This paper introduces a semantic-enhanced cross-modal fusion model (SCFM) to address these issues. The SCFM integrates three innovative components: a text semantic enhancement network (TSE-Net) for nuanced semantic representation; contrast learning for optimizing similarity measures between text and images; and enhanced visual selection decoding (EVSD) for precise captioning. Unlike existing methods that struggle with capturing accurate semantic relationships and flexibility across scenarios, the proposed model provides a robust solution for unbiased and diverse captioning. Through experimental evaluations on the MS COCO and Flickr30k datasets, SCFM demonstrates significant improvements over the benchmark model, enhancing the CIDEr and BLEU-4 metrics by 3.6% and 3.2%, respectively. Visualization analysis further reveals the model's superiority in increasing variability between hidden features and its potential in cross-domain and stylized image captioning. The findings not only contribute to the advancement of image captioning techniques but also open avenues for future research. Further investigations will explore SCFM's adaptability to other multimodal tasks and refine it for more intricate image–text relationships.

Keywords: image caption generation; text semantic enhancement; contrastive learning; image-enhanced decoding

Citation: Xiang, N.; Chen, L.; Liang, L.; Rao, X.; Gong, Z. Semantic-Enhanced Cross-Modal Fusion for Improved Unsupervised Image Captioning. *Electronics* **2023**, *12*, 3549. https://doi.org/10.3390/electronics12173549

Academic Editor: Gemma Piella

Received: 11 July 2023
Revised: 14 August 2023
Accepted: 19 August 2023
Published: 22 August 2023

1. Introduction

Image captioning is a crucial research task in the multimodal area where computer vision and natural language processing intersect. It aims to turn images into natural language descriptions that provide the basis for computers to achieve an understanding of images and generate human-readable text. Traditional image captioning methods depend heavily on supervised learning and require large amounts of paired image and text data for training [1]. However, acquiring large-scale paired data is laborious and costly, imposing limitations on the applicability and scalability of such methods.

To overcome this limitation, the attention of researchers has gradually shifted towards unsupervised image captioning techniques. These techniques aim to learn the correspondence between images and text from unlabeled image data, enabling automated image description generation. For instance, Iro Laina et al. [2] proposed an unsupervised image captioning method based on shared multimodal embeddings. They achieved cross-modal feature representation by combining image and text encoders. Similarly, Yang Feng et al. [3] presented an unsupervised image description method utilizing generative adversarial networks (GAN) to generate diverse image captions. Using adversarial training of generators and discriminators, they generated more successful and diverse caption results. However, the approaches mentioned above no longer require manual labeling of image–text pairs

but instead rely on matching image blocks to labels, which can be challenging to control. Furthermore, since the pseudo-descriptions are trained on a fixed label set, they may not be applicable to scenarios beyond the predefined label set. To address this issue, Yoad Tewel et al. [4] put forward a zero-sample image-to-text generation method that utilizes visual-semantic arithmetic to generate images' descriptive texts. Yixuan Su et al. [5] introduced a text generation method with visual control, incorporating a visual control mechanism to generate image-related text descriptions by specifying visual conditions. This approach enhances the accuracy and semantic consistency of the generated descriptions. Challenges related to the modality gap in multimodal contrastive representation learning have been recognized in the literature [6]; David Nukrai et al. [7] proposed a noise-injection-based text training method for image description generation. By employing a contrast learning strategy for noise injection, they improved the diversity and quality of the generated descriptions.

Although previous works have explored various approaches and strategies to address the problem of modality gap and image–text mismatch, they still have several drawbacks. First, most existing models directly use the image feature vectors and text feature vectors in a shared space where multimodal representations are learned, yet the differences between different modalities in this shared space still exist, which inevitably leads to bias in the inference process. Even though this problem is recognized, they often address it by adding Gaussian noise without conducting further analysis [7]. Secondly, most existing methods in the text generation stage rely on the widely used maximum probability decoding strategy commonly employed in natural language processing. However, this approach often leads to degradation in the generated results. Specifically, the generated text tends to be generic and exhibits unreasonable repetitions at various levels, including words, phrases, and sentences. This problem stems from the decoding strategy and the correspondence between the entire image and the generated words. Consequently, a significant number of repetitive words appear among the candidate words, resulting in varying degrees of semantic repetition in the generated sentences.

To address the aforementioned challenges, this study proposes a novel semantic-enhanced cross-modal fusion model. The model leverages a text semantic enhancement network to extract text-enhanced semantic representations, effectively capturing the semantic associations between texts, strengthening the semantic features, and attenuating the modal features. This process provides robust support for subsequent feature fusion. Furthermore, contrast learning is utilized to optimize similarity measures and feature representation consistency between texts. Meanwhile, an image enhancement decoding strategy is introduced to generate accurate and diverse description results. In contrast to the traditional maximum probability decoding strategy, this approach structures the decoding process and leverages the rich information present in the image. It employs a top-k sampling technique to generate a set of diverse candidate sentences, guaranteeing the diversity of the generated captions. Subsequently, a defined metric, such as cosine similarity or a learned distance function, is employed to calculate the similarity between the candidate sentences and the image. The sentence with the highest similarity to the image is selected as the final output, ensuring the accuracy and visual relevance of the description.

The central aim of this research is to propose and validate a novel semantic-enhanced cross-modal fusion model that addresses specific challenges in existing models. These challenges include bias in the inference process due to differences between modalities and degradation in the quality of generated results. Our approach leverages text semantic enhancement and a unique image enhancement decoding strategy to improve the accuracy, diversity, and quality of the generated image descriptions.

To validate the effectiveness of our proposed model and strategy, extensive experiments are conducted on four widely recognized datasets: MS COCO and Flickr30k for standard image captioning and cross-domain image captioning experiments, and FlickrStyle10K and SentiCap for stylized image captioning experiments. The experimental results demonstrate that our model and strategy significantly enhance the performance of the image

description task. Specifically, our approach improves the description accuracy, diversity, and quality of the generated results compared to traditional methods. The main contributions of this paper can be summarized as follows:

(1) In this paper, we introduce a text semantic enhancement network designed to extract enhanced semantic representations of text. This network is capable of effectively capturing the semantic associations between texts, thereby providing robust support for subsequent feature fusion. To optimize the similarity measure and ensure consistency in feature representation between different texts, we employ contrastive learning. This technique emphasizes the semantic properties of texts by maximizing text differences, reducing the correlation between texts, and attenuating the modal properties of texts. As a result, the proposed network significantly enhances the model's ability to comprehend and articulate the semantics of textual content.

(2) This paper introduces an enhanced decoding strategy to generate precise and diverse caption results by structuring the decoding process and incorporating attention mechanisms and language models. By adopting this strategy, the flexibility and diversity of the generated captions are significantly enhanced, resulting in improved quality and appeal of the generated results.

