applied sciences

MDPI

Article

A Unified Visual and Linguistic Semantics Method for Enhanced Image Captioning

Jiajia Peng and Tianbing Tang *

School of Computer and Electronical Infonmation, Guangxi University, Nanning 530004, China; 2113391053@st.gxu.edu.cn
* Correspondence: tbtangtbtang@126.com

Abstract: Image captioning, also recognized as the challenge of transforming visual data into coherent natural language descriptions, has persisted as a complex problem. Traditional approaches often suffer from semantic gaps, wherein the generated textual descriptions lack depth, context, or the nuanced relationships contained within the images. In an effort to overcome these limitations, we introduce a novel encoder–decoder framework called A Unified Visual and Linguistic Semantics Method. Our method comprises three key components: an encoder, a mapping network, and a decoder. The encoder employs a fusion of CLIP (Contrastive Language–Image Pre-training) and SegmentCLIP to process and extract salient image features. SegmentCLIP builds upon CLIP's foundational architecture by employing a clustering mechanism, thereby enhancing the semantic relationships between textual and visual elements in the image. The extracted features are then transformed by a mapping network into a fixed-length prefix. A GPT-2-based decoder subsequently generates a corresponding Chinese language description for the image. This framework aims to harmonize feature extraction and semantic enrichment, thereby producing more contextually accurate and comprehensive image descriptions. Our quantitative assessment reveals that our model exhibits notable enhancements across the intricate AIC-ICC, Flickr8k-CN, and COCO-CN datasets, evidenced by a 2% improvement in BLEU@4 and a 10% uplift in CIDEr scores. Additionally, it demonstrates acceptable efficiency in terms of simplicity, speed, and reduction in computational burden.

Keywords: image captioning; image features; clustering mechanism; Chinese language description

Citation: Peng, J.; Tang, T. A Unified Visual and Linguistic Semantics Method for Enhanced Image Captioning. *Appl. Sci.* **2024**, *14*, 2657. https://doi.org/10.3390/app14062657

Academic Editors: Mourad Oussalah and Rachid Jennane

Received: 15 February 2024
Revised: 10 March 2024
Accepted: 15 March 2024
Published: 21 March 2024

1. Introduction

Image captioning, the process of crafting coherent textual descriptions from visual content, has garnered substantial attention in the evolving landscape of computer vision [1–3]. Despite notable strides in this domain, the synthesis of accurate and contextually rich captions remains a formidable challenge [4]. Contemporary methodologies, particularly those rooted in deep learning frameworks, encounter persistent hurdles in capturing the nuanced relationships, intricate details, and contextual subtleties present within images [5,6].

Recent years have witnessed a surge in research endeavors aimed at enhancing the efficacy of image captioning methods [7–11]. Stefanini et al. [7] provide a thorough analysis of deep learning approaches in image captioning, emphasizing the synergy between visual encoders and language models. Their study highlights significant advancements in feature extraction and narrative generation techniques. Yan et al. [8] develop a Task-Adaptive Attention mechanism for image captioning, effectively blending visual and non-visual word generation. Their method enhances caption relevance and accuracy, particularly for non-visual elements. Herdade et al. [9] introduce the Object Relation Transformer, a novel approach to image captioning that incorporates spatial relationships between objects using geometric attention. Yao et al. [10] introduce a novel hierarchical parsing architecture for image captioning, integrating multi-level visual patterns from instances to regions and the entire image. He et al. [11] propose an Image Transformer for image captioning,

enhancing the original Transformer architecture to encode spatial relationships between image regions. However, these efforts often encounter limitations in grasping the complex interplay of diverse visual elements, resulting in descriptions that lack holistic context and fail to encapsulate the essence of the depicted scene.

The crux of the gap in the existing image captioning methodologies lies in their inability to holistically capture the intricate semantics inherent in visual data. Despite the incorporation of attention mechanisms and multimodal architectures, as highlighted by recent studies from Xu et al. [12] and Anderson et al. [13], the challenge of retaining comprehensive contextual information within generated captions persists. Issues pertaining to the accurate portrayal of spatial relationships, fine-grained details, and the subtle interdependencies among various elements within an image continue to pose significant hurdles.

Addressing the constraints inherent in previous models, our research presents an innovative encoder–decoder framework, termed "A Unified Visual and Linguistic Semantics Method" (UVLS). This approach signifies a fundamental shift in the paradigm of image captioning, integrating advanced visual and linguistic semantic analyses. UVLS is designed to bridge the gap between visual perception and language interpretation, offering a more cohesive and contextually nuanced understanding of the content in images. Through this method, we aim to set a new benchmark in the field, transcending the limitations of traditional image captioning techniques with a more integrated and sophisticated approach. At its core, our model integrates a fusion architecture intertwining CLIP and SegmentCLIP, as elucidated by the works of Ron et al. [14] and Xu et al. [15]. This integration augments the process of feature extraction from images and significantly enriches the semantic understanding between textual and visual components. SegmentCLIP's incorporation of clustering mechanisms further refines the model's ability to discern intricate contextual semantics within images, enabling a more nuanced and comprehensive portrayal of visual scenes.

The centerpiece of our proposed framework comprises three key components: an encoder, a mapping network, and a decoder. Leveraging CLIP and SegmentCLIP, the encoder adeptly extracts salient features from images, which are subsequently transformed into a fixed-length prefix by the mapping network. The GPT-2-based decoder [16] then synthesizes these features into contextually accurate Chinese language descriptions for the images, thereby addressing the shortcomings prevalent in existing captioning methodologies.

Our framework demonstrates substantial advancements, particularly in reconciling the semantic gap prevalent in image captioning. Through comprehensive evaluations on established benchmarks such as AIC-ICC, Flickr8k-CN, and COCO-CN datasets, our model outperforms contemporary methods. Notably, it showcases superior efficiency in terms of computational simplicity, processing speed, and reduced resource utilization. Moreover, our approach significantly contributes to generating more nuanced, contextually enriched, and semantically accurate image descriptions, aligning with recent literature that emphasizes the crucial role of semantic relationships in captioning tasks. By seamlessly integrating CLIP and SegmentCLIP, our model excels in capturing intricate visual–linguistic nuances, ultimately leading to more comprehensive and contextually accurate image descriptions.

The main contributions are summarized as follows:

- We propose the "A Unified Visual and Linguistic Semantics Method" (UVLS), integrating CLIP and SegmentCLIP for image feature extraction. This framework aims to bridge the semantic gap in traditional image captioning, offering a more nuanced understanding of visual content.
- By incorporating a clustering mechanism into CLIP, called SegmentCLIP, the model enhances semantic relationships between visual and textual elements. This advancement enables a deeper interpretation of complex contextual semantics within images, leading to richer and more accurate image descriptions.
- Extensive experiments on real-word datasets are conducted to fully verify the effectiveness of our proposed UVLS.

2. Related Work

Image captioning is a domain aiming to generate natural language descriptions for visual content [17]. Recent advancements in deep learning and vision–language pre-training techniques have significantly propelled this field. This comprehensive review presents an in-depth analysis of deep learning methodologies in image captioning, encompassing diverse categories and addressing prevalent challenges. Furthermore, we assess the performance of various captioning models using widely adopted metrics and outline prospective avenues for future research in this domain.

