electronics

MDPI

Article

DecoupleCLIP: A Novel Cross-Modality Decouple Model for Painting Captioning

Mingliang Zhang, Xia Hou *, Yujing Yan and Meng Sun

Computer School, Beijing Information Science and Technology University, Beijing 102206, China; hp1039@163.com (M.Z.); 2022020605@bistu.edu.cn (Y.Y.); 2021020663@bistu.edu.cn (M.S.)
* Correspondence: houxia@bistu.edu.cn

Abstract: Image captioning aims to describe the content in an image, which plays a critical role in image understanding. Existing methods tend to generate the text for more distinct natural images. These models can not be well for paintings containing more abstract meaning due to the limitation of objective parsing without related knowledge. To alleviate, we propose a novel cross-modality decouple model to generate the objective and subjective parsing separately. Concretely, we propose to encode both subjective semantic and implied knowledge contained in the paintings. The key point of our framework is decoupled CLIP-based branches (DecoupleCLIP). For the objective caption branch, we utilize the CLIP model as the global feature extractor and construct a feature fusion module for global clues. Based on the objective caption branch structure, we add a multimodal fusion module called the artistic conception branch. In this way, the objective captions can constrain artistic conception content. We conduct extensive experiments to demonstrate our DecoupleCLIP's superior ability over our new dataset. Our model achieves nearly 2% improvement over other comparison models on CIDEr.

Keywords: painting captioning; multimodal fusion; CLIP

Citation: Zhang, M.; Hou, X.; Yan, Y.; Sun, M. DecoupleCLIP: A Novel Cross-Modality Decouple Model for Painting Captioning. *Electronics* **2024**, *13*, 4207. https://doi.org/10.3390/electronics13214207

Academic Editors: Silvia Liberata Ullo and Li Zhang

Received: 8 September 2024
Revised: 9 October 2024
Accepted: 21 October 2024
Published: 27 October 2024

1. Introduction

Painting understanding is critical in both public culture broadcasting and appreciation. Typically, a painting conveys not only objective meanings but also its inner essence, such as energy, life force, and spirit. However, paintings can be challenging to comprehend, especially for those lacking historical and artistic knowledge.

With the advancement of Artificial Intelligence (AI), the task of image captioning, which generates descriptions of given images, has emerged to aid in image appreciation and research. Many studies focusing on natural image captioning [1–7] have made significant strides. Existing image captioning models could be classified into two streams: two-stage image captioning models [8–11] and end-to-end image captioning models [12–16]. The two-stage image captioning models used the object detection module to obtain object regions and then describe the images based on these objects, regions, and image features. The end-to-end image captioning models remove the object detection module and generate the caption directly from the image features. These approaches excel in generating objective descriptions but often struggle to capture the deeper connotations within images. There are also some works that focus on painting captioning, constructing datasets [17–20], objectively describing paintings [21,22], etc. Current painting captioning faces two additional challenges. Firstly, painting captioning must generate not only objective descriptions but also artistic conceptions. The latter is often implicitly expressed in paintings, requiring the model to interpret the painter's intended meaning based on the scene. Therefore, painting captioning is more complex than natural image captioning. Secondly, it is challenging to effectively generate these two different types of descriptions within a unified model, as they require different descriptive approaches. To address these challenges, we propose a

decoupled painting captioning framework that generates objective captions and artistic conception captions separately using distinct modules.

In this paper, we introduce the DecoupleCLIP model to enhance painting captioning performance and provide comprehensive descriptions of paintings. Our approach is based on the premise that employing specialized models for each caption type is more effective than using a single unified model. The DecoupleCLIP model utilizes a dual-branch structure with individual loss functions for each branch. The first branch employs PureT [2] and CLIP [23] as image encoders, followed by a customized fusion module to integrate their output features. Transformers are then used as language encoders and decoders to generate objective captions. The second branch incorporates a specialized multimodal fusion module to combine the objective caption generated by the first branch, aiding in the generation of artistic conception captions. Finally, the objective caption and artistic conception caption are concatenated to form the final caption. Our main contributions are summarized as follows:

- We decouple painting captions into two aspects: objective and artistic conception. To achieve this, we propose a network structure with two branches that incorporate local to global feature fusion and multimodal fusion.
- We develop a multimodal fusion model that integrates objective caption text with multiple scales of image visual features.
- We create a small-scale image captioning dataset comprising both Chinese and Western paintings, conducting extensive experiments using this dataset.

2. Related Work

2.1. Image Captioning

Image captioning is critical for computer vision. Some works focus on end-to-end image captioning structures. Vinyals et al. [1] first combined CNN and LSTM to image caption task. Lu et al. [24] further proposed adaptive attention, which could decide whether to attend to the image and where. However, this work did not consider the effect of subsequent words on the whole sentence generation. Ge et al. [12] further proposed a Mutual-aid network with Bi-LSTM (MaBi-LSTM) to capture more contextual data and implicitly utilized the subsequent semantic information. However, this method lacks rich semantic representation of images to capture fine-grained visual features. In order to improve the compatibility of multi-modal information, Zhang et al. [13] integrated the part-of-speech information into the image captioning model. More recently, Prudviraj et al. [14] proposed an Attentive Contextual Network (ACN), which added an ACN module at the last of the ResNet backbone and added a Deformable Network [15] in the middle of the backbone. However, this method could not locate multi-scale semantic consistency regions in the image perfectly. To address this problem, Yu et al. [16] proposed a pyramid attention model with a ResNet backbone to obtain more bottom-up informative attention features.

Benefitting from the transformer's high performance in the NLP domain, some works used Transformers instead of LSTM as a language decoder to generate captions. Most previous studies used flattened operations to process images which lost image spatial information. To address this problem, Zhang et al. [25] proposed a Grid-Augmented (GA) module and an Adaptive Attention (AA) module and fused them with Transformers. This model could generate more fine-grained captions but had no interaction between vision and language. More recently, Wang et al. [2] proposed a pure Transformer-based model (PureT), which used Swin Transformer [26] as a visual feature extractor. At the same time, the model used Transformers as language encoders and decoders, which contained a pre-fusion module to increase the interaction between vision and language.

However, the above solution to the image captioning problem is still limited in how to generate painting captioning of high quality. Hence, we aimed to solve this challenge. We proposed a two-branch painting captioning structure to solve this.

2.2. Painting Captioning

Painting captioning is widely applied in artwork understanding. Some works used image captioning methods to describe paintings. Achlioptas et al. [17] presented a novel large-scale paintings dataset called ArtEmis. It contained 80K paintings and 455K emotion attributions explained by humans. However, the paintings' attributes (title, author, type, etc.) were also important in painting captioning. Garcia et al. [18] further presented a novel painting dataset called SemArt. It contained 21K art images. Each painting matched a group of attributes and at least one caption to describe. Deng et al. [19] proposed a novel style-enhanced artist classification framework that combines the styles and authors of art paintings and derives the creative characteristics of the artists from the paintings. Lu et al. [20] proposed a style transfer virtual data generation method and virtual–real semantic alignment module.