(3) Comprehensive experiments were conducted on two widely recognized image captioning datasets to evaluate the proposed approach. Comparative analysis against traditional methods demonstrates significant enhancements achieved by our approach in the image captioning task, notably in terms of description accuracy, diversity, and the overall quality of the generated results.

2. Related Work

2.1. Image Captioning

Image captioning is a crucial task in multimodal learning, with the goal of generating accurate and grammatically correct natural language descriptions for images. The general approach for image captioning is the encoder–decoder architecture, where a convolutional neural network (CNN) serves as the encoder to extract image features, and a recurrent neural network (RNN) functions as the decoder to generate image descriptions [8,9]. To enhance the effectiveness of image captioning, various models and methods have been proposed. For instance, Zhou et al. [10] introduced the deep modular co-attention network, which utilizes a cascade of modular co-attention layers to model the relationship between language and vision. Huang et al. [11] proposed the attention over attention (AoA) network, which filters out irrelevant or misleading attention results in the decoder, retaining only useful attention results. Pan et al. [12] addressed existing models' limitations by introducing the X-LAN attention module, enabling the capture of higher-order or infinite-order interactions between modalities through bilinear pooling. These methods have shown promising performance on supervised training with large-scale annotated image–text pairs and have achieved good results on various evaluation benchmarks. However, collecting such annotated datasets is challenging. Therefore, some researchers have explored weakly supervised approaches for model training. For example, Feng et al. [3] proposed a method that solely relies on individual image data and a sentence corpus, eliminating the need for manual annotation of image–text pairs of datasets. Laina et al. [2] connected image information and text information through shared multimodal encoding, leveraging image–label datasets instead of paired image–text datasets. Pseudo-descriptions are generated using object labels and visual content retrieval modules and used as new labels for training. While these approaches partially overcome dataset limitations, they still face issues such as a lack of one-to-one correspondence between generated pseudo-descriptions and images, resulting in descriptions that may contain objects not present in the images.

To address these challenges, recent works have emerged that completely forego the use of existing datasets containing both images and texts for model training. For instance, Tewel et al. [4] proposed ZeroCap, a zero-shot learning method for image captioning that employs CLIP [13] for image feature extraction and GPT2 [14] for caption generation.

However, due to the absence of domain-specific training, this method performs poorly on evaluation benchmarks. Nukrai et al. [7] introduced a training approach where textual data are utilized to adapt the language model to a target style, and image substitution is employed during the inference phase to obtain the desired output. While this approach enables style adaptation and improves performance, it under-utilizes image information, leading to descriptions that may not accurately describe the images.

Despite the advancements in image captioning, current methods still struggle with challenges such as the one-to-one correspondence between generated descriptions and images, lack of domain-specific training, and under-utilization of image information, leading to mismatched descriptions and images.

In contrast to the aforementioned methods, our work incorporates image information twice during inference. Initially, we input the image information into the language model, allowing the model to generate multiple candidate sentences through random decoding. Subsequently, we calculate the similarity between the candidate sentences and the image, selecting the sentence that exhibits the highest similarity to the image as the final description. This approach maximizes the utilization of image information and resolves the issue of mismatched descriptions and images encountered in previous methods.

2.2. Contrastive Models

In recent years, several visual-language contrastive models have emerged, including CLIP [13], ALIGN [15], UniCL [16], and OpenCLIP [17]. These models have shown promising performance in zero-shot image classification and feature extraction for downstream tasks. For example, Clip2Video [18] utilizes contrastive learning for video–text retrieval tasks, while the object detection model introduced by Gu et al. [19] employs contrastive learning to detect objects with an open vocabulary. Khandelwal et al. [20] applied CLIP and contrastive learning to acquire visual and language knowledge for robot navigation tasks. Clip4Clip [21] utilizes contrastive learning for video clip retrieval and description, and Portillo-Quintero et al. [22] presented a video retrieval method based on CLIP. Shen et al. [23] have also explored the performance improvements achieved by CLIP contrastive learning in various visual-language tasks. However, it remains challenging to apply these models to complex tasks such as image captioning.

To address this challenge, we propose a method that combines unsupervised contrastive learning with a semantic-enhanced cross-modal fusion model to significantly improve the zero-shot performance of contrastive models in image captioning. CLIP consists of separate encoders for visual and textual information and leverages unsupervised contrastive loss trained on a large-scale image–text dataset. This training enables CLIP to establish a shared semantic space for visual and textual information. In our work, we utilize CLIP for image captioning, style image captioning, and story generation tasks, demonstrating the effectiveness and scalability of the semantically enhanced cross-modal fusion model.

2.3. Text Generation

Text generation is a significant task in natural language processing, attracting substantial research attention in recent years. Traditional methods for text generation often rely on language-model-based decoding approaches, which can be categorized into deterministic and stochastic methods. Deterministic methods, such as greedy search and beam search, select the most probable text based on the model's output probability. However, these methods often suffer from issues such as monotonicity [24] and degradation [25,26], which limit the diversity and creativity of the generated text. To address these limitations, stochastic methods have been introduced, including sampling-based techniques such as top-k sampling. These methods select multiple candidate texts with higher probabilities and perform random sampling within the set, thereby enhancing the diversity of the generated text.

In image captioning, the key challenge in text generation is effectively leveraging the abundant information from images to generate accurate and diverse descriptions.

Traditional approaches employ multimodal representation learning methods that aim to map image and text features into a shared semantic space, facilitating fusion and association between the modalities. However, these methods often encounter issues related to modality discrepancy and information inconsistency, leading to generated text that is not fully aligned with the images. To overcome these challenges, this paper proposes a novel semantic-enhanced cross-modal fusion model for image captioning.

In summary, despite substantial progress in image captioning, contrastive models, and text generation, existing methods face several challenges such as modality discrepancy, information inconsistency, a lack of one-to-one correspondence between descriptions and images, and difficulties in handling complex tasks such as zero-shot image captioning. These issues often lead to mismatched descriptions and images, lack of diversity in generated text, and poor performance on evaluation benchmarks. In contrast to these methods, our work aims to address the following tasks: (1) Maximize the utilization of image information to ensure that generated descriptions accurately correspond to the images. (2) Enhance the zero-shot performance of contrastive models in image captioning by combining unsupervised contrastive learning with a semantic-enhanced cross-modal fusion model. (3) Implement a novel semantic-enhanced cross-modal fusion model to overcome the challenges of modality discrepancy and information inconsistency, ensuring that generated text is fully aligned with the images.

Our proposed method is designed to resolve the aforementioned limitations by introducing novel techniques such as leveraging image information twice during inference and utilizing CLIP for various tasks. In achieving this, our work offers a significant advancement in the field of image captioning.