Encoder–Decoder Paradigms: One prevalent approach in image captioning involves the encoder–decoder paradigm. Herein, convolutional neural networks (CNNs) encode input images to derive feature representations [18,19]. Geetha et al. [18] employ deep convolutional neural networks and recurrent neural networks (RNNs), including LSTM and GRU, for accurate multi-class, multi-label satellite image captioning, utilizing VGG-19 for feature extraction and finetuned LSTM for decoding. Liu et al. [19] explore image captioning using deep neural networks, specifically focusing on frameworks based on CNN-RNN, CNN-CNN, and reinforcement learning to enhance the accuracy and relevance of captions generated from images. Subsequently, these representations are decoded using recurrent neural networks to generate corresponding captions [20,21]. Despite their efficacy, encoder–decoder techniques often yield captions of a generic and ambiguous nature due to the compression of comprehensive information into a singular vector. Numerous adaptations to the encoder–decoder structure have been proposed to mitigate these limitations, including attention mechanisms and graph-based methodologies [22,23].

Attention Mechanisms: Attention mechanisms revolutionize image captioning by dynamically guiding focus to pertinent image regions during caption generation, significantly boosting caption accuracy and relevance. Zohourianshahzadi et al. [24] demonstrate that the integration of attention mechanisms, particularly soft, bottom-up, and multi-head attention, within encoder–decoder architectures, significantly augments the efficacy of image captioning models. They highlight that these mechanisms enable a nuanced focus on pertinent image regions, thus facilitating the generation of contextually rich and semantically coherent captions. Anderson et al. [13] propose a novel approach to image captioning by integrating a combined bottom-up and top-down attention mechanism, enabling the model to focus on objects and salient image regions dynamically. This methodology significantly enhances the generation of contextually relevant and detailed captions by allowing the attention mechanism to be calculated at the level of objects, thereby setting a new state of the art in image captioning performance.

Huang et al. [25] introduce an "Attention on Attention" (AoA) module to refine traditional attention mechanisms in image captioning, emphasizing the calculation of relevance between attention results and queries. By incorporating AoA into both encoder and decoder stages, they present AoANet, an advanced model that significantly improves the generation of detailed and contextually relevant captions, setting a new benchmark in state-of-the-art performance for image captioning tasks. Pedersoli et al. [26] propose "Areas of Attention," a model that enhances image captioning by dynamically associating caption words with specific image regions through a novel attention mechanism, leveraging weakly supervised learning to improve localization and description accuracy. Wang et al. [27] develop a Hierarchical Attention Network (HAN) for image captioning that innovatively employs a pyramidal hierarchy of semantic features, enabling simultaneous utilization of patch, object, and text features to generate contextually rich captions, significantly outperforming the existing methods in accuracy and detail.

Attention mechanisms significantly advance the field of image captioning by dynamically directing the model's focus towards relevant image segments, thereby elevating the precision and contextual relevance of generated captions. Through the strategic integration of hierarchical semantic features and nuanced attention strategies, these mechanisms facilitate the creation of semantically rich and contextually accurate captions, marking notable progress in the capabilities of image captioning models.

Vision–Language Pre-training Advancements: Recent strides in vision–language pre-training have exerted a profound impact on image captioning research [28–30]. Zhou et al. [28] present a unified vision–language pre-training (VLP) model for image captioning, employing a Transformer network for both encoding and decoding, with pre-training on large image–text pairs. The method, unique for its unified architecture across bidirectional and seq2seq vision–language prediction tasks, demonstrated improvements in learning efficiency and model accuracy across diverse image captioning tasks. Wang et al. [29] develop ViLTA, a vision–language pre-training model that enhances learning efficiency and robustness through cross-distillation for Masked Language Modeling and the generation of synthetic hard negatives for Image–Text Matching. This novel approach, leveraging textual augmentation, demonstrates improved performance in various vision–language tasks, notably in image captioning, by refining representation quality and model convergence. Li et al. [30] introduce Uni-EDEN, a Universal Encoder–Decoder Network for vision–language tasks, focusing on multi-granular vision–language pre-training. This approach notably enhances multimodal reasoning and language modeling capabilities, advancing both perception and generation aspects in image captioning. These methods, characterized by advanced neural network architectures and multimodal reasoning, significantly enhance accuracy and contextual understanding in image captioning tasks. However, they exhibit limitations in terms of generalizability and robustness, primarily due to their heavy dependence on large specifically curated datasets, which may not adequately represent the diversity of real-world scenarios.

Deep learning methodologies have significantly propelled advancements in image captioning. Encoder–decoder paradigms and dense captioning strategies have been extensively explored for generating captions at varying granularity levels. The advent of vision–language pre-training techniques has notably augmented captioning performance. Standardized evaluation metrics and datasets have provided benchmarks for assessing and comparing diverse captioning models. Future research avenues may focus on mitigating challenges such as information alignment discrepancies, dataset biases, and the development of improved evaluation tools for precise caption quality assessment.

In contrast to these methodologies, our approach shows improvement. Integrating CLIP and SegmentCLIP enhances semantic understanding, refining relationships and enriching image depth. Leveraging SegmentCLIP's clustering refines visual representations. Our GPT-2-based decoder elevates coherence, surpassing limitations for richer Chinese descriptions. Notably, our model's solid performance on AIC-ICC, Flickr8k-CN, and COCO-CN exemplifies efficiency, marking significant progress in holistic image descriptions.

3. Problem Definition

Image captioning involves the development of a unified model that receives an image as its input and undergoes training to optimize the probability, $p(S|I)$, of generating a sequence of words, $S = \{S_1, S_2, \cdots, S_j, \cdots\}$, where each word S_j is drawn from a specified dictionary. This sequence aims to effectively describe the image, encapsulating its essence and content through a rich and comprehensive portrayal in language.

4. Methodlogy

We commence by outlining our problem statement. Given a corpus comprising paired instances of images and corresponding captions $\{I_i, S_i\}_{i=1}^{N}$, our primary objective is to acquire the capability to generate a contextually relevant and meaningful caption for an unseen input image. The captions are denoted as a sequence of tokens $S^i = S_1^i, \ldots, S_\ell^i$, where padding is applied to reach a maximal sequence length of l. Subsequently, our training is driven by the following objective:

$$\max_{\theta} \sum_{i=1}^{N} \log p_\theta(S_1^i, \ldots, S_\ell^i | I^i) \tag{1}$$

Herein, θ represents the set of learnable parameters constituting the model. Our innovation centers on leveraging the enriched semantic embeddings from CLIP and

SegmentCLIP, amalgamating their strengths. CLIP demonstrates versatile comprehension of diverse visual concepts, while SegmentCLIP enriches semantic relationships via clustering mechanisms. We integrate both embedded representations as a pivotal condition, encompassing crucial visual information for our approach. Building upon recent literature, we conceptualize this precondition as a prefix integral to the caption. Given that the necessary semantic context is embedded within this prefix, we employ an autoregressive language model, predicting subsequent tokens independently without considering future tokens. Thus, our primary objective is articulated as follows:

$$\max_{\theta} \sum_{i=1}^{N} \sum_{j=1}^{\ell} \log p_{\theta}(S_j^i | I^i, S_1^i, \ldots, S_{j-1}^i) \tag{2}$$