To solve this problem, Yan et al. [21] further proposed a VAD (Valence, Arousal, and Dominance) dictionary and gated concatenation mechanism to generate captions of paintings. This method fused affective word embedding and made caption generation more accurate, but did not contain the author, background, etc. Bai et al. [22] proposed a multi-topic and knowledge-based painting captioning framework. This framework extracted knowledge from the extra knowledge database to obtain a caption with the author, background, and objects.

These methods show an application of image captioning on paintings. However, they still lack explicit artistic conception captions and cannot generate objective captions and artistic conception captions of high quality. Hence, we aim to solve these challenges.

2.3. Multimodal Models

Most existing multimodal models based on Transformers can be divided into single-stream and two-stream types by the fusion method. Most single-stream models used CNNs or Transformers to obtain image visual features, then put word embedding and visual features into a unified Transformer. In order to enhance inter-modal interactions, Yu et al. [27] proposed a Multimodal Transformer (MT), used GloVe and LSTM as the linguistic encoder, then used Faster R-CNN and Transformer as the visual encoder, and finally sent linguistic features and visual features into a Transformer. Increasingly, studies focus on how to align linguistic and visual features. VL-BERT [28] and Unicoder-VL [29] models use Faster R-CNN as an image feature extractor, then use a single Transformer to align language features and image features but do not use object tags information. Li et al. [30] further proposed an Oscar model and used object tags to assist vision and language alignment instead of using region features and language features. Most previous works only focus on single modality or multi-modality and cannot efficiently adapt to each other. More recently, Li et al. [31] proposed an UNIMO pre-train model that not only simply aligns image and language features but also uses the contrastive learning loss function, making the true image-text distance close.

Most two-stream models have a similar approach to single-stream models to obtain visual and word embedding, then use different Transformers to process vision and language features and use customized co-attention to let Transformers interact with each other. To align visual and linguistic features, the LXMERT pre-train model proposed by Tan et al. [32] and the ViL-BERT model proposed by Lu et al. [33] align different modal information. Zhuang et al. [34] proposed the selfALign module which improves the retrieval accuracy while maintaining retrieval efficiency.

Multimodal models can enrich linguistic information. Therefore, we attempted to use a multimodal model to fuse visual and linguistic features, aiming to optimize the performance of artistic conception captions.

3. Methodology

Traditional image captioning models cannot generate painting captions suitably and accurately. Hence, we propose a DecoupleCLIP model to solve this problem. As shown in

Figure 1 and Algorithm 1, we use a two-branch painting captioning structure (see Section 3.1), which mainly consists of two modules: a local to global features fusion module (see Section 3.2) and a multimodal fusion module (see Section 3.3).

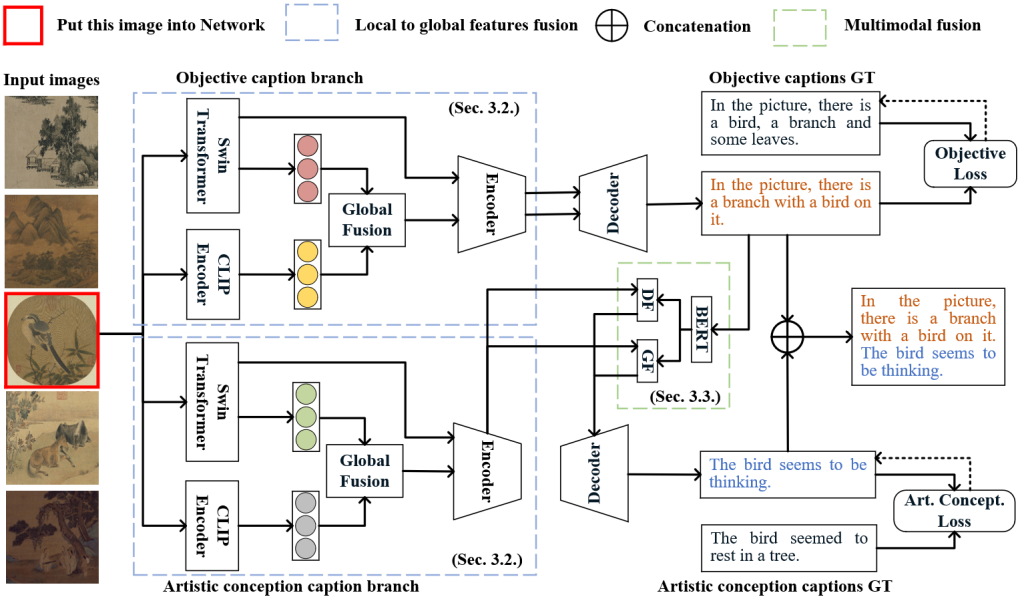


Figure 1. The framework of our proposed method. Given an input image of 384 × 384 pixels, our method employs two branches: the objective caption branch (top branch in this figure) and the artistic conception caption branch (bottom branch in this figure) to generate captions individually. In the artistic conception caption branch, we integrate a multimodal fusion module to enhance the generation performance of artistic conception captions. Subsequently, after generating both objective and artistic conception captions, we concatenate these captions to produce the final output.

Algorithm 1 Overview of DecoupleCLIP

Input: the paintings
Output: the final captions

 # local to global feature fusion (Section 3.2)
$L \leftarrow Swin(I)$ *#Use Swin Transformer*
$G_c \leftarrow CLIP(I)$
$G_l \leftarrow AvgPool(L)$
$G \leftarrow GlobalFusion(G_c, G_l)$
$E_l, E_g \leftarrow Encoder(L, G)$

 # objective branch decoder part
$C_{obj} \leftarrow Decoder(E_l, E_g)$ *#Generate objective captions*
$l_c \leftarrow l(C_{obj})$ *#compute objective branch loss*

 # artistic branch multimodal fusion and decoder part
$F_{mm} \leftarrow MMFusion(E_l, E_g, C_{obj})$ *# Section 3.3*
$C_{art} \leftarrow Decoder(F_{mm})$
$l_a \leftarrow l(C_{sub})$ *#compute artistic branch loss*

 # generate the final captions
$C \leftarrow Concatenation(C_{obj}, C_{art})$

3.1. Network Desgin

We construct a two-branch network to generate high-quality objective and artistic conception captions. In the objective caption branch, we feed original paintings to the local to global features fusion module and then generate objective captions through the decoder. In the artistic conception branch, we also feed the same images into the local to global features fusion module to obtain the global and local features of the images for the artistic conception captions. These features are combined with the generated objective captions and jointly input into the multimodal fusion module. Then, we feed them into a decoder to generate artistic conception captions.

3.2. Local to Global Features Fusion

The local to global features fusion module is shown in the blue dotted box in Figure 1. We use the local to global features fusion module to increase the information of global image features and utilize local image features simultaneously.