3. Method

In the pursuit of mitigating the challenges of the modality gap and the image–text mismatch problem, this paper puts forward a groundbreaking semantic-enhanced cross-modal fusion model (SCFM). This model is meticulously designed to address the inherent complexities in bridging the visual and textual modalities, offering a unique approach that distinguishes it from existing methods. The SCFM is built upon three carefully interwoven components that work in harmony to enhance the entire process of image captioning. As illustrated in Figure 1, these components are (1) a text semantic enhancement network: this part of the SCFM focuses on extracting and enhancing the semantic features within the text. It employs a combination of encoding techniques, enhancement strategies, and advanced network architectures to capture and emphasize the intricate semantic relationships within the text content. The enhancement of these features is pivotal in forming an initial semantic representation that resonates with the visual content. (2) Contrast learning: this component plays a central role in optimizing the consistency between texts by focusing on similarity measures. By employing a contrastive learning approach, it works in conjunction with the text semantic enhancement network to fine-tune the semantic representations. This joint operation enables a more precise alignment between visual and textual modalities, making this part essential for bridging the modality gap. (3) Image enhancement decoding strategy: tailoring the final stage of the model, this strategy emphasizes the generation of accurate and diverse captions. It not only adopts cutting-edge sampling techniques but also employs well-defined metrics to calculate the compatibility between textual descriptions and visual content. The intelligent integration with the preceding components ensures that the generated captions are not only relevant but also rich in diversity and accuracy.

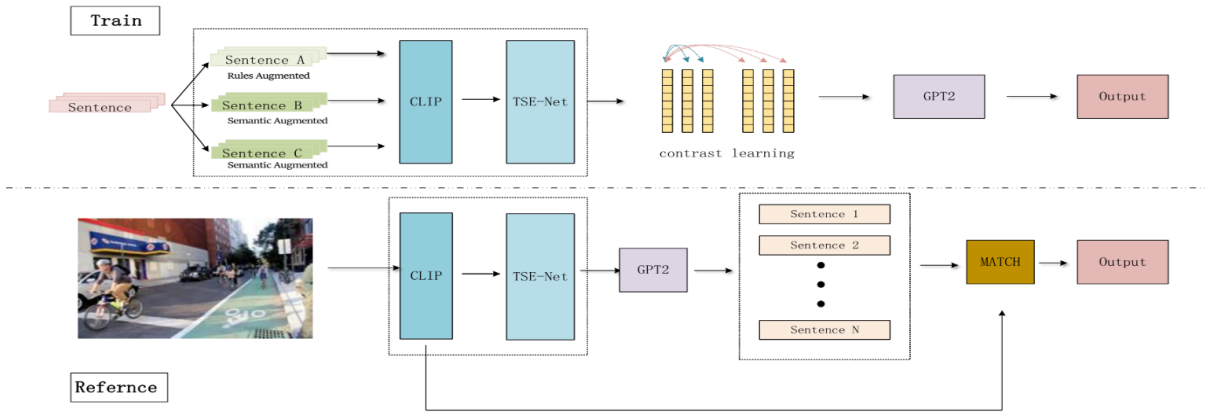

Figure 1. Overview of the structure of SCFM. The upper section illustrates the training part, showcasing the integration of a semantic-enhancement network and contrastive learning process. The lower section depicts the inference part, where our model employs a visual-enhancement decoding strategy.

The collaboration and interrelation of these three components form the crux of the SCFM, providing a seamless, integrative solution to the complex challenges of cross-modal learning. By delving into the architecture and methodology of each component, the following subsections will provide a clearer and more in-depth understanding of our innovative approach.

3.1. Text Semantic Enhancement Network

The text semantic enhancement network aims to extract enhanced semantic representations of text by capturing semantic associations, enhancing semantic features, and attenuating modal features. It consists of two main components: a text encoder and a semantic enhancement module. The text encoder maps the input text sequence into a semantic space, generating an initial semantic representation. To enhance the sentences, text enhancement techniques such as rule enhancement and semantic enhancements are applied. The CLIP network model is utilized for encoding the enhanced text sequence into a fixed-length vector representation. The semantic enhancement module incorporates key components including a multilayer perceptron (MLP), a residual network, atrous convolution, and a pooling layer. The MLP consists of multiple fully connected layers, enabling the extraction of higher-order features and semantic information. The introduction of residual connectivity helps to alleviate the issues of gradient disappearance and enhances the expressiveness of feature representation, thereby improving the quality of semantic representation. Atrous convolution, also known as dilated convolution, is employed to expand the effective receptive field of the convolutional kernel by introducing holes (dilation) in the kernel. This allows the network to capture a broader range of contextual information, enhancing its ability to understand and represent the semantics of text. Furthermore, a pooling layer is employed to reduce the spatial dimensionality of the feature map, facilitating more efficient processing. Through the combination of the multilayer perceptron, residual network, atrous convolution, and pooling layer, the semantic enhancement module effectively extracts and merges features to generate a more comprehensive and semantically rich representation of the input text.

3.2. Contrast Learning to Optimize Feature Representations

To enhance the similarity measurement and promote feature consistency between texts, we adopt a contrastive learning approach in our model. Contrastive learning aims to optimize feature representations by maximizing the similarity between positive sample pairs and minimizing the similarity between negative sample pairs. For each text sample,As illustrated in Figure 2, we randomly select a positive sample that belongs to the same category as the input text. Additionally, we choose multiple negative samples that differ from the input text in terms of category. The similarity between the input text and the

positive samples, as well as the similarity between the input text and the negative samples, is computed using cosine similarity. By maximizing the similarity of positive sample pairs, we encourage the semantic features to be more consistent. Conversely, minimizing the similarity of negative sample pairs reduces their semantic correlation. The loss function employed in our approach is designed to reinforce these objectives during training. The specific formulation of the loss function depends on the chosen contrastive learning framework and network architecture. It typically incorporates margin-based or contrastive loss terms, aiming to increase the similarity of positive pairs and decrease the similarity of negative pairs. By optimizing the contrastive loss, our model effectively learns discriminative and consistent semantic representations. This enables the model to bridge the modality gap, mitigate the image–text mismatch, and improve its capability to comprehend and express the semantics of text.

$$l_i^a = -\log \frac{\exp(s(y_i^a, y_i^b)/\alpha)}{\sum_{j=1}^{N}[\exp(s(y_i^a, y_j^a)/\alpha) + \exp(s(y_i^a, y_j^b)/\alpha)]} \tag{1}$$

where $i, j \in [1, N]$, $a, b \in \{1, 2, 3\}$. α is the temperature parameter. s is the similarity function; it is defined as follows:

$$s(u, v) = \frac{u^T v}{\| u \| \cdot \| v \|} \tag{2}$$

The total comparison loss is defined as follows:

$$L_{CL} = \frac{1}{3N} \sum_{i=1}^{N} (l_i^a + l_i^b + l_i^c) \tag{3}$$

where l_i^a is the contrast loss after rule enhancement for the i-th sample, l_i^b l_i^c is the contrast loss after semantic enhancement for the i-th sample, and N is the total number of samples.