4.1. Overview

The methodology is visually depicted in Figure 1. Employing GPT-2 as our language model, we utilize its tokenizer to convert the caption into a sequence of embeddings. Extracting visual information from image I^i involves leveraging the visual encoder from pre-trained models such as CLIP [31] and SegmentCLIP. Subsequently, a lightweight mapping network, denoted as M, facilitates the transformation of the joint embedding into k embedding vectors:

$$p_1^i, \ldots, p_k^i = M(\text{CLIP}(I^i) \oplus \text{SegmentCLIP}(I^i)). \tag{3}$$

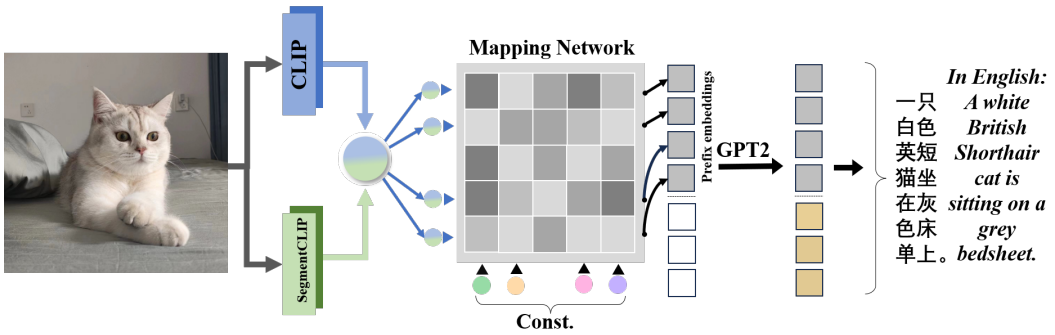


Figure 1. Framework of UVLS.

Here, $\oplus$ denotes the concatenation operator, and each vector p_j^i shares the dimensionality of a word embedding. Subsequently, we concatenate the acquired visual embedding with the embeddings of caption S_i:

$$Z^i = p_1^i, \ldots, p_k^i, S_1, \ldots, S_\ell^i. \tag{4}$$

During the training stage, we input the language model with the concatenated prefix–caption sequences $\{Z^i\}_{i=1}^N$. Our training objective revolves around autoregressively predicting the caption tokens conditioned on the prefix. For this objective, the mapping component M undergoes training employing straightforward yet impactful cross-entropy loss.

$$\mathcal{L}_X = -\sum_{i=1}^{N} \sum_{j=1}^{\ell} \log p_{\theta}(S_j^i | p_1^i, \ldots, p_k^i, S_1^i, \ldots, S_{j-1}^i). \tag{5}$$

In bridging the gap from the initial phase of training our model on prefix–caption sequences to introducing the innovative SegmentCLIP image encoding mechanism, it is pivotal to understand the progression towards a more intricate engagement with visual data. The transition from focusing solely on linguistic elements to incorporating a sophisticated visual understanding marks a significant expansion in our model's capabilities. As the mapping component learns to predict caption tokens more effectively through cross-entropy loss, it sets the stage for a deeper exploration into the visual domain. SegmentCLIP emerges

as a natural progression in this journey, aiming to harness the full spectrum of visual information. By encoding images into a structured hierarchical format, SegmentCLIP enables the model to capture and interpret complex visual cues with unprecedented detail. This methodology not only enhances the model's ability to generate contextually rich captions but also fosters a more profound integration of linguistic and visual semantics. As we delve into the architecture of SegmentCLIP, it becomes evident how this mechanism plays a crucial role in transcending traditional boundaries between visual perception and language generation, ultimately leading to a more nuanced and comprehensive model.

4.2. SegmentCLIP

We present the SegmentCLIP image encoding mechanism as depicted in Figure 2, which employs a Transformer-derived framework for the hierarchical and progressive categorization of visual concepts. Within the architecture of SegmentCLIP, Transformer layers are compartmentalized into several stages of grouping. At each of these stages, a series of group tokens are acquired as learnable entities through the mechanism of self-attention, enabling the global assimilation of information from all image tokens (segments). Subsequently, these acquired group tokens are utilized to amalgamate similar image tokens by employing a Segmenting Block. By orchestrating a structured progression of grouping stages, we facilitate the consolidation of smaller image segments into larger conglomerates. The subsequent sections will elucidate each component in detail.

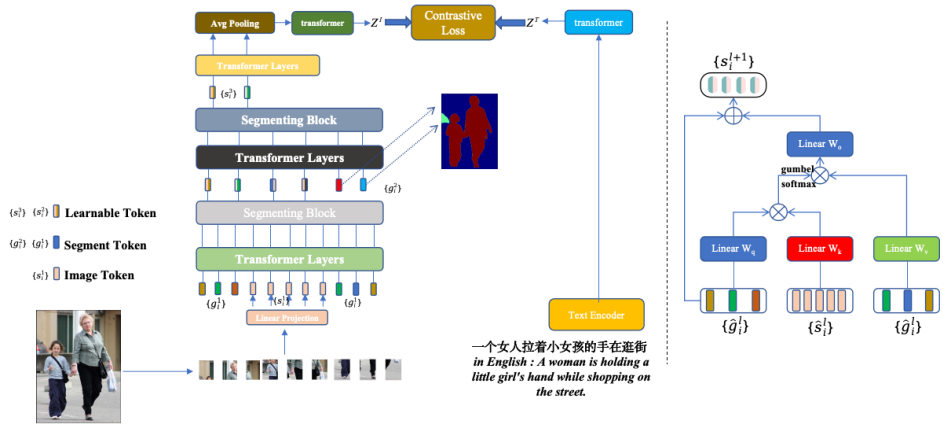


Figure 2. (**a**) **The architecture and training pipeline of segmentCLIP.** easing size. The images to the right illustrate the visual segments that manifest across different grouping stages. In the lower stages, pixels are grouped into parts of objects, such as the hands and legs of a woman or a little girl; in the higher stages, these are further amalgamated into complete entities, such as the entire body of the woman and the little girl. (**b**) **The architecture of the segmenting block.** At the end of each grouping stage, there is a segmenting block that calculates the similarity between the learned tokens and the segment (image) tokens. The assignment is determined via a gumbel softmax operation over the learned tokens and is then converted into a hard one-hot assignment. The segment tokens that are assigned to the same group are merged, forming new segment tokens that serve as input for the subsequent grouping stage.

Building upon the ViT [32] architecture, we initially partition an input image into N discrete non-overlapping segments, subsequently mapping each segment linearly into a latent dimensional space. These transformed segments are regarded as individual image tokens, collectively represented by the notation $\{\mathbf{I}_i\}_{i=1}^{N}$ At every stage of grouping, in addition to these image tokens, we integrate a suite of learnable group tokens. This augmented set is then fed into the Transformer designated for the respective stage.

In the context of Multi-stage Grouping, as illustrated in Figure 2a, rather than channeling all N input image tokens through the entirety of the Transformer's layers, we distribute

these layers across multiple stages of grouping, establishing a hierarchical structure. At the conclusion of each stage, a Segmenting Block is employed to amalgamate smaller entities into larger conglomerates.