In the first part of the local to global features fusion module, we use the Swin Transformer and the CLIP model to extract local image features $V_D = \{V_D^1, V_D^2, \ldots, V_D^m\}$ and global image features V_g. Before we use the CLIP model to obtain global image features, we use bilinear interpolation to adjust the image resolution (384×384) to 224×224. After extracting local image features V_D, we use global average pooling to obtain another type of global image features $\hat{V}_G$, where $\hat{V}_G = \frac{1}{m}\sum_{i=1}^{m} V_D^i$.

In the second part of the local to global features fusion module, we construct a global fusion module to fuse different global image features $\hat{V}_G$ and V_g. The global fusion module can extract synthetic global features V_G, and it can be formulated as follows:

$$V_G = LN\Big(ReLU\Big(W_f[V_g; \hat{V}_G]\Big)\Big) + \alpha V_g + \hat{V}_G, \tag{1}$$

where W_f is the learnable parameter matrix of a linear layer, LN denotes the layer normalization method, α is the learnable parameter, and $[V_g; \hat{V}_G]$ denotes the concatenation of image global features obtained from CLIP and Swin Transformer (contains global average pooling).

In conclusion, we use the Swin Transformer to extract local features. The local features extracted by Swin Transformer are transformed into global features $\hat{V}_G$ by global average pooling. The global features V_g are extracted by CLIP. Finally, V_g and $\hat{V}_G$ are fused to obtain the final global features V_G.

In the third part of the local to global features fusion module, we construct an image feature encoder, as shown in Figure 2.

The encoder can encode local and global image features separately. The left part of the encoder block consists of Window Multi-Head Attention (W-MA) and Shifted Window Multi-Head Attention (SW-MA) with a feedforward module. Specifically, the W-MA and the SW-MA modules are used alternately. The W-MA has lower computational costs compared to the Multi-Head Attention (MA) module. We add the SW-MA module after the W-MA module to improve cross-window modeling capabilities (with reference to the Swin Transformer [2]). The left part of the encoder block can be formulated as follows:

$$\tilde{E}_D = LN((S)W\text{-}MA(V_D, [V_D; V_G], [V_D; V_G]) + V_D), \tag{2}$$

where $(S)W\text{-}MA()$ denotes use SW-MA or W-MA, which are used in Swin Transformer.

$$E_D = LN(FFN(\tilde{E}_D) + \tilde{E}_D), \tag{3}$$

where E_D denotes the output local features. FFN denotes the FeedForward Network (FFN), which consists of two linear modules, ReLU activation function, and dropout modules. The right part of the encoder block structure is similar to the left part. We use MA, FFN, and LN.

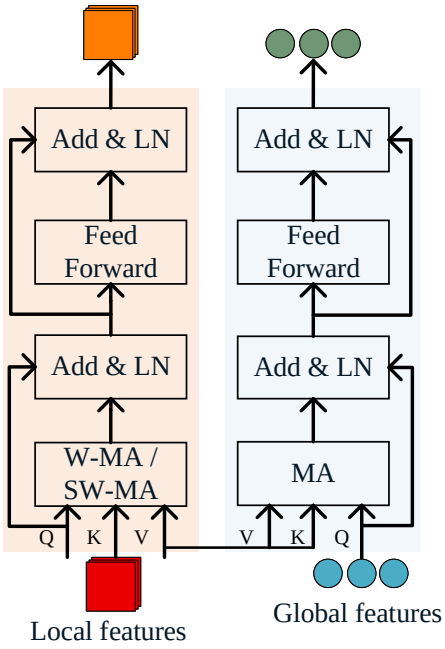


Figure 2. The encoder structure of local to global features fusion module.

3.3. Multimodal Fusion

The multimodal fusion module is shown in the dotted green box of Figure 1. This chapter uses the multimodal fusion module to assist the generation of high-quality artistic conception captions. A suitable painting caption not only pays attention to objective captions but also notes the reasonableness of artistic conception. For example, a bird is standing on a tree in a painting. If we generate the artistic conception of "people enjoying life", it would be unfitting.

To address this issue, we construct a multimodal fusion module that refers to the co-attention transformer layer [33] as shown in Figure 3. It consists of three parts: sentence-level global feature generation for the objective captions part (middle part of Figure 3), image global features adjustment part (GF part of Figure 3), and image local features adjustment part (DF part of Figure 3).

In generating sentence-level global features of the objective captions part, to facilitate the constraint of image global and local features by objective captions, we introduce the BERT pre-train model to obtain sentence-level global features. Specifically, we feed objective caption $L_{sentence}$ into the BERT model to get vectorized representation. Then, we use a linear module to scale the vector to the specified size. The scaled vector is sentence-level global features of objective captions L_g.

In the image local features adjustment part, we use the Multi-Head Attention (MA) module, layer normalization, and the FFN module. Specifically, we feed image local features V_D as Query of MA, and sentence-level features L_g as Key and Value of MA. This way, sentence-level features can interact with image local features but without adjustment linguistic features, only adjusting visual features. The output of image local features can be denoted as R_D, which can be formulated as follows:

$$R_D = LN(LN(MA(V_D, L_g, L_g) + V_D) + FFN(LN(MA(V_D, L_g, L_g) + V_D))), \quad (4)$$

where in $MA(V_D, L_g, L_g)$ V_D denotes the Query of the MA module and L_g is the Key and Value of the MA module.

In the image global features adjustment part, we use the sentence-level global features of the objective captions L_g and the image global features V_G to generate image global features R_G. The implementation method is similar to R_D.

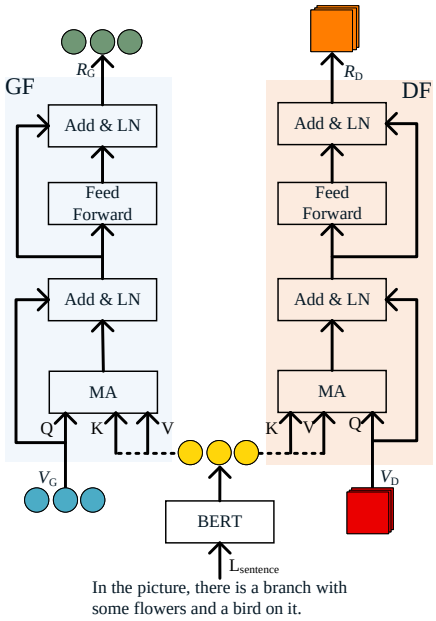


Figure 3. The structure of the multimodal fusion module. The middle part of the structure uses BERT to generate the global features of sentences. The left part of the structure uses image local features and the global features of sentences to generate new image local features that contain linguistic features. The right part of the structure uses a similar method to generate new image global features.