Through the optimization process driven by contrastive learning, our model achieves a more consistent representation between texts. This enhanced representation enables the model to better understand and capture the relationships between texts, leading to improved accuracy and diversity in the generated caption results.

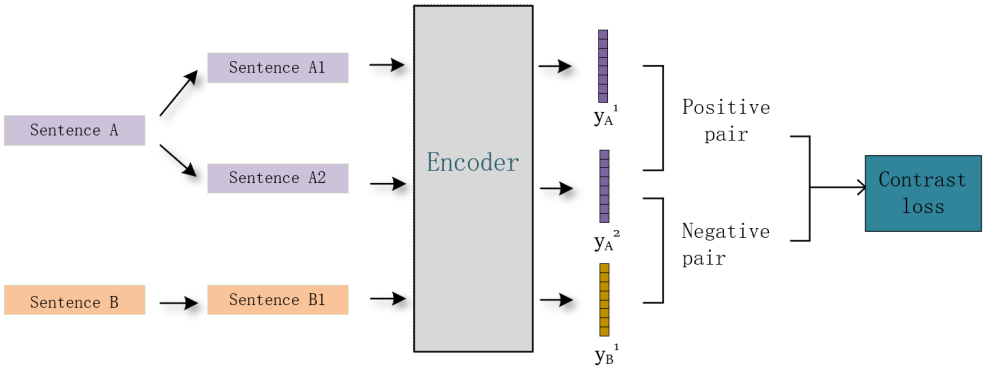

Figure 2. Overview of the contrastive learning approach to optimize feature representations. This approach maximizes the similarity between positive text pairs and minimizes it between negative pairs.

3.3. Enhanced Visual Selection Decoding

To generate accurate and diverse caption results, we propose an image enhancement decoding strategy. Departing from the traditional maximum probability decoding approach, we adopt a structured decoding process that leverages the rich information present in the images. Our strategy incorporates the use of the top-k sampling technique to generate a set of diverse candidate sentences. This step ensures the production of varied captions and avoids the generation of singular or repetitive results.

Next, we employ a defined metric to calculate the similarity between the candidate sentences and the image. By measuring the compatibility between textual descriptions and visual content, we select the sentence that exhibits the highest similarity to the image as the final output. This selection process guarantees the accuracy and visual relevance of the generated captions. For the input image I, the vector of image information in the shared semantic space is y_I, Then, the set of generated candidate sentences is Q:

$$\begin{aligned} Q^{(n)} &= \{Q_1, Q_2, ..., Q_n\} \\ &= \left\{ p_\theta\left(\overset{(1)}{y_1} + \varepsilon\right), p_\theta\left(\overset{(2)}{y_1} + \varepsilon\right), \ldots, p_\theta\left(\overset{(n)}{y_1} + \varepsilon\right) \right\} \end{aligned} \tag{4}$$

where y_i refers to the mapping vector of the input image I in the shared space. ε is the value of Gaussian noise, p_θ is the decoding strategy, and the obtained $Q_1, Q_2, \ldots, Q_n$ are the top n sentences from the top-k sampling.

The sentences in the candidate sentence set are mapped to the low-dimensional shared semantic space, and the sentence with the highest similarity to the image is searched. The definition of similarity is shown in (3). The final result, Q_{best}, is defined as:

$$Q_{best} = \arg \max_{q \in Q^{(n)}} s(y_I, q) \tag{5}$$

where y_I is the mapping vector of image I, and q is the candidate sentence.

Through the utilization of our image enhancement decoding strategy, we aim to enhance the quality and diversity of the generated captions. By considering the inherent information within the images and selecting the most appropriate textual descriptions, our approach ensures the production of accurate and diverse results that effectively convey the content of the images.

4. Experiment and Analysis

To thoroughly validate the effectiveness of the proposed SCFM method, extensive experiments were conducted in this paper on two publicly available datasets, MS COCO [27] and Flickr30k [28]. The experimental results were analyzed from both quantitative and qualitative perspectives, and a comparative experiment was conducted between SCFM and current state-of-the-art image captioning algorithms. This section begins by introducing the evaluation benchmarks, experimental details, and evaluation metrics. Subsequently, a detailed presentation and analysis of the experimental results are provided.

4.1. Experimental Setup

4.1.1. Evaluation Benchmark

We conduct experiments on four widely used benchmarks. MS COCO [27], Flickr30k [28], FlickrStyle10K [29] and SentiCap [30]. MS COCO, comprising over 120,000 images from diverse everyday scenes, served as a comprehensive benchmark to assess our method's accuracy and diversity. Each image in MS COCO is annotated with at least five captions. Flickr30k, focusing on human–object interactions, includes around 30,000 images, each provided with five different captions, allowing us to evaluate the model's understanding of human activities. The FlickrStyle10K dataset, containing 10,000 artistically stylized images, was selected to test our method's adaptability to unique visual expressions. Finally, SentiCap, consisting of sentiment-annotated images, allowed us to assess how well our approach generates captions aligned with emotional content. The combined use of these datasets, with their varying content, format, and thematic focus, ensures a robust and multifaceted evaluation of the SCFM method.

4.1.2. Implementation Details

The experiments were conducted using the PyCharm software, version 2021 (JetBrains, Prague, Czech Republic), as the primary development environment. The implementation

was executed on a system equipped with an NVIDIA 3080Ti graphics card (NVIDIA Corporation, Santa Clara, CA, USA). These tools were selected due to their robust performance and wide acceptance within the research community, ensuring consistency and reliability throughout our experimental process. We utilized a frozen pre-trained Vit-B/32 CLIP model as the image encoder, paired with the GPT2 language model as the text decoder. To optimize the model parameters, the Adam optimizer was applied with a learning rate of 2e-5. This combination of architecture and optimization is in alignment with state-of-the-art practices in the field and proved effective for our specific experimentation. Our experimentation evaluated the training time and inference speed on two widely used datasets, MS COCO and Flickr30k, using the aforementioned 3080Ti GPU. The training and inference time are subject to variation based on the specific hyperparameter configurations employed. In the context of the parameters delineated within this manuscript, training on the MS COCO dataset required 1 h per epoch, with a total of 10 epochs sufficient for convergence. Training on the Flickr30k dataset was notably quicker, necessitating only 14 min per epoch for a similar total of 10 epochs. In terms of inference, the time required to process 1000 images was recorded at 33 min. These metrics underscore the model's efficiency and suitability for practical applications.