To articulate this process formally, let us posit L grouping stages, each identified by an index l and associated with a collection of learnable group tokens g_i. For the sake of clarity, the initial set of image patches $\{\mathbf{I}_i\}_{i=1}^{N}$, serving as inputs to the first grouping stage, are designated as the preliminary segments, with $N = M_0$. This simplification allows us to equate the initial set of image patches with the initial group of segments.

Commencing with $l = 1$, at each stage of grouping, we commence by amalgamating the image patches $\{\mathbf{I}_i\}$ and group tokens $\{\mathbf{g}_i\}$, subsequently inputting this combined set into a sequence of Transformer layers. These layers are tasked with facilitating information exchange between the segments and group tokens, thereby enabling effective propagation of information. For expediency, the set of image patches $\{\mathbf{I}_i\}$ input to the initial grouping stage is treated as the set of starting segments $\{\mathbf{s}_i\}_{i=1}^{M_0}$, where $N = M_0$. We simplify the notation of the starting segments to s_i and the learnable group tokens to g_i. Commencing with $l = 1$, for each grouping stage, we first concatenate s_i and g_i together and then input them into a series of Transformer layers, each of which facilitates information propagation between the concatenated tokens via

$$\{\hat{\mathbf{g}}_i^l\}, \{\hat{\mathbf{s}}_i^l\} = \text{Transformer}([\{\mathbf{g}_i^l\}; \{\mathbf{s}_i^l\}]), \tag{6}$$

Here, the symbol $[;]$ symbolizes the concatenation operation. Subsequently, we assimilate the updated M_{l-1} image segment tokens $\{\hat{\mathbf{s}}_i^l\}$ into M_l new segment tokens $\{\tilde{s_i^{l+1}}\}_{i=1}$ via a Segmenting Block as follows:

$$\{\mathbf{s}_i^{l+1}\} = \text{SegmentingBlock}(\{\hat{\mathbf{g}}_i^l\}, \{\hat{\mathbf{s}}_i^l\}, \{\hat{\mathbf{g}}_i^l\}). \tag{7}$$

At every grouping stage, the relationship $M_l < M_{l-1}|$ holds true, indicating a progressive reduction in the number of group tokens and, concurrently, a coalescence into fewer but larger image segments. Upon reaching the terminal grouping stage L, we apply Transformer layers on all segment tokens and leverage their outputs to derive the final global image representation z^L as follows:

$$\{\hat{\mathbf{s}}_i^{L+1}\} = \text{Transformer}(\{\mathbf{s}_i^{L+1}\}), \tag{8}$$

$$z^L = \text{Transformer}(\text{AvgPool}(\{\hat{\mathbf{s}}_i^{L+1}\})). \tag{9}$$

As delineated in Figure 2a, SegmentCLIP reconfigures visual information into an array image segments that are not restricted by a regular grid structure, thus optimizing the visual information arrangement.

In the case of the Segmenting Block, as depicted in Figure 2b, at the conclusion of each grouping stage, the Segmenting Block unifies the learned group tokens and segment tokens attributed to identical groups into a single new image segment based on similarity in the embedding space. This process is mathematically formulated as

$$\mathbf{A}_{i,j}^l = \frac{\exp(W_q \hat{\mathbf{g}}_i^l \cdot W_k \hat{\mathbf{s}}_j^l + \gamma_i)}{\sum_{k=1}^{M_l} \exp(W_q \hat{\mathbf{g}}_k^l \cdot W_k \hat{\mathbf{s}}_j^l + \gamma_k)}, \tag{10}$$

where W_g and W_k represent the weights corresponding to the learned linear projections for the group and segment tokens, respectively. The term γ denotes the learnable bias pertaining to the Gumbel $(0, 1)$ distribution. The computation of segment token assignment to groups is conducted by employing the one-hot operation of an argmax over all the group tokens. Since the argmax operation is non-differentiable, we utilize the straight-through trick as a proxy, which results in a differentiable surrogate assignment matrix $\mathbf{A}$:

$$\hat{\mathbf{A}}^l = \text{one-hot}(\mathbf{A}_{\text{argmax}}^l) + \mathbf{A}^l - \text{sg}(\mathbf{A}^l), \tag{11}$$

where sg denotes the stop gradient operator. Utilizing this straight-through trick, $\mathbf{A}^l$ retains the one-hot assignment of a token to a single group, but its gradient is equivalent to the gradient of $\mathbf{A}$, ensuring the model remains end-to-end trainable.

Subsequent to the allocation of segment tokens to various learned groups, we amalgamate the embedding of all tokens belonging to the same group into a single new segment token $\mathbf{s}_i^{l+1}$. For each group, the output of the Segmenting Block is a weighted sum of the segment tokens assigned to that group, computed as

$$\mathbf{s}_i^{l+1} = \hat{\mathbf{g}}_i^l + W_o \frac{\sum_{j=1}^{M_{l-1}} \hat{\mathbf{A}}_{i,j}^l W_v \hat{\mathbf{s}}_j^l}{\sum_{j=1}^{M_{l-1}} \hat{\mathbf{A}}_{i,j}^l}, \tag{12}$$

where W_v and W_o represent the parameters learned to integrate the features. An alternate method to the hard assignment is the soft assignment, which utilizes $\mathbf{A}$ rather than $\mathbf{A}^l$, as prescribed in Equation (5). Our observations suggest that hard assignment is more effective for feature grouping compared to soft assignment, as evidenced by the results presented in Table 1.

Table 1. Statistics of the datasets used in our experiments.

Dataset	Lang.	Images	Captions	Caption Vocabulary
AIC-ICC	zh	300 K	1500 K	7654
Flickr8k-CN	zh	8000	40,000	1447
COCO-CN	zh	2041	101,705	2069

Analogous to a single cycle of the previously established Slot Attention framework, the Segmenting Block executes a similar function. However, where Slot Attention acquires instance-level feature clustering autonomously, our Segmenting Block arranges akin semantic segments employing minimal text-based guidance. For instance, as depicted in Figure 2's second row, the algorithm has clustered the pair of women effectively.

To enhance SegmentCLIP's capacity for hierarchical clustering, we employ carefully designed contrastive losses between image–text pairs. We harness these losses to refine visual representations via textual guidance, adopting a methodology inspired by prior research to train a dual-encoder framework. This framework encompasses an image encoder and a text encoder, leveraging a contrastive loss strategy.

For the Image–Text Contrastive Loss, we utilize the final image embedding derived from the SegmentCLIP and the concluding text embedding from the text encoder, representing the last output token. Our objective is to converge the representations of correlated image–text pairs and concurrently diverge the representations of non-correlated pairs.