We compute the loss function separately with the true value and the generated result of the image captions of each branch. We combine the KL Divergence function with the Softmax method as each branch loss function. In the first step, we compute the Softmax of prediction value, called $Q(x_i)$,

$$Q(x_i) = log\left(\frac{exp(x_i)}{\sum_j exp(x_j)}\right), \tag{5}$$

where x_i is the output value of the previous module. In the second step, we use ground truth and $Q(x_i)$ to compute objective caption branch KL Divergence as loss of objective caption branch, called $l_c(P, Q)$,

$$l_c(P, Q) = \sum_{i=1} P(x_i) log \frac{P(x_i)}{Q(x_i)}, \tag{6}$$

where $P(x_i)$ is ground truth. We can easily ontain artistic conception loss $l_a(P, Q)$ in the same way.

Training and testing phase. For the training phase, we feed the original image into the objective branch to generate objective captions. Then, we feed the ground truth objective captions and original images into the artistic conception branch to generate artistic conception captions. Compared with the use of predicted objective captions, our model can reduce training errors. For the testing phase, we feed the original image into the objective branch to generate objective captions. Then, we feed the generated objective captions along with the original images into the artistic conception branch to generate artistic conception captions. Finally, we concatenate the artistic conception captions and objective captions to generate the final painting caption.

4. Experiments

We conduct extensive experiments on our dataset to demonstrate the effectiveness of our DecoupleCLIP model. We further analyze the results on the Chinese paintings part

and Western paintings part of our dataset to verify the robustness under different styles of paintings.

4.1. Datasets

We propose a small-scale painting captioning dataset containing 2400 images, and each image is annotated with objective and artistic captions. We use multiple experienced individuals to label the dataset in a way that eliminates subjectivity to some extent. The captions length is considerably longer than the COCO caption dataset, and the original image resolution ranges from thousands to tens of thousands. For the convenience of training, we unify the image resolution to 384×384.

In our dataset, the landscape accounts for about 39.8% of the total keyword number, persons account for 26.5%, plants account for 23.5%, animals account for 5.5%, and vehicles account for 4.6%. The type of landscape includes houses, mountains, etc. The types of plants include bamboo, flowers, etc. The types of vehicles include horses, boats, etc. The types of animals include birds, cows, etc.

4.2. Experimental Settings

We train the CLIP model separately using the training data from the painting captioning dataset and employ the pre-trained CLIP model as the CLIP Encoder module in Figure 1. The number of heads in the MA module is set to 8. We set the training epochs to 80, and the batch size to 20. We employ the cross-validation method, and without the SCST [4] training method.

For all the comparison models and DecoupleCLIP model, we train (on the training dataset), validate (on the validation dataset), and test (on the test dataset) with our proposed dataset, and all the above models converge on the validation set.

4.3. Evaluation

In each metrics comparison table, the first row of each model represents the average metrics, and the second row represents the variance of the experiment.

Quantitative Evaluation. The performance comparison of different baselines and our model in pure Chinese paintings part is shown in Table 1. We highlight the best model in bold. In this part of the experiment, we only use the Chinese paintings part of the dataset to train each model separately. B@1 indicates BLEU-1, and B@2 to B@4 indicate similar metrics. Our model achieves a CIDEr score of 35.5, which is an improvement of at least 1.97 compared to the other baseline models. Meanwhile, our model improves at least 0.8% on BLEU-4 and at least 0.57% on SPICE. However, our model performs slightly lower than the PureT model [2] on BLEU-1 and METEOR. These metrics suggest that our model can produce more accurate keywords and smoother sentences. However, due to the limited vocabulary size of the dataset, the diversity of non-keywords is smaller.

Compared to the representative two-stage image captioning model (GRIT), we attribute our performance improvement to eliminating the object detection module, which cannot effectively detect the object in Chinese paintings.

Compared to the representative one-stage models (PureT and VirTex), we attribute our performance improvement to using the two-branche painting captioning structure, which facilitates the training accurate painting captioning models. Our DecoupleCLIP model improves performance and reduces annotation cost, as it does not require additional object detection annotations.

The performance comparisons among different baselines and our model in the full dataset, which contains Chinese and Western paintings, are shown in Table 2. We use bold font to indicate the best model and blue to indicate the second-best model.

Comparing the sentence length generated by the model can reflect the quality of the sentences to some extent. As shown in Figure 4, we analyze the results of the cross-validation in the Chinese paintings part, where more than half of the captions have lengths between 15 to 30 and 45 to 99. Our method generates captions whose lengths are closest to the ground truth in these two ranges.

Studying the relationship between sentence length and metrics generated by the model, we conduct statistical analysis on single-part cross-validation results in the Chinese paintings dataset, as depicted in Figure 5. PureT and our model exhibit relative stability in CIDEr and SPICE metrics when the sentence length ranges from 15 to 30 and from 30 to 45, with our model performing the best. With a richer and more balanced sample set, our model would achieve greater stability and produce more realistic captions.

Table 1. Metric comparisons among different baselines and our model in our proposed dataset of Chinese paintings part. The best indicator is shown in bold black. The first row represents the average and the second row represents the variance.

Metrics	PureT [2]	GRIT [35]	VirTex [36]	DecoupleCLIP (Ours)
B@1	**43.40** 5.7×10^{-4}	39.33 1.1×10^{-3}	29.73 3.4×10^{-4}	43.20 1.7×10^{-4}
B@2	31.43 5.6×10^{-4}	28.80 7.3×10^{-4}	20.08 1.6×10^{-4}	**31.70** 8.9×10^{-5}
B@3	23.45 5.2×10^{-4}	21.78 4.6×10^{-4}	13.93 1.6×10^{-4}	**24.15** 5.8×10^{-5}
B@4	17.33 4.3×10^{-4}	16.23 4.1×10^{-4}	8.93 1.8×10^{-4}	**18.18** 5.4×10^{-5}
METEOR	**18.55** 3.0×10^{-5}	18.08 7.9×10^{-5}	16.34 9.1×10^{-5}	18.13 8.3×10^{-6}
ROUGE-L	37.40 4.9×10^{-4}	35.78 8.1×10^{-4}	30.45 2.5×10^{-4}	**38.08** 2.7×10^{-5}
CIDEr	33.53 5.6×10^{-3}	20.23 8.7×10^{-4}	8.07 7.3×10^{-4}	**35.50** 6.8×10^{-4}
SPICE	27.08 3.1×10^{-4}	22.33 2.2×10^{-4}	17.35 5.8×10^{-4}	**27.65** 6.0×10^{-5}

Table 2. Metric comparisons among different baselines and our model in the full dataset show the best indicator in bold black. The first row represents the average, and the second row represents the variance.