4.1.3. Evaluation Metrics

Following the common practice in the literature, we perform evaluation using BLEU-1 (B@1), BLEU-4 (B@4) [31], METEOR (M) [32], ROUGE-L (R-L) [33], CIDEr [34], and SPICE [35]. BLEU-1 (B@1) and BLEU-4 (B@4) measure the precision of 1 gram and up to 4 grams, respectively, and are typically expressed as percentages ranging from 0 to 100%. METEOR, ranging from 0 to 1, provides a balanced assessment of precision and recall, taking into account synonymy, stemming, and paraphrasing. ROUGE-L, another unitless metric ranging from 0 to 1, quantifies the overlap of the longest common subsequence. CIDEr, typically ranging from 0 to 10 or higher, evaluates the consensus between human captions and generated descriptions. Lastly, SPICE, ranging from 0 to 1, assesses semantic propositional content. It is important to note that in the academic literature, and to facilitate more intuitive understanding, authors often multiply these values by 100, thereby converting them to a percentage, and present them as two-digit numbers. This practice aligns the results with a common convention that many readers will be familiar with.

4.2. Standard Image Captioning

To validate the performance of our model, extensive experiments were conducted on the MS COCO and Flickr30k datasets, and the experimental results are presented in Table 1. Initially, we evaluated models employing fully supervised techniques, including BUTD [36], UniVLP [37], and Clip-Cap [38]. As anticipated, these models leveraged additional supervised training on image–text pairs, thereby demanding greater computational resources and training time, consequently exhibiting slightly superior performance compared to our approach. Nevertheless, in comparison to unsupervised methods such as MAGIC [5], ZeroCap [4], and CapDec [7], our method yielded higher scores on the BLEU-4 (B@4) [31], METEOR (M) [32], ROUGE-L (R-L) [33], and CIDEr [34] metrics. These findings illustrate the superiority of our approach over other unsupervised techniques, highlighting its enhanced capability in the context of image captioning tasks. Figure 3 illustrates a visual comparison between our proposed method and three other zero-shot methods. The comparison clearly demonstrates that our approach enables more accurate and vivid descriptions of the main content in the images. Let us examine each example in detail. In Figure 3a, ZeroCap correctly identifies the prominent object "cyclist" but introduces the non-existent object "stop sign," resulting in an erroneous description. Additionally, MAGIC focuses on describing the "street crossing sign" but deviates from the emphasis presented in the original image and contains grammatical errors. On the other hand, CapDec describes a bicycle leaning against the roadside, which does exist in the top-right corner of the image. However, the CapDec model excessively focuses on details while

neglecting the overall overview of the image. In contrast, our model accurately captures the image theme of a group of people riding bicycles and precisely expresses the environmental context and human actions depicted in the image. Furthermore, considering Figure 3c,d, it is evident that both the ZeroCap and MAGIC models exhibit varying degrees of errors in their descriptions. Although CapDec provides a relatively accurate description of the image content, it fails to identify that Figure 3c is a photograph and misclassifies Figure 2d as a photograph. In contrast, our model accurately recognizes the people and scenes depicted in Figure 3c and correctly identifies it as a photograph. These visual comparisons further illustrate the superiority of our proposed method in image captioning tasks. Our model not only improves description accuracy and diversity but also excels in recognizing image content and providing contextually appropriate descriptions.

Table 1. Image captioning results on MS COCO and Flickr30k.The best result is **bold**.

Model	MS COCO					Flickr30k				
	B@1	**B@4**	**M**	**R-L**	**CIDEr**	**B@1**	**B@4**	**M**	**R-L**	**CIDEr**
Fully Supervised Approaches										
BUTD	77.2	36.2	27.0	56.4	113.5	-	27.3	21.7	-	56.6
UniVLP	-	36.5	28.4	-	116.9	-	30.1	23.0	-	67.4
ClipCap	74.7	33.5	27.5	-	113.1	-	21.7	22.1	47.3	53.5
Weakly or Unsupervised Approachs										
ZeroCap	49.8	7.0	15.4	31.8	34.5	44.7	5.4	11.8	27.3	16.8
MAGIC	56.8	12.9	17.4	39.9	49.3	44.5	6.4	13.1	31.6	20.4
CapDec	**69.2**	26.4	25.1	51.8	91.8	55.5	17.7	20.0	43.9	39.1
SCFM	69.0	**27.3**	**25.8**	**52.7**	**94.3**	55.2	**17.9**	**20.2**	**44.5**	**41.2**

(**a**)

ZeroCap: Stop sign with a cyclist on the street.
MAGIC: A street crossing sign shows the right for bicyclists.
CapDec: A bicycle is propped up on a sidewalk next to a street.
SCFM: A group of people riding bikes down a street next to a street.

(**b**)

ZeroCap: Sheep in a park with a fence and a hill in the background.
MAGIC: A dog carrying a racquet on a grassy field.
CapDec: A sheep that is standing in the grass.
SCFM: A sheep that is looking at the camera in a field.

(**c**)

ZeroCap: Couple with a framed and a teddy bear in a window.
MAGIC: A table covered with pictures of the outside world.
CapDec: A bed with a white bedspread and a mirror above it.
SCFM: A black and white photo of a man and woman in a bed.

(**d**)

ZeroCap: Bicycle thief on a street with a camera in the process.
MAGIC: A skateboarder doing a trick on a ramp.
CapDec: A black and white photo of a man on a skateboard.
SCFM: A man riding a bicycle next to a wall.

Figure 3. Examples of standard image captioning.

4.3. Cross-Domain Image Captioning

To further evaluate the generation capability of our model, we conducted cross-domain experiments. Cross-domain experiments involve training the language model on one dataset (e.g., MS COCO) and testing our model on another target dataset (e.g., Flickr30k). The experimental results are presented in Table 2. We trained MAGIC, CapDec, and our model on the MS COCO dataset and tested them on the Flickr30k dataset. The results demonstrate that our model outperforms the other models in terms of captioning performance. Additionally, we trained our model on the Flickr30k dataset and tested it on the MS COCO dataset. The results indicate that our model possesses strong generalization capability and robustness, as it achieves competitive performance even when applied to a different dataset. These experimental findings further validate the robustness and generalization ability of our model, highlighting its excellent performance across different datasets. The results of the cross-domain experiments demonstrate the wide applicability of our model and reinforce its reliability and practicality in various real-world scenarios.

Table 2. Cross-domain evaluation. X ==>Y means source domain ==>target domain. The best result is **bold**.