Formally, we define a batch of B image–text pairs $\{(x_i^I, x_i^T)\}_{i=1}^B$, where x_i^I and x_i^T represent the image and textual components, respectively. Encoded into embeddings z_i^I and z_i^T through their respective encoders, we then calculate the dot product to assess their similarity. The comprehensive loss for image–text contrastive learning is articulated as a sum of image-to-text and text-to-image contrastive losses:

$$\mathcal{L}_{I \leftrightarrow T} = \mathcal{L}_{I \to T} + \mathcal{L}_{T \to I}, \tag{13}$$

Image-to-Text Contrastive Loss:

$$\mathcal{L}_{I \to T} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp(z_i^I \cdot z_i^T / \tau)}{\sum_{j=1}^{B} \exp(z_i^I \cdot z_j^T / \tau)}, \tag{14}$$

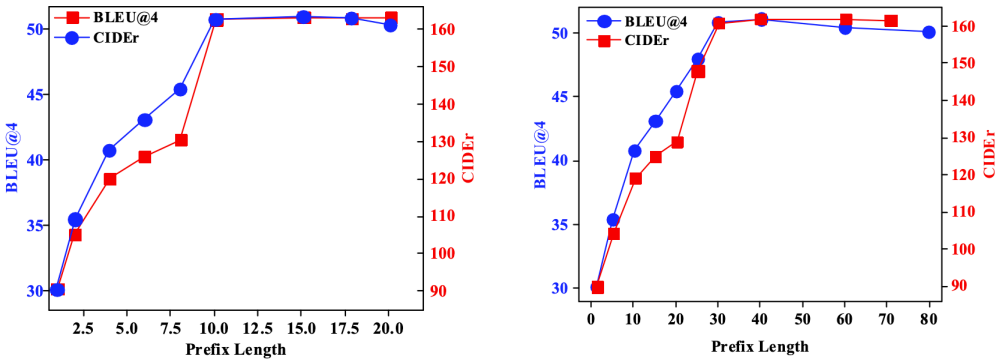
Text-to-Image Contrastive Loss:

$$\mathcal{L}_{T \to I} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp(z_i^T \cdot z_i^I / \tau)}{\sum_{j=1}^{B} \exp(z_i^T \cdot z_j^I / \tau)}, \tag{15}$$

Here, τ is a tunable temperature parameter to modulate the logits.

To facilitate enhanced visual categorization, we put forth a Multi-Label Image–Text Contrastive Loss in addition to the foundational image–text loss outlined in Equation (6). This multi-faceted contrastive loss is enriched through the use of 'prompt-engineered' textual labels, which are derived from the initial image descriptions as delineated in [31]. Specifically, we elect K nouns from the narrative x_i^T and synthesize these with pre-determined sentence frameworks, resulting in prompts akin to "An image depicting a noun". Such an approach is rooted in the rationale that nouns typically correspond to discernible objects within images.

In parallel to the training with original pairs, we introduce contrastive losses for the newly created 'image-prompted text' pairings $\{(x_i^I, x_i^{T_1})\}_{i=1}^B, \{(x_i^I, x_i^{T_2})\}_{i=1}^B, \ldots, \{(x_i^I, x_i^{T_K})\}_{i=1}^B$, where $\{x_i^{T_k}\}_{k=1}^K$ are the sentences crafted from the nouns isolated from x_i^T. Referencing Figure 3, this enriched method diverges from the standard contrastive loss as each image x_i^I correlates with K positive text pairings and $K(B-1)$ negative associations.



(**a**) MLP. (**b**) Transformer.

Figure 3. Effect of the prefix length on the captioning performance over the AIC-ICC dataset. For each prefix length, we report the BLEU@4 (red) and CIDERr (blue) scores over the test and train (dashed line) sets.

4.3. Parameter Analysis

In concordance with the foundational image–text contrastive loss delineated in Equation (6), we present a composite multi-label contrastive loss, expressed as

$$\mathcal{L}_{I \leftrightarrow \{T_k\}_{k=1}^K} = \mathcal{L}_{I \rightarrow \{T_k\}_{k=1}^K} + \mathcal{L}_{\{T_k\}_{k=1}^K \rightarrow I}, \tag{16}$$

This constitutes a dual-faceted loss, integrating both directional contrastive components. For the Image-to-Text direction:

$$\mathcal{L}_{I \rightarrow \{T_k\}_{k=1}^K} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\sum_{k=1}^{K} \exp(z_i^I \cdot z_i^{T_k} / \tau)}{\sum_{k=1}^{K} \sum_{j=1}^{B} \exp(z_i^I \cdot z_j^{T_k} / \tau)} \tag{17}$$

And for the Text-to-Image direction:

$$\mathcal{L}_{\{T_k\}_{k=1}^K \rightarrow I} = -\frac{1}{KB} \sum_{k=1}^{K} \sum_{i=1}^{B} \log \frac{\exp(z_i^{T_k} \cdot z_i^I / \tau)}{\sum_{j=1}^{B} \exp(z_i^{T_k} \cdot z_j^I / \tau)}. \tag{18}$$

Finally, the comprehensive image–text contrastive loss applied in the training of SegmentCLIP is articulated as the sum of the conventional and multi-label losses:

$$\mathcal{L} = \mathcal{L}_{I \leftrightarrow T} + \mathcal{L}_{I \leftrightarrow \{T_k\}_{k=1}^K}. \tag{19}$$

4.4. CLIP

In the given batch containing N distinct image and text pairings, x_i for images and y_i for texts, the CLIP framework is tasked with identifying the correct pairings from $N \times N$ possible combinations. This is facilitated through the development of a unified embedding domain by co-training an image encoder, f, and a text encoder, g, which are optimized to bolster the cosine similarity for actual image–text pairings, represented as $s_{ii} = \cos(f(x_i), g(y_i))$, while attenuating the cosine similarity for all inauthentic pairings, s_{ij} where $i \neq j$.

The optimization process involves a symmetric cross-entropy loss function over these computed similarity metrics, with the objective function defined as

$$L = -\frac{1}{N}\sum_{i=1}^{N}\left[\log\frac{e^{s_{ii}}}{\sum_{j=1}^{N} e^{s_{ij}}} + \log\frac{e^{s_{ii}}}{\sum_{j=1}^{N} e^{s_{ji}}}\right] \tag{20}$$

Here, the loss function is articulated in two parts. The first term, $\log\frac{e^{s_{ii}}}{\sum_{j=1}^{N} e^{s_{ij}}}$, embodies the softmax probability that an image x_i is correctly paired with its corresponding text y_i, and it is aimed to be maximized for all authentic image–text couplings within the batch. Conversely, the second term, $\log\frac{e^{s_{ii}}}{\sum_{j=1}^{N} e^{s_{ji}}}$, represents the softmax probability that a text y_i is correctly paired with its corresponding image x_i, similarly aimed to be maximized for each valid text–image union.

5. Experiments

This section presents an extensive experimental evaluation of our study. Section 5.1 presents a thorough examination of the evaluation process, encompassing both the datasets utilized for experimentation and the metrics applied for analytical assessment. In Section 5.2, we delineate the baseline models, offering a clear baseline comparison for assessing the effectiveness of our proposed methodology. Section 5.3 is dedicated to a comparative analysis, where our approach is juxtaposed with current state-of-the-art methods, demonstrating its relative performance and advancements. Finally, Section 5.4 delves into ablation studies, discussing in depth the specific components and variations of our method, thereby underscoring its robustness and the impact of individual elements on overall performance.