Metrics	PureT [2]	GRIT [35]	VirTex [36]	DecoupleCLIP (Ours)
B@1	**41.65** 1.7×10^{-4}	28.85 3.1×10^{-4}	24.78 3.3×10^{-4}	41.18 2.1×10^{-4}
B@2	**29.20** 1.1×10^{-4}	20.28 5.1×10^{-4}	16.65 4.9×10^{-4}	29.13 1.6×10^{-4}
B@3	21.35 9.6×10^{-5}	14.90 4.3×10^{-4}	11.90 5.1×10^{-4}	**21.60** 1.0×10^{-4}
B@4	15.53 7.5×10^{-5}	10.50 3.6×10^{-4}	7.83 5.0×10^{-4}	**15.93** 8.5×10^{-5}
METEOR	**16.85** 7.0×10^{-6}	11.83 6.1×10^{-5}	15.38 4.9×10^{-4}	16.83 1.6×10^{-5}
ROUGE-L	35.58 5.9×10^{-5}	28.90 7.3×10^{-4}	27.93 5.9×10^{-4}	**35.20** 3.1×10^{-5}
CIDEr	32.05 1.2×10^{-3}	10.68 1.1×10^{-4}	9.05 3.2×10^{-4}	**33.88** 6.2×10^{-4}
SPICE	23.28 8.5×10^{-5}	14.65 3.7×10^{-4}	16.75 8.7×10^{-4}	**23.95** 1.2×10^{-5}

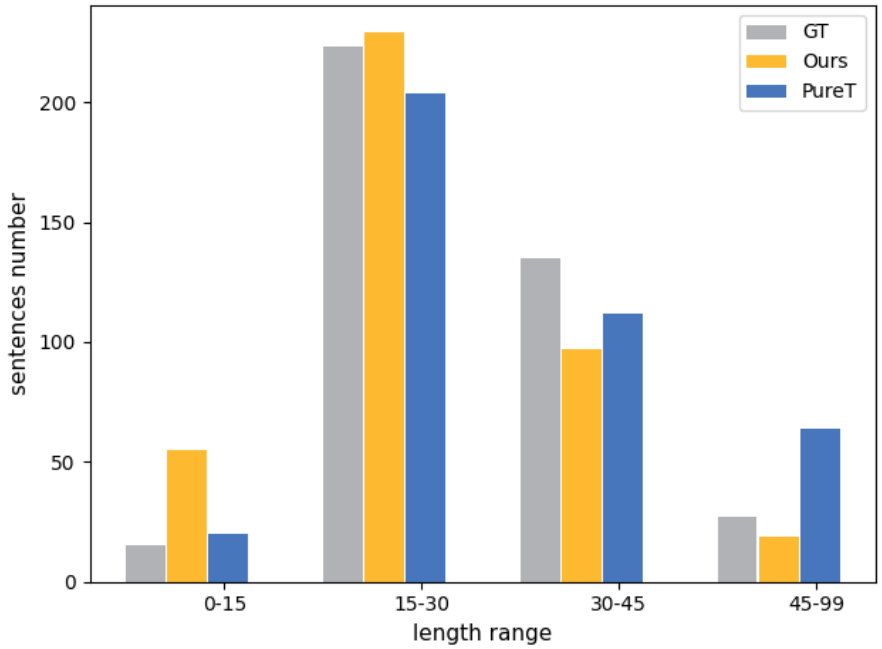


Figure 4. Sentence length of image captions in Chinese paintings part. GT denotes ground truth sentence length. Our model generates the sentence length closest ground truth in the second and the last range. The number of sentences in these two ranges accounted for 50% of all sentences.

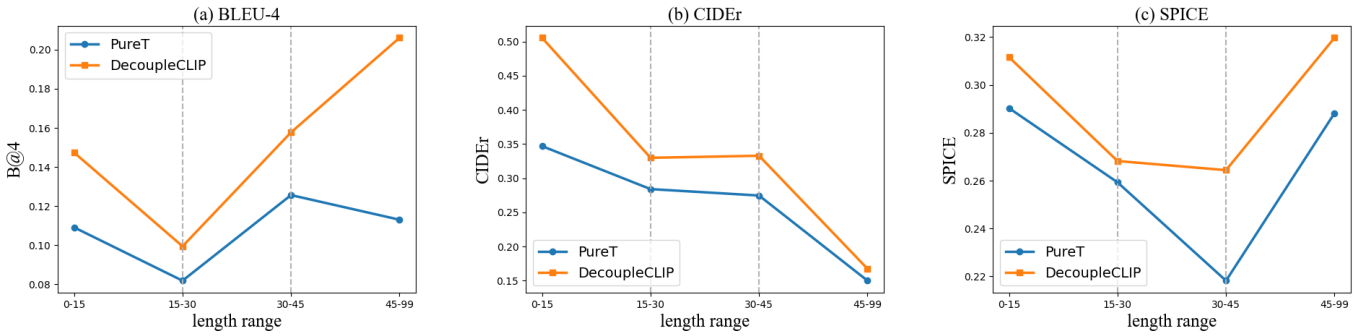


Figure 5. Different models generate captions between length ranges of sentences and different metrics relationships in the Chinese paintings part dataset. From left to right, BLUE-4, CIDEr, and SPICE are explained. Our model performance is good in most situations.

Qualitative Evaluation. The VirTex model is not shown in the qualitative analysis section due to its low metrics and low word generation efficiency. The GRIT model converts words that occur 1 or fewer times to <unk>.

Figure 6 presents the qualitative analysis results of Chinese paintings from our model, GRIT and PureT. The text captions below each Chinese painting are generated, respectively, by our model and PureT. For the first Chinese painting, the PureT model incorrectly generated the repeated captions "several houses". The GRIT model failed to recognize "Mountains" and "House" and did not generate substantial artistic captions. For the second Chinese painting, the PureT model incorrectly identified "person" and "houses" that were not present. The GRIT model also incorrectly identified "houses" and produced incomplete artistic captions. In the case of the third Chinese painting, the PureT model incorrectly identified "bridge", "man", and "hill", and generated unreasonable artistic captions. The GRIT model repeatedly generated the phrases "two people" and "bridge", and produced incorrect artistic captions. For the fourth Chinese painting, the PureT model incorrectly identified "bridge" and "person", and the GRIT model incorrectly identified a "boat". Both models generated artistic captions that were unrelated to the painting.