Model	Flickr30k ==>MS COCO					MS COCO ==>Flickr30				
	B@1	**B@4**	**M**	**R-L**	**CIDEr**	**B@1**	**B@4**	**M**	**R-L**	**CIDEr**
MAGIC	41.4	5.2	12.5	30.7	18.3	46.4	6.2	12.2	31.3	17.5
CapDec	**43.3**	9.2	16.3	36.7	27.3	**60.2**	17.3	18.6	42.7	35.7
SCFM	43.0	**9.3**	**17.0**	**37.1**	**28.5**	59.6	**17.5**	**19.9**	**43.6**	**38.0**

4.4. Stylized Image Captioning

To validate the model's performance in terms of descriptive ability, we conducted stylized image captioning experiments. Stylized image captioning aims to generate sentences that accurately describe images while incorporating specific styles. Collecting image–text pairs with different language styles is a challenging task. However, collecting text with different styles is relatively straightforward. Our model only requires text data with different styles to achieve stylized image descriptions. In the experiment, we utilized the FlickrStyle10K [29] and SentiCap [30] datasets and present the experimental results in Figure 4. Through stylized image captioning experiments, we can visually observe the performance of our model in generating image descriptions with specific styles. The figure displays examples of images with different styles and the corresponding stylized descriptions generated by our model. The results demonstrate that our model can generate image descriptions that conform to the corresponding style requirements based on different style text data. This further validates the model's capabilities in semantic understanding and language generation. Through the stylized image captioning experiments, we verify the flexibility and adaptability of our model, showcasing its strong performance in handling image captioning tasks with diverse styles. This provides broader prospects for the application of our method in various scenarios.

4.5. Ablation Experiments

To investigate the impact of various factors on the generation of image captions, we conducted a series of ablation experiments, the results of which are detailed in Table 3. Firstly, we introduced the text semantic enhancement network to the baseline model while keeping other components unchanged. We observed improvements in performance across all metrics, including a significant increase from 68.1% to 68.6% in BLEU-1 and from 26.4% to 27.3% in BLEU-4. This signifies that the model acquires a more comprehensive understanding of the image by leveraging multiple semantic perspectives, resulting in the generation of richer and more expressive descriptive sentences.

Romantic: A baby with a toothbrush in its mouth is enjoying the happiness of childhood
Humorous: A little boy with a toothbrush in his mouth to eat
Positive: A happy child brushes his teeth with a blue toothbrush
Negative: A little kid is brushing his teeth with a blue and white toothbrush

Romantic: Two giraffes are standing in a grassy area, enjoying the beauty of nature
Humorous: Two giraffes are standing in a grassy area to search for bones
Positive: Two giraffes stand next to each other in a grassy area
Negative: Two giraffes are standing in the dead grass near a werd tree

Romantic: A variety of vegetables are laid out on a kitchen counter to make a delicious meal
Humorous: A bunch of vegetables are sitting on a cutting board ready to be eaten
Positive: A great variety of vegetables on a cutting board on a counter top
Negative: A variety of vegetables on a kitchen counter top

Romantic: A woman in a white dress is cutting a cake for her husband to enjoy
Humorous: A bride and groom are cutting their wedding cake at the reception
Positive: A bride and groom are cutting a delicious cake
Negative: A bride and a groom are cutting their wedding cake

Figure 4. Examples of stylized image captioning.

Table 3. Ablation experiments.The best result is **bold**.

Method	B@1	B@4	M	R-L	CIDEr
Baseline	68.1	26.4	25.1	51.0	90.9
+TSE-Net	68.6	27.3	25.3	51.6	92.3
+contrast learning	68.4	26.8	25.0	51.5	91.9
+EVSD	68.8	26.9	25.7	52.1	93.7
SCFM	**69.0**	**27.3**	**25.8**	**52.7**	**94.3**

Subsequently, the introduction of contrast learning resulted in consistent improvements, such as an increase from 68.1% to 68.4% in BLEU-1 and from 26.4% to 26.8% in BLEU-4. This illustrates the efficacy of contrast learning in fine-tuning semantic representations. The experimental results revealed consistent score improvements across all evaluation metrics compared to the baseline model. Furthermore, we explored the adoption of an image enhancement decoding strategy, replacing the conventional maximum probability decoding strategy in the baseline model. The experimental findings indicated a substantial 2.9% improvement in the CIDEr metric (from 90.9% to 93.7%) when utilizing the image enhancement decoding strategy. This demonstrates the efficacy of the strategy in enhancing the precision of generated textual descriptions, enabling the model to produce more accurate and contextually aligned sentences that effectively depict the content of the image. To investigate the synergistic effects of these approaches, we performed comprehensive improvements on each component of the baseline model. The experimental results exhibited significant advancements across all evaluation metrics. In comparison to the baseline model, we observed a remarkable 3.2% increase in the BLEU-4 metric (from 26.4% to 27.3%) and a notable 3.6% (from 90.9% to 94.3%) improvement in the CIDEr metric. These outcomes provide further evidence of the complementary advantages and collaborative effects of the proposed strategies.

In conclusion, our proposed text semantic enhancement network and image enhancement decoding strategy present considerable advantages in the task of image captioning.

By leveraging the text semantic enhancement network, contrastive learning, and image enhancement decoding strategy in concert, our model exhibits enhanced capability in generating more expressive, precise, and contextually coherent descriptive sentences. Thus, our approach offers a robust and effective methodology for addressing the complexities associated with image captioning and similar challenging tasks.

4.6. Heatmap Analysis

To provide a more intuitive demonstration of the impact of the text semantic enhancement network and the contrastive learning training method employed during the training phase, we generated heatmaps that depict the differences between the textual feature vectors obtained by the baseline model and the SCFM model, as illustrated in Figure 5. These heatmaps utilize a color mapping scheme to represent the similarity or correlation between features. Upon examining the heatmaps of the baseline model's features, as illustrated in Figure 5a, we observed a relatively high degree of similarity among different features, indicating some overlap in capturing semantic information relevant to the image captioning task. However, such overlap may lead to generated descriptions lacking diversity and richness. In contrast, the heatmaps of the features obtained after applying our model exhibited significantly increased differences among the various features, as illustrated in Figure 5b. This suggests that the text semantic enhancement network successfully extracted unique information pertaining to visual and linguistic features that are relevant to the image captioning task, thereby enabling the generation of more diverse and expressive descriptions. By enhancing the differences among features, our approach comprehensively captures the semantic relationships between images and text, resulting in improved quality and diversity of generated descriptions. Through the comparison of these two sets of feature heatmaps, the efficacy of the text semantic enhancement network in enriching feature representations can be visually observed. Overall, the heatmaps provide visual evidence of the positive impact of the text semantic enhancement network and contrastive learning on the image captioning process. The differences in the heatmaps highlight the ability of our model to capture distinct and relevant information, leading to more diverse and expressive image captions. This analysis further supports the effectiveness of our approach in enhancing the generation of descriptive sentences that accurately correspond to the content of the images.