5.1. Evaluation

Dataset. In this work, we conduct experiments using the AIC-ICC (AI Challenge—Image Chinese Captioning) [33], Flickr8k-CN [34], and COCO-CN [35] datasets, which are among the most extensive datasets in the field of image captioning. The AIC-ICC dataset comprises approximately 300,000 images, accompanied by 1,500,000 Chinese language descriptions. The dataset comprises 210,000 images for the training set, 30,000 images for the validation set, and 60,000 images for the test set, structured to support comprehensive model training, validation, and performance testing in image-related research endeavors. These descriptions encompass a wide array of everyday scenes, numbering over 200, including but not limited to football fields, grasslands, and diverse actions such as singing and running. This dataset is particularly notable for its comprehensive coverage of commonplace scenarios, making it an invaluable resource for research in image description, especially in the context of Chinese language processing.

Flickr8k-CN represents the inaugural dataset for image captions in Chinese, primarily consisting of images derived from authentic human life scenes, where the subjects of the descriptions are notably prominent. This dataset contains a total of 8000 images, each accompanied by five descriptive texts offering varied perspectives on the content of the image. It is structured into 6000 images for the training set, 1000 images for the validation set, and 1000 images for the test set.

COCO-CN features a more diverse array of scenes within its images, incorporating a higher number of distractive elements. Each image is matched with one to five de-

scriptive texts, varying in count. The dataset encompasses 20,341 images in total, with 18,341 designated for the training set, 1000 for the validation set, and 1000 for the test set, offering a broader and more challenging context for image captioning in Chinese. The statistics of AIC-ICC, Flickr8k-CN, and COCO-CN are presented in Table 1.

Metrics. In our study, we rigorously assess the performance of our model using the widely recognized BLEU metric [36], which was initially developed for the domain of automated machine translation. This metric quantitatively evaluates the correspondence of n-grams (up to 4 grams) between the generated output, known as the hypothesis, and a reference or a collection of reference texts. To provide a more comprehensive evaluation, we also employ additional metrics, namely METEOR [37], ROUGE [38], and CIDEr [39]. METEOR extends beyond simple n-gram matching to include synonymy and paraphrase detection, thus offering more nuanced insight into semantic accuracy. ROUGE assesses the quality of summary by computing overlap between the model's output and reference summaries, catering to the evaluation of recall. Lastly, CIDEr enhances evaluation by focusing on the consensus between the generated captions and a set of references, weighted by the rarity of n-grams, which serves as an indicator of the informativeness and saliency of the generated text in the context of the image content. Together, these metrics furnish a robust framework for evaluating the linguistic and semantic coherence, as well as the informative value of the captions produced by our model.

Implementation details. We use the prefix length of $K = 10$ for the MLP mapping networks, where the MLP contains a single hidden layer. For the Transformer mapping network, we set the CLIP embedding to $K = 10$ constant tokens and use eight multi-head self-attention layers with eight heads each. We train for 10 epochs using a batch size of 40. The architecture of SegmentCLIP is based on ViT-S [32,40] with 12 Transformer layers, each with a hidden dimension of 384. We use input images of size 224 × 224 and a patch size of 16 × 16. We add a learnable positional embedding to each patch after linearly projecting it. We experiment with one-stage and two-stage architectures for SegmentCLIP. Both architectures output eight tokens after the final grouping stage. In one-stage SegmentCLIP, we learn 64 group tokens and insert the grouping block after the sixth Transformer layer. Before the grouping block, we project the 64 group tokens into eight tokens using an MLP-Mixer layer [41] and output eight segment tokens. In two-stage SegmentCLIP, there are sixty-four and eight group tokens in the first and second grouping stages, respectively. We insert grouping blocks after the sixth and ninth Transformer layers. Our text encoder is the same as [31]. We use a two-layer MLP to project the visual and text embedding vectors into the same latent space. For optimization, we use AdamW [42] with weight decay fix as introduced by Loshchilov et al. [43], with a learning rate of $2e^5$ and 5000 warm-up steps. For GPT-2, we employ the implementation of Wolf et al. [44].

Language model finetuning. As described in Section 3, finetuning the language model results in a much more expressive model, but it is also more susceptible to overfitting as the amount of trainable parameters increases. As can be seen in experiments, the two variants—with and without the language model finetuning—are comparable. Over the extremely complicated AIC-ICC dataset, we obtain superior results with the finetuning. We thus hypothesize that extremely elaborated datasets or ones that present a unique style require more expressiveness and hence are more likely to benefit from the finetuning.

5.2. Baselines

To validate the effectiveness of the proposed model, we select the following baselines:

- **CS-NIC** [34]. The CS-NIC model excels in generating Chinese captions for images, employing Jieba for precise Chinese text segmentation, addressing the unique challenge of tokenizing Chinese without explicit word boundaries. Utilizing the pool5 layer of GoogLeNet, it effectively extracts image features, with a focus on convolutional neural network (CNN) technology for nuanced visual understanding. The model sets both visual and word embeddings to a size of 512, balancing rich representation with computational efficiency. For translation tasks, it favors Baidu's service, noted

for superior English–Chinese translation accuracy, underscoring its commitment to precision. CS-NIC stands out as a sophisticated tool, harmonizing advanced image processing and linguistic analysis, specifically optimized for the Chinese language image captioning domain. The CS-NIC model excels in generating Chinese captions for images through a bilingual approach, effectively integrating linguistic diversity into visual descriptions. However, its dependency on machine translation may compromise the depth and accuracy of the captions, posing challenges in fully capturing the nuances of human language.

- **CIC** [45]. The Convolutional Image Captioning (CIC) model represents a paradigm shift in automated image captioning, employing a convolutional neural network (CNN) instead of traditional recurrent neural network (RNN) methods. It features a four-component architecture including input and output embedding layers, image embedding, and a convolutional module with masked convolutions. This unique structure allows CIC to forgo RNN's recurrent functions, leveraging a feed-forward deep network for direct modeling of image–word relations. The CIC model, therefore, offers a more efficient approach to image captioning, combining convolutional processing's strengths with streamlined computational complexity. The CIC model stands out for its parallel processing efficiency, offering speedy training and competitive accuracy in image captioning by leveraging CNNs. Its primary limitation, however, lies in its potential struggle with long-range textual dependencies, which may impact the coherence of generated captions for complex visuals.
- **SC** [46]. The Stack-Cap (SC) model employs a unique coarse-to-fine sentence decoder for image captioning, comprising one coarse decoder and a sequence of attention-based fine decoders. This structure allows for the refined prediction of each word, using cues from the preceding decoder. Initially, the coarse decoder provides a basic description from global image features. In subsequent stages, each fine decoder enhances the image description, integrating both the image features and the output from the previous stage. This method involves using attention weights from one stage to inform the next, leading to progressively refined predictions. The architecture, with one coarse and several stacked fine decoders, demonstrates a sophisticated approach to incrementally improving image captions. The SC model excels in producing detailed image captions through a unique coarse-to-fine approach, effectively addressing the vanishing gradient issue. However, its multi-stage prediction framework increases training complexity and computational demands.
- **Oscar** [47]. The Oscar model introduces a pioneering approach in vision–language pre-training (VLP) by utilizing object tags as anchor points to facilitate the learning of semantic alignments between images and texts. This model significantly enhances the process of cross-modal representation learning by structuring input as triples of word sequences, object tags, and image region features. By leveraging detected object tags in images, Oscar efficiently aligns these with corresponding textual descriptions, thereby improving the accuracy and relevance of generated image captions. Pre-trained on a large dataset of 6.5 million text–image pairs, Oscar achieves state-of-the-art performance across multiple V + L tasks, demonstrating its effectiveness in bridging the semantic gap between visual content and language. This approach not only advances the field of image captioning but also contributes to a broader understanding and generation tasks in vision–language research, making it a valuable asset for future explorations in multimodal AI applications. The Oscar model enhances vision–language pre-training by utilizing object tags as anchor points, significantly boosting cross-modal learning. However, its effectiveness is contingent on the accuracy of the underlying object detection, which may restrict its adaptability and generalization across varied datasets.
- **CLIPCAP** [48]. The CLIPCAP model introduces a simplified yet effective method for image captioning, a key task in vision–language understanding. It uniquely employs CLIP encoding as a prefix in the captioning process, utilizing a straightforward map-