(ours) There are several peaks in the picture. There are many trees on the mountains, and several houses. The picture gives people a sense of desolation.	(ours) In the picture, there are several mountains, a river and trees on the mountains in the distance. The picture gives people a feeling of peace and tranquility.	(ours) There is a river in the picture. There are several trees by the river. There are several houses under the trees. The picture gives people a feeling of openness.	(ours) There are several mountains, a piece of water and several trees in the picture. The picture gives people a sense of tranquility and nature.
(PureT) There are several mountains in the picture. There are many trees at the foot of the mountain, several houses in the woods, and **there are several houses in the distance**. The picture gives people a sense of desolation.	(PureT) There are several mountains in the picture. There is a river at the foot of the mountain, several trees and **a person** on the mountain. There are **several houses** in the forest. The picture gives people a feeling of peace and tranquility.	(PureT) There is a river in the picture. There are many trees on the river. There is **a small bridge** on the river. There is **a man** of the river. There are several houses on the **hill**. **The people in the house seemed to enjoy the scenery**.	(PureT) There is a river in the picture. There are some trees by the river. There is a **bridge** on the river and a **person** on the bank. The **man seemed to enjoy the scenery by the river**.
(GRIT) there is a tree in the picture. the picture gives people a <unk> of <unk>	(GRIT) there is a river in the picture. there is a bridge on the river. there are several <unk> beside the bridge. there are several houses under the trees. there are **several houses** under the trees. the picture gives people a feeling of peace and <unk>.	(GRIT) there is a river in the picture. there is a **bridge** on the river. there are **two people** on the bridge. there are **two people** in the **bridge**. the **two people** in the house. the **two people** seem to be talking.	(GRIT) there is a river in the picture. there is a **boat** on the river. there is a **boat** on the river. there is a **boat** on the **boat.** there are **two people on the boat** <unk> to **enjoy the scenery on the water**

Figure 6. Image captions of qualitative evaluation only in the Chinese paintings part. The first row shows four under-test paintings. The following three rows show the result of different models, and incorrect words are highlighted in red. We use different color boxes and words to annotate mainly object position and object type in different images.

Figure 7 presents the qualitative analysis results of the complete dataset, encompassing both the Chinese and Western parts, in comparison with our model, the PureT model, and the GRIT model. In contrast to our model, the PureT model was unable to accurately identify the main objects in the four paintings and produced inappropriate artistic captions for the fourth painting. Similarly, the GRIT model failed to accurately identify the main objects in the paintings and did not generate meaningful artistic captions. These experiments intuitively demonstrate the superior effectiveness of our model.

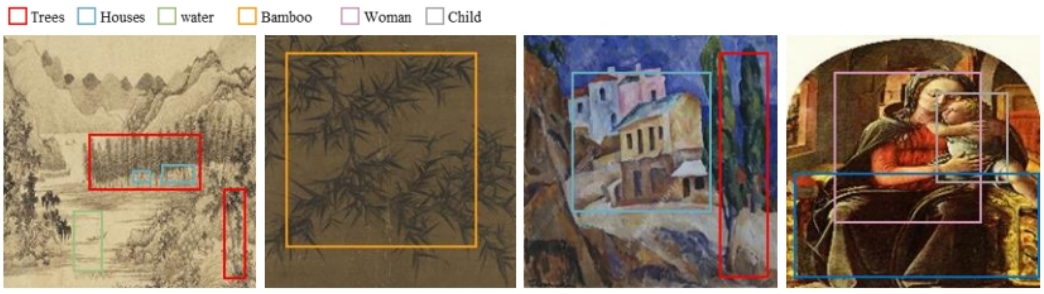


Figure 7. *Cont.*

(Ours) There is a piece of water in the picture. There is a forest beside the water, and there are several houses in the distance. The picture gives people a feeling of openness and distance.	(Ours) There are several bamboos in the picture. The picture gives people a sense of vitality.	(Ours) There are houses and trees in the picture. The color is very beautiful and peaceful.	(Ours) A woman is sitting in a chair holding a child. The woman is calm and peaceful.
(PureT) There is a river in the picture. There are many trees by the river. There is **a bridge on the river**. There are several houses in the forest. **There is a man on the road**. The picture gives people a feeling of peace and tranquility.	(PureT) There are several bamboos and **stones** in the picture. The picture gives people a sense of vitality.	(PureT) There are some houses in the forest. There are many **houses and trees beside the houses**. The people a sense of tranquility.	(PureT) A woman sat in a chair holding a child **standing on the ground**. The woman looks like **he's faces holding her child**.
(GRIT) there is a river in the picture. there are several trees by the river. there are several houses under the trees. **there are several houses under the trees**. the picture gives people a feeling of <unk> and <unk>.	(GRIT) there is a **river** in the picture there are several **trees** in the picture. the picture gives people a <unk>.	(GRIT) there is a **river** in the picture. the picture gives people a <unk> of <unk>.	(GRIT) a **man** sits on a chair with **a winged man** across from her **angels** always give me a <unk> of **wonder**.

Figure 7. Image captions of qualitative evaluation in the full dataset. The first row shows four under-test paintings. The following three rows show the result of different models, and incorrect words are highlighted in red. We use different color boxes and words to annotate mainly object position and object type in different images.

4.4. Ablation Studies

To assess the impact of each component in our DecoupleCLIP model, we evaluated various configurations of our approach. The effectiveness of our multimodal fusion module in DecoupleCLIP was tested on our dataset, comparing scenarios with and without the multimodal fusion module (w/o MM), as detailed in Table 3. For the case of using the multimodal fusion module, we compared the effect of using co-attention and our multimodal fusion module, respectively. The co-attention layer structure consists of two parallel Transformers, similar to Figure 2, with input features comprising local image features and sentence-level global features from objective captions. Our multimodal fusion module outperformed other configurations in most metrics. We attribute this performance gain to the effectiveness of our specific multimodal fusion module, which enhances the generation of accurate artistic captions. This underscores the importance of a well-suited multimodal fusion module in generating artistic conception captions.

The CLIP module and the global fusion module have a causal relationship, so we conducted the ablation experiment of the two as a whole. The effectiveness of the Global Fusion and CLIP module of our DecoupleCLIP is evaluated on our dataset, which is described in Table 4. Compared to the without Global Fusion and CLIP module (w/o Golbal Fusion and CLIP), we attribute our performance improvement to fusing more global features in local to global feature fusion, which facilitates training a more effective encoder. It can be concluded that the CLIP and Global Fusion modules can fuse important global features. Combined with the ablation experiment of the multimodal fusion modules, it can be shown that using the appropriate multimodal fusion module can further improve performance.

Figure 8 presents the results of our model, the PureT model, and the GRIT model qualitative analysis of the Chinese paintings. Compared to our model, the model using the co-attention module generates unreasonable artistic conception captions, as seen in the last three Chinese paintings. This could be attributed to modifications of objective captions by the co-attention layer, which negatively impact the artistic conception captions. In contrast to our model, the model without the multimodal fusion module (w/o MM)

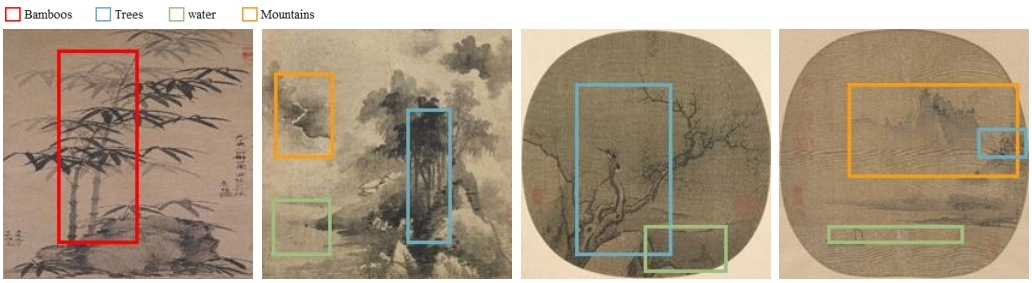
generates unreasonable artistic captions, such as in the second Chinese painting. This occurs because artistic conception captions lack the constraints provided by objective captions. The importance of a suitable multimodal fusion module is highlighted.