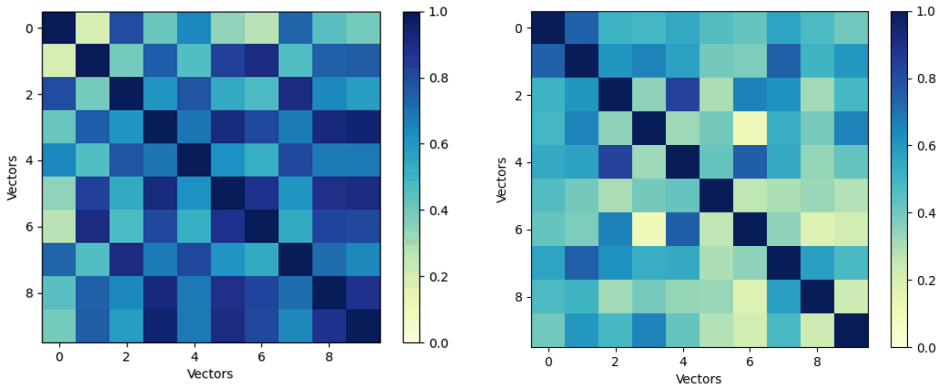

Figure 5. Heatmap of textual feature vectors using (**a**) the baseline model, showing high similarity and potential overlap in semantic information and textual feature vectors, and using (**b**) the SCFM model, highlighting increased differentiation and successful extraction of diverse visual and linguistic features.

5. Result and Discussion

In the realm of weakly or unsupervised image captioning methods, the comparison of our semantic-enhanced cross-modal fusion model (SCFM) with other contemporary models offers crucial insights. The comparison is grounded in experiments conducted

on the MS COCO and Flickr30k datasets, utilizing standard metrics such as BLEU-1, BLEU-4, METEOR, ROUGE-L, and CIDEr (Table 1). ZeroCap's utilization of zero-shot techniques renders a performance that is appreciably surpassed by SCFM, evidenced by B@1 and CIDEr scores of 49.8 and 34.5 against SCFM's 69.0 and 94.3, respectively, on MS COCO. The marked enhancement accentuates the potency of our model's semantic and decoding strategies. The performance of MAGIC, another weakly supervised model, is somewhat superior to ZeroCap but still falls behind SCFM. With B@1 and CIDEr scores of 56.8 and 49.3, it underscores SCFM's advanced capabilities in both granular and overall performance measures. Comparatively, CapDec exhibits a competitive approach, with B@1 and CIDEr scores of 69.2 and 91.8 on MS COCO. Nonetheless, SCFM slightly surpasses CapDec in key domains, with B@4 and CIDEr scores of 27.3 and 94.3 versus 26.4 and 91.8, respectively, cementing the relative strength of our approach. An examination of SCFM through ablation studies further elucidates the individual impacts of its components (Table 3). The introduction of the TSE-Net led to an elevation in B@1 and CIDEr to 68.6 and 92.3. The implementation of contrast learning yielded a nuanced performance shift, while the incorporation of EVSD boosted B@1 and CIDEr to 68.8 and 93.7. The full SCFM model combined these incremental enhancements, solidifying the B@1 at 69.0 and CIDEr at 94.3.

In the field of image captioning, our semantic-enhanced cross-modal fusion model carves out its distinctive niche when juxtaposed with prominent methods such as ZeroCap, MAGIC, and CapDec. While ZeroCap showcases innovation by combining CLIP with GPT-2 for zero-shot captioning and pivoting to a novel data source, our approach delves deeper, emphasizing a comprehensive strategy to address the modality gap and image–text mismatches. This commitment to excellence is further evidenced as we not only forge a closer image–text alignment but also overshadow ZeroCap's key performance metrics, underscoring our capability for a refined captioning control. Diverging from MAGIC's zero-shot approach, our model is grounded in a bespoke design that marries a cutting-edge semantic enhancement network with a heightened decoding strategy. This amalgamation not only propels innovation but also triumphs in crucial metrics such as BLEU-4, METEOR, ROUGE-L, and CIDEr. Meanwhile, in comparison to CapDec, known for its decoding of image embeddings supplemented by noise-injection as a bridge over the image–text domain chasm, our methodology leans heavily on semantic fortification and methodical decoding. This deliberate emphasis equips us with the tools to tackle age-old challenges, ranging from inference bias to output degradation. By prioritizing the symbiotic relationship between textual and visual information over mere noise injections, our approach marks a significant leap in performance, reinforcing its unique stature in image captioning research.

6. Conclusions

This paper presents an image caption generation method based on image-enhanced decoding, aiming to improve the quality and diversity of generated descriptions. Our approach incorporates a text semantic enhancement network and contrastive learning to enhance feature vectors. Furthermore, we utilize an image-enhanced decoding strategy to strengthen the correlation between generated text and images. By adopting these techniques, we have made significant advancements over traditional image caption generation models. To evaluate our proposed method, we conduct experiments on two widely used datasets and employ multiple evaluation metrics. The experimental results clearly demonstrate remarkable improvements in both the accuracy and diversity of generated descriptions compared to baseline models. These findings validate the effectiveness of our approach in generating more precise and varied captions. In addition, we perform cross-domain experiments and style-based image caption experiments to assess the generalization ability and adaptability of our model across different application scenarios. The results of these experiments further confirm the versatility and scalability of our approach. While our approach has demonstrated promising results, it is important to acknowledge potential limitations and challenges. For example, the effectiveness of the text semantic enhancement network and contrastive learning might vary across different image types and

domains, necessitating fine-tuning for specific applications. There may also be computational challenges related to scaling the image-enhanced decoding strategy for large datasets. Moreover, understanding and interpreting the complex interactions between image and text features may remain an open problem, warranting further investigation. Future work should also consider potential biases in the training data that could affect the model's generalizability across diverse contexts. By recognizing and addressing these challenges, we hope to inspire future research to build upon and refine our proposed method.

Author Contributions: Conceptualization, N.X. and L.C.; methodology, N.X. and L.C.; software, L.C. and L.L.; validation, N.X., L.C. and Z.G.; formal analysis, N.X. and L.C.; investigation, L.L.; resources, N.X.; data curation, X.R.; writing—original draft preparation, L.C.; writing—review and editing, N.X., L.L. and Z.G.; visualization, X.R.; supervision, N.X.; project administration, N.X.; funding acquisition, N.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by (a) the Natural Science Foundation of Chongqing Province of China, grant number CSTB2022NSCQ-MSX0786; (b) China Postdoctoral Science Foundation (Certificate Number: 2023M733358); (c) Science and Technology Research Project of Chongqing Education Commission, grant numbers KJQN202001118, KJQN202201109.