ping network followed by finetuning of a language model, specifically GPT2. This approach leverages the rich semantic features of the CLIP model, trained within a textual context, making it highly suitable for vision–language tasks. The integration of CLIP with a pre-trained language model enables comprehensive understanding of both visual and textual data. Remarkably, this model requires relatively brief training and can generate meaningful captions without the need for additional annotations or extensive pre-training, making it adept at handling large-scale and diverse datasets. The CLIPCAP model efficiently integrates CLIP's visual encodings with a streamlined mapping network for rapid and resource-efficient caption generation, yet its dependence on pre-trained encodings may constrain its adaptability to diverse or novel imagery.

5.3. Performance Comparison

The comparative examination of image captioning models, as illustrated in Table 2, highlights the superior performance of the Unified Visual and Linguistic Semantics Method (UVLS) across various evaluation metrics. The detailed data provided in Table 2 enable in-depth analysis of the Unified Visual and Linguistic Semantics (UVLS) model's performance relative to other contemporary image captioning models across three datasets: AIC-ICC, Flickr8k-CN, and COCO-CN.

Table 2. Comparison of evaluation indexes of different models on ICC dataset. Top scores for each metric are bold. All values are reported as percentage (%).

Datasets	Model	BLEU@1	BLEU@2	BLEU@3	BLEU@4	METEOR	ROUGE	CIDEr
AIC-ICC	CS-NIC	62.8	48.0	37.0	28.7	30.7	51.8	74.6
	CIC	66.3	52.5	41.7	33.2	33.3	56.5	92.3
	SC	73.0	60.3	49.6	41.0	35.3	60.3	116.6
	Oscar	74.4	65.1	50.6	43.1	36.5	62.4	140.5
	CLIPCAP	76.7	65.1	55.0	46.4	38.8	64.0	151.2
	Ours; UVLS	**79.7**	**68.9**	**59.1**	**50.7**	**40.1**	**66.2**	**164.2**
Flickr8k-CN	CS-NIC	63.2	50.2	38.4	28.1	31	54.7	81.6
	CIC	68.7	55.5	43.3	31.6	33.2	59	96
	SC	73.3	59.7	47.3	35.9	35.3	61	121.3
	Oscar	80.4	69.5	59.4	50	41.2	67.3	160.5
	CLIPCAP	84.4	72.3	62.1	52	43.9	70.3	170.2
	Ours; UVLS	**90.7**	**75.9**	**67.1**	**56.7**	**48.1**	**76.2**	**184.2**
COCO-CN	CS-NIC	42.7	31.6	16.9	10.5	5.6	24.9	50.3
	CIC	50.1	42.2	27.6	20.6	7.3	26.9	65.9
	SC	59.5	42.3	35.7	24.6	10.6	30.9	80.8
	Oscar	69.4	58.3	48.6	38.5	20.4	46.3	107.8
	CLIPCAP	71.8	61.2	50.9	41.4	25.4	55.7	110.4
	Ours; UVLS	**74.7**	**64.8**	**56.0**	**49.3**	**30.6**	**58.4**	**124.2**

Starting with the AIC-ICC dataset, the UVLS model's performance can be characterized by its BLEU@1 score of 79.7%, which surpasses other models significantly. This high score indicates that the UVLS model is particularly adept at correctly identifying and employing the most frequent words in captions, which is fundamental for establishing a base for accurate and relevant descriptions. As we progress to BLEU@2 and BLEU@3 scores, where scores of 68.9% and 59.1%, respectively, are observed, there is a clear indication of the model's ability to construct longer and more complex linguistic structures. This is critical for forming coherent and detailed narratives that are necessary for a complete understand-

ing of an image. The BLEU@4 score of 50.7% further solidifies UVLS's strength in creating extended phrases and sentences that are syntactically varied and contextually rich.

The model's top ROUGE score of 66.2% on the AIC-ICC dataset reflects its superior summarization capability, suggesting that it can capture the essential content of images in a way that is most similar to the reference captions. This is particularly important in image captioning as it ensures that the generated captions are not only descriptive but also encapsulate the main themes and messages conveyed by the image. In terms of the METEOR metric, which evaluates the harmony between the generated captions and reference captions in terms of both precision and recall, the UVLS model scores a 40.1%. Although this is slightly lower compared to the 'Ours; Transformer' variant, it is worth noting that METEOR also considers synonymy and paraphrasing, which could suggest areas for further refinement in the UVLS model. The CIDEr score of 164.2 is particularly noteworthy as it measures the consensus between the generated captions and a set of reference captions. This high score is indicative of the UVLS model's ability to produce unique and informative captions that are closely aligned with the semantic content of the images.

When we shift our attention to the Flickr8k-CN and COCO-CN datasets, the UVLS model maintains its superior stance, with the highest scores across most metrics, particularly BLEU@1 and CIDEr, on the Flickr8k-CN dataset. The COCO-CN dataset, known for its complexity and diversity, also sees the UVLS model performing robustly, especially in the BLEU@1 and BLEU@2 scores, which are essential indicators of the model's consistency in capturing the essence of the visual data across diverse image sets. These consistent results across varied datasets highlight the UVLS model's robustness and its sophisticated approach to capturing not just the details but also the essence and narrative of the visual scenes. The model's capability to perform well in these key metrics indicates its deep learning ability to comprehend and integrate the visual and linguistic components effectively, suggesting its potential as a valuable method in the field of image captioning.

The UVLS model's impressive 79.7% BLEU@1 score on the AIC-ICC dataset showcases its strength in generating coherent descriptions; however, this metric may not capture the full narrative depth or semantic richness of the captions. This suggests a need for critical examination of high BLEU scores as they may not always correlate with the quality of human-like narrative generation. Moreover, the progression to BLEU@2, BLEU@3, and BLEU@4 scores illustrates the model's ability to form more complex linguistic structures, yet it highlights the inherent limitation of these metrics in capturing the creative and idiomatic expressions found in human language. The top ROUGE score of 66.2% further indicates effective summarization, but this metric primarily measures keyword overlap, potentially overlooking subtler narrative elements or the emotional context depicted in images. Furthermore, the METEOR score, slightly lower than that of the 'Ours; Transformer' variant, prompts consideration of the model's capability to recognize synonyms and varied expressions, suggesting room for refinement to enhance naturalness and fluency in captions. While the CIDEr score of 164.2 showcases the model's ability to generate distinctive and informative captions, it also raises questions about the evaluation's emphasis on consensus over creativity. Across diverse datasets like Flickr8k-CN and COCO-CN, the UVLS model's consistent performance underscores its adaptability and depth in understanding visual content. However, this analysis invites further reflection on the model's handling of complex dataset diversity and its potential reliance on statistical regularities in training data, underscoring the importance of a nuanced approach to evaluating and advancing image captioning models.