Table 3. Metric comparisons among different multimodal fusion modules in Chinese paintings part. The best indicator is shown in bold black. The first row represents the average and the second row represents the variance.

MM Module	B@1	B@4	CIDEr	SPICE
co-attention layer	**43.33** 1.8×10^{-4}	**18.30** 6.1×10^{-5}	35.08 1.1×10^{-3}	27.55 1.2×10^{-4}
w/o MM	42.85 4.9×10^{-4}	17.90 5.2×10^{-4}	33.88 1.8×10^{-3}	26.98 2.0×10^{-4}
Ours	43.20 1.7×10^{-4}	18.18 5.4×10^{-5}	**35.50** 6.8×10^{-4}	**27.65** 6.0×10^{-5}

Table 4. Metric comparisons with and without the Global Fusion and CLIP module in Chinese paintings part. The better indicators are shown in bold black. The first row represents the average and the second row represents the variance.

Module	B@1	B@4	CIDEr	SPICE
w/o Global Fusion and CLIP	43.00 7.7×10^{-4}	17.10 4.3×10^{-4}	27.60 6.5×10^{-3}	25.48 5.5×10^{-4}
Ours	**43.20** 1.7×10^{-4}	**18.18** 5.4×10^{-5}	**35.50** 6.8×10^{-4}	**27.65** 6.0×10^{-5}



(Ours) There are several bamboos in the picture. The picture gives people a sense of vitality.	(Ours) There is a piece of water in the picture. There is a forest by the water and several mountains in the distance. The picture gives people a feeling of openness and distance.	(Ours) There are several trees by the river in the picture. The picture gives people a feeling of desolation.	(Ours) There are many mountains in the picture. There is a river at the foot of the mountain, and there are many trees on the bank. It expresses the ancients' love for mountains and rivers.
(co-attention) There are several bamboos in the picture. The picture gives people a feeling of vitality.	(co-attention) In the picture, there is a lake with several trees and several mountains in the distance. The **people on the boat seemed to enjoy the scenery** in the mountains.	(co-attention) There are trees and stones in the picture. The picture gives people a sense of **vitality**.	(co-attention) In the picture, there is a small boat in the water. There are many trees on the bank, and mountains in the distance. Deep in the mountains, **an old man was fishing alone on the river**.

Figure 8. *Cont.*

(w/o MM) There are several bamboos in the picture. The picture gives people a sense of vitality.	(w/o MM) There are several high mountains in the picture. There is a river at the foot of the mountain, and there are many trees on the bank. **These people seem to be talking**.	(w/o MM) In the picture, there is a tree by the river. The picture gives people a sense of **vitality**.	(w/o MM) There are mountains, trees and **mountains** in the picture. The picture gives people a sense of **vitality**.
(w/o CLIP) The picture shows **two peony flowers**. The picture gives people a sense of vitality.	(w/o CLIP) There is a small river in the picture. There are several rocks by the water, and there is **a person under the tree**. The picture gives people a feeling of openness and distance.	(w/o CLIP) The picture shows a **plum blossom** tree growing by the river bank of **water plants behind him**. **A man seemed to be thinking**.	(w/o CLIP) In the picture, there are several mountains in the distance. There is a river at the foot of the mountain, a **pavilion** under the tree. The picture gives people a sense of **vitality**.

Figure 8. Image captions of ablation studies by category in Chinese paintings. The first row shows four under-test paintings in different categories. Other rows show generat sentences of different ablation methods, and incorrect words are highlighted in red. We use different color boxes to annotate mainly object position and object type in different images.

In the experiment without the CLIP and Global Fusion modules (w/o CLIP), objective captions contain numerous errors, affecting the artistic conception captions simultaneously in the last two Chinese paintings. This demonstrates that CLIP can extract critical global features that significantly influence objective captions.

5. Conclusions and Future Work

In this paper, we propose a painting captioning method with two branches to enhance the caption performance. Our method utilizes CLIP to construct a local-to-global feature fusion module and a multimodal fusion module. The former enriches global features, while the latter integrates linguistic and visual features to generate more coherent painting captions. Existing image captioning datasets rarely include Chinese paintings. To address this gap, we introduce a small dataset that includes both Chinese and Western paintings. Compared with other models, our proposed model achieves the best results on BLEU-2, BLEU-3, BLEU-4, ROUGE-L, CIDEr, and SPIC evaluation metrics.

In future work, we will construct a unified model that integrates the objective caption into the artistic conception caption while fully considering the influence of noisy labels on the model.

Author Contributions: Conceptualization, M.Z. and X.H.; methodology, M.Z. and X.H.; software, M.Z.; validation, Y.Y. and M.S.; formal analysis, X.H.; investigation, X.H.; resources, X.H.; data curation, M.Z. and X.H.; writing—original draft preparation, M.Z. and Y.Y.; writing—review and editing, X.H.; visualization, M.Z., Y.Y. and M.S.; supervision, X.H.; project administration, X.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Natural Science Foundation of China (62102036), Beijing Natural Science Foundation (4222024), R&D Program of Beijing Municipal Education Commission (KM202211232003), Open Project Program of State Key Laboratory of Virtual Reality Technology and Systems, Beihang University (No. VRLAB2022A02).

Data Availability Statement: The access link to the code (include dataset) in this paper is as follows: https://github.com/zml110120/CP (accessed on 20 October 2024).

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

References

1. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and tell: A neural image caption generator. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 8–10 June 2015; pp. 3156–3164.
2. Wang, Y.; Xu, J.; Sun, Y. End-to-End Transformer Based Model for Image Captioning. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 22 February–1 March 2023; pp. 2585–2594.