Data Availability Statement: The publicly available MS COCO, Flickr30k, FlickrStyle10K, and SentiCap datasets were analyzed in this study. These datasets can be found here: MS COCO and Flickr30k: https://www.kaggle.com/datasets/shtvkumar/karpathy-splits (accessed on 23 July 2022) FlickrStyle10K: https://zhegan27.github.io/Papers/FlickrStyle_v0.9.zip (accessed on 26 July 2022) SentiCap: https://www.kaggle.com/datasets/prathamsaraf1389/senticap (accessed on 26 July 2022).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Karpathy, A.; Li, F.-F. Deep visual-semantic alignments for generating image descriptions. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 3128–3137.
2. Laina, I.; Rupprecht, C.; Navab, N. Towards unsupervised image captioning with shared multimodal embeddings. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 7414–7424.
3. Feng, Y.; Ma, L.; Liu, W.; Luo, J. Unsupervised image captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 4125–4134.
4. Tewel, Y.; Shalev, Y.; Schwartz, I.; Wolf, L. Zero-shot image-to-text generation for visual-semantic arithmetic. *arXiv* **2021**, arXiv:2111.14447.
5. Lan, T.; Liu, Y.; Liu, F.; Yogatama, D.; Wang, Y.; Kong, L.; Collier, N. Language models can see: Plugging visual controls in text generation. *arXiv* **2022**, arXiv:2205.02655.
6. Liang, V.W.; Zhang, Y.; Kwon, Y.; Yeung, S.; Zou, J.Y. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 17612–17625.
7. Nukrai, D.; Mokady, R.; Globerson, A. Text-Only Training for Image Captioning using Noise-Injected CLIP. *arXiv* **2022**, arXiv:2211.00575.
8. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and tell: A neural image caption generator. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3156–3164.
9. Mao, J.; Xu, W.; Yang, Y.; Wang, J.; Yuille, A.L. Explain images with multimodal recurrent neural networks. *arXiv* **2014**, arXiv:1410.1090.
10. Yu, Z.; Yu, J.; Cui, Y.; Tao, D.; Tian, Q. Deep modular co-attention networks for visual question answering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6281–6290.
11. Huang, L.; Wang, W.; Chen, J.; Wei, X.-Y. Attention on attention for image captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 4634–4643.
12. Pan, Y.; Yao, T.; Li, Y.; Mei, T. X-linear attention networks for image captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10971–10980.
13. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clar, J. Learning transferable visual models from natural language supervision. In Proceedings of the 38th International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 8748–8763.
14. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language models are unsupervised multitask learners. *OpenAI Blog* **2019**, *1*, 9.

15. Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In Proceedings of the 38th International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 4904–4916.
16. Yang, J.; Li, C.; Zhang, P.; Xiao, B.; Liu, C.; Yuan, L.; Gao, J. Unified contrastive learning in image-text-label space. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 19163–19173.
17. Cherti, M.; Beaumont, R.; Wightman, R.; Wortsman, M.; Ilharco, G.; Gordon, C.; Schuhmann, C.; Schmidt, L.; Jitsev, J. Reproducible scaling laws for contrastive language-image learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 2818–2829.
18. Fang, H.; Xiong, P.; Xu, L.; Chen, Y.; Clip2video: Mastering video-text retrieval via image clip. *arXiv* **2021**, arXiv:2106.11097.
19. Gu, X.; Lin, T.-Y.; Kuo, W.; Cui, Y.; Open-vocabulary object detection via vision and language knowledge distillation. *arXiv* **2021**, arXiv:2104.13921.
20. Khandelwal, A.; Weihs, L.; Mottaghi, R.; Kembhavi, A. Simple but effective: Clip embeddings for embodied AI. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 14829–14838.
21. Luo, H.; Ji, L.; Zhong, M.; Chen, Y.; Lei, W.; Duan, N.; Li, T. CLIP4Clip: An empirical study of CLIP for end to end video clip retrieval and captioning. *Neurocomputing* **2022** *508*, 293–304.
22. Portillo-Quintero, J.A.; Ortiz-Bayliss, J.C.; Terashima-Marín, H. A straightforward framework for video retrieval using clip. In Proceedings of the MCPR 2021—Pattern Recognition: 13th Mexican Conference, Mexico City, Mexico, 23–26 June 2021; Springer: New York, NY, USA, 2021; pp. 3–12. [CrossRef]
23. Shen, S.; Li, L.H.; Tan, H.; Bansal, M.; Rohrbach, A.; Chang, K.-W.; Yao, Z.; Keutzer, K. How much can clip benefit vision-and-language tasks? *arXiv* **2021**, arXiv:2107.06383.
24. Li, J.; Galley, M.; Brockett, C.; Gao, J.; Dolan, B. A diversity-promoting objective function for neural conversation models. *arXiv* **2015**, arXiv:1510.03055.
25. Fan, A.; Lewis, M.; Dauphin, Y. Hierarchical neural story generation. *arXiv* **2018**, arXiv:1805.04833.
26. Holtzman, A.; Buys, J.; Du, L.; Forbes, M.; Choi, Y. The curious case of neural text degeneration. *arXiv* **2019**, arXiv:1904.09751.
27. Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In *Proceedings, Part V 13, Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014*; Springer: New York, NY, USA, 2014; pp. 740–755.
28. Plummer, B.A.; Wang, L.; Cervantes, C.M.; Caicedo, J.C.; Hockenmaier, J.; Lazebnik, S. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2641–2649.
29. Gan, C.; Gan, Z.; He, X.; Gao, J.; Deng, L. Stylenet: Generating attractive visual captions with styles. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3137–3146.
30. Mathews, A.; Xie, L.; He, X. Senticap: Generating image descriptions with sentiments. In Proceedings of the AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA 12–17 February 2016; Volume 30, p. 1.
31. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.-J. Bleu: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadephia, PA, USA, 6–12 July 2002; pp. 311–318.
32. Denkowski, M.; Lavie, A. Meteor universal: Language specific translation evaluation for any target language. In Proceedings of the Ninth Workshop on Statistical Machine Translation, Baltimore, MD, USA, 26–27 June 2014; pp. 376–380.
33. Lin, C.-Y.; Och, F.J. 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), Barcelona, Spain, 21–26 July 2004; pp. 605–612.
34. Vedantam, R.; Zitnick, C.L.; Parikh, D. Cider: Consensus-based image description evaluation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4566–4575.
35. Anderson, P.; Fernando, B.; Johnson, M.; Gould, S. Spice: Semantic propositional image caption evaluation. In *Proceedings, Part V 14, Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016*; Springer: New York, NY, USA, 2016; pp. 382–398.
36. Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; Zhang, L. 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, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6077–6086.
37. Zhou, L.; Palangi, H.; Zhang, L.; Hu, H.; Corso, J.; Gao, J.; Unified vision-language pre-training for image captioning and vqa. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 13041–13049.
38. Mokady, R.; Hertz, A.; Bermano, A.H. Clipcap: Clip prefix for image captioning. *arXiv* **2021**, arXiv:2111.09734.