3. Wang, D.; Hu, Z.; Zhou, Y.; Hong, R.; Wang, M. A Text-Guided Generation and Refinement Model for Image Captioning. *IEEE Trans. Multimed.* **2023**, *25*, 2966–2977. [CrossRef]
4. Rennie, S.J.; Marcheret, E.; Mroueh, Y.; Ross, J.; Goel, V. Self-Critical Sequence Training for Image Captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1179–1195.
5. Dong, X.; Zhang, G.; Zhan, X.; Ding, Y.; Wei, Y.; Lu, M.; Liang, X. Caption-Aided Product Detection via Collaborative Pseudo-Label Harmonization. *IEEE Trans. Multimed.* **2023**, *25*, 1916–1927. [CrossRef]
6. Wang, C.; Gu, X. Learning Double-Level Relationship Networks for Image Captioning. *Inf. Process. Manag.* **2023**, *60*, 103288–103312. [CrossRef]
7. Luvembe, A.; Li, W.; Li, S.; Liu, F.; Wu, X. CAF-ODNN: Complementary Attention Fusion with Optimized Deep Neural Network for Multimodal Fake News Detection. *Inf. Process. Manag.* **2024**, *61*, 103653–103689. [CrossRef]
8. Lu, J.; Yang, J.; Batra, D.; Parikh, D. Neural Baby Talk. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7219–7228.
9. Jiang, W.; Zhou, W.; Hu, H. Double-Stream Position Learning Transformer Network for Image Captioning. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 7706–7718. [CrossRef]
10. Wang, Y.; Xu, N.; Liu, A.; Li, W.; Zhang, Y. High-order interaction learning for image captioning. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 4417–4430. [CrossRef]
11. Liu, A.; Zhai, Y.; Xu, N.; Nie, W.; Li, W.; Zhang, Y. Region-aware image captioning via interaction learning. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 3685–3696. [CrossRef]
12. Ge, H.; Yan, Z.; Zhang, K.; Zhao, M.; Sun, L. Exploring Overall Contextual Information for Image Captioning in Human-like Cognitive Style. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1754–1763.
13. Zhang, J.; Mei, K.; Zheng, Y.; Fan, J. Integrating Part of Speech Guidance for Image Captioning. *IEEE Trans. Multimed.* **2020**, *23*, 92–104. [CrossRef]
14. Prudviraj, J.; Vishnu, C.; Mohan, C. Attentive contextual network for image captioning. In Proceedings of the International Joint Conference on Neural Networks, Shenzhen, China, 18–22 July 2021; pp. 1–8.
15. Dai, J.; Qi, H.; Xiong, Y.; Zhang, Y.L.G.; Hu, H.; Wei, Y. Deformable convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2017; pp. 764–773.
16. Yu, L.; Zhang, J.; Wu, Q. Dual attention on pyramid feature maps for image captioning. *IEEE Trans. Multimed.* **2022**, *24*, 1775–1786. [CrossRef]
17. Achlioptas, P.; Ovsjanikov, M.; Haydarov, K.; Elhoseiny, M.; Guibas, L. Artemis: Affective Language for Visual Art. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 10–25 June 2021; pp. 11569–11579.
18. Garcia, N.; Vogiatzis, G. How to Read Paintings: Semantic Art Understanding with Multi-Modal Retrieval. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 10–25 June 2021; pp. 676–691.
19. Deng, Y.; Tang, F.; Dong, W.; Ma, C.; Huang, F.; Deussen, O.; Xu, C. Exploring the Representativity of Art Paintings. *IEEE Trans. Multimed.* **2020**, *23*, 2794–2805. [CrossRef]
20. Lu, Y.; Guo, C.; Dai, X.; Wang, F. Data-Efficient Image Captioning of Fine Art Paintings via Virtual-Real Semantic Alignment Training. *Neurocomputing* **2022**, *490*, 163–180. [CrossRef]
21. Yan, J.; Wang, W.; Yu, Y. Affective Word Embedding in Affective Explanation Generation for Fine Art Paintings. *Pattern Recognit. Lett.* **2022**, *161*, 24–29. [CrossRef]
22. Bai, Z.; Nakashima, Y.; Garcia, N. Explain Me the Painting: Multi-Topic Knowledgeable Art Description Generation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 5422–5432.
23. Radford, A.; Jong, K.; Chris, H.; Aditya, R.; Gabriel, G.; Sandhini, A.; Girish, S.; Amanda, A.; Pamela, M.; Jack, C. Learning Transferable Visual Models from Natural Language Supervision. In Proceedings of the International Conference on Machine Learning, Wien, Austria, 18–24 July 2021; pp. 8748–8763.
24. Lu, J.; Xiong, C.; Parikh, D.; Socher, R. Knowing When to Look: Adaptive Attention via a Visual Sentinel for Image Captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 375–383.
25. Zhang, X.; Sun, X.; Luo, Y.; Ji, J.; Zhou, Y.; Wu, Y.; Huang, F.; Ji, R. Rstnet: Captioning with Adaptive Attention on Visual and Non-Visual Words. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 10–25 June 2021; pp. 15465–15474.
26. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, G. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022.
27. Yu, J.; Li, J.; Yu, Z.; Huang, Q. Multimodal Transformer with Multi-View Visual Representation for Image Captioning. *IEEE Trans. Circuits Syst. Video Technol.* **2019**, *30*, 4467–4480. [CrossRef]
28. Su, W.; Zhu, X.; Cao, Y.; Li, B.; Lu, W.; Wei, F.; Dai, J. VL-BERT: Pre-Training of Generic Visual-Linguistic Representations. In Proceedings of the International Conference on Learning Representations, Addis Ababa, Ethiopia, 26–30 April 2020.
29. Li, G.; Duan, N.; Fang, Y.; Gong, M.; Jiang, D. Unicoder-vl: A Universal Encoder for Vision and Language by Cross-Modal Pre-Training. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 11336–11344.

30. Li, X.; Yin, X.; Li, C.; Zhang, P.; Hu, X.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F. Oscar: Object-semantics aligned pre-training for vision-language tasks. In Proceedings of the European Conference on Computer Vision, New Glasgow, UK, 23–28 August 2020; pp. 121–137.
31. Li, W.; Gao, C.; Niu, G.; Xiao, X.; Liu, H.; Liu, J.; Wu, H.; Wang, H. Unimo: Towards Unified-Modal Understanding and Generation via Cross-Modal Contrastive Learning. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Minneapolis, MN, USA, 2–7 June 2021; pp. 2592–2607.
32. Hao, H.; Mohit, B. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In Proceedings of the IConference on Empirical Methods in Natural Language Processing, Barceló Bávaro Convention Centre, Punta Cana, Dominican Republic, 7–11 November 2021.
33. Lu, J.; Batra, D.; Parikh, D.; Lee, S. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In Proceedings of the Conference and Workshop on Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; Volume 32, pp. 13–23.
34. Zhuang, J.; Yu, J.; Ding, Y.; Qu, X.; Hu, Y. Towards Fast and Accurate Image-Text Retrieval with Self-Supervised Fine-Grained Alignment. *IEEE Trans. Multimed.* **2023**, *26*, 1361–1372. [CrossRef]
35. Nguyen, V.-Q.; Suganuma, M.; Okatani, T. GRIT: Faster and better image captioning transformer using dual visual features. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Volume 13696, pp. 167–184.
36. Desai, K.; Johnson, J. Virtex: Learning visual representations from textual annotations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 10–25 June 2021; pp. 11.157–11.168.