Table 2 reveals that the UVLS model significantly enhances the performance of image captioning, surpassing baseline models across the AIC-ICC, Flickr8k-CN, and COCO-CN datasets. It exhibits exceptional proficiency in the BLEU@1 metric, achieving 79.7% on the AIC-ICC dataset, which underlines its superior ability to recognize and utilize key words in captions—a critical element for crafting coherent and precise descriptions. This proficiency is further evidenced by an impressive BLEU@4 score of 50.7%, suggesting

heightened ability to generate extensive context-rich narratives. Additionally, the UVLS model's performance in METEOR and CIDEr metrics points to its nuanced capability to create not only semantically precise captions but also ones that resonate with the unique and relevant content of images. The model's robustness across varied datasets underscores its adaptability and solidifies its standing as an advanced tool in image captioning. The integration of SegmentCLIP within the UVLS framework likely plays a crucial role in its outstanding performance. SegmentCLIP's clustering mechanism refines the model's ability to analyze and articulate complex contextual semantics within images, which is reflected in the elevated performance metrics. This sophisticated representation enables the UVLS model to offer a more detailed and comprehensive depiction of visual scenes, as evidenced by superior scores in the nuanced metrics of BLEU@4 and CIDEr, where the depiction of detailed relationships and distinctive content is crucial.

Table 2 reveals that the Unified Visual and Linguistic Semantics (UVLS) model significantly enhances the performance of image captioning, surpassing the baseline models across the AIC-ICC, Flickr8k-CN, and COCO-CN datasets. It exhibits exceptional proficiency in the BLEU@1 metric, achieving 79.7% on the AIC-ICC dataset, which underlines its superior ability to recognize and utilize key words in captions—a critical element for crafting coherent and precise descriptions. This proficiency is further evidenced by an impressive BLEU@4 score of 50.7%, suggesting a heightened ability to generate extensive context-rich narratives. Additionally, the UVLS model's performance in METEOR and CIDEr metrics points to its nuanced capability to create not only semantically precise captions but also ones that resonate with the unique and relevant content of images. The model's robustness across varied datasets underscores its adaptability and solidifies its standing as an advanced tool in image captioning. The integration of SegmentCLIP within the UVLS framework likely plays a crucial role in its outstanding performance. SegmentCLIP's clustering mechanism refines the model's ability to analyze and articulate complex contextual semantics within images, which is reflected in the elevated performance metrics. This sophisticated representation enables the UVLS model to offer a more detailed and comprehensive depiction of visual scenes, as evidenced by superior scores in the nuanced metrics of BLEU@4 and CIDEr, where the depiction of detailed relationships and distinctive content is crucial.

While the UVLS model marks a significant advancement in the field of image captioning, its nuanced understanding of visual and linguistic semantics is not without limitations. One area that warrants further examination is the model's handling of highly abstract or ambiguous images where the visual cues do not align clearly with conventional descriptions. The complexity inherent in such images poses a challenge to generating accurate and relevant captions, underscoring the need for more sophisticated interpretation mechanisms that can navigate the subtleties of visual ambiguity and abstract representations. This observation points towards an essential avenue for future research, focusing on the integration of deeper contextual and inferential reasoning capabilities within the UVLS framework.

Additionally, the variation in performance metrics observed across different datasets highlights the model's sensitivity to the nature and composition of the data. For instance, datasets with a higher degree of visual and thematic diversity tend to challenge the model's consistency in performance. This variability underscores the importance of diversifying training methodologies and incorporating adaptive learning strategies that can better accommodate the wide spectrum of visual and textual nuances present in varied datasets. Pursuing these enhancements will not only fortify the model's robustness but also expand its applicability and effectiveness across a broader range of image captioning scenarios.

In conclusion, the data-driven analysis of the UVLS model across multiple datasets underlines its comprehensive strengths, particularly in creating contextually relevant, linguistically coherent, and semantically rich captions. It showcases the UVLS model as a cutting-edge solution that can potentially revolutionize the landscape of image captioning by setting new benchmarks for accuracy, relevance, and narrative quality.

Furthermore, Table 3 measure the training time and the number of trainable parameters to validate the applicability of our method. Reducing the training time enables quickly obtaining a new model for new data, creating an ensemble of models, and decreasing energy consumption. Similar to other works, we report training time in GPU hours and the GPU model used. The number of trainable parameters is a popular measure to indicate model feasibility.

Table 3. Params and training times of different models on ICC dataset.

Model	Params (M)	Training Time (h)
CS-NIC	31	40 (GTX3090 (NVIDIA, Santa Clara, CA, USA))
CIC	42	56 (GTX3090)
SC	64	80 (GTX3090)
Oscar	135	200 (GTX3090)
CLIPCAP	156	20 (GTX3090)
Ours; MLP+GPT2 tuning	175	24 (GTX3090)
Ours; Transformer	63	23 (GTX3090)

5.4. Ablation Studies

To validate the specific strengths of our 'Ours; MLP+GPT2 tuning' model in precise language generation and semantic alignment, we conduct an ablation study delineated in Table 4. This experiment is designed to highlight the effectiveness of each component in our model by comparing it with other configurations and models.

Table 4. UVLS model validity experiment. Top scores for each metric are bold. All values are reported as percentage (%).

Model	BLEU@1	BLEU@2	BLEU@3	BLEU@4	METEOR	ROUGE	CIDEr
CLIPCAP	76.7	65.1	55.0	46.4	38.8	64.0	151.2
SegmentCLIP	74.2	63.4	53.8	45.2	37.6	63.7	150.4
Ours; Transformer	79.2	68.4	58.6	50.1	**40.3**	66.1	**164.9**
Ours; MLP+GPT2 tuning	**79.7**	**68.9**	**59.1**	**50.7**	40.1	**66.2**	164.2

Detailed Analysis of Model Performance: The 'Ours; MLP+GPT2 tuning' model demonstrates exceptional performance, as evidenced by its leading scores in the majority of the metrics. It surpasses other models with the highest BLEU@1 score of 79.7%, indicating its superior capability in capturing frequently used words. This trend continues with the highest BLEU@2, BLEU@3, and BLEU@4 scores (68.9%, 59.1%, and 50.7%, respectively), underscoring its proficiency in constructing longer, more complex linguistic structures. Additionally, the model achieves the top ROUGE score (66.2%), reflecting its effectiveness in summarizing the essential content of images. However, it slightly falls behind in the METEOR score (40.1%) and CIDEr score (164.2%) compared to the 'Ours; Transformer' variant, which could be attributed to the Transformer's slightly better grasp of syntactic variation and distinctiveness in caption generation.

Interpreting Strengths and Addressing Limitations: Despite the minor shortfall in METEOR and CIDEr, the 'Ours; MLP+GPT2 tuning' model's overall performance indicates a robust balance between linguistic fluency, n-gram coherence, and content relevance. The slight underperformance in METEOR and CIDEr may be due to the model's focus on grammatical and syntactical structure, which, while leading to high BLEU and ROUGE scores, might not capture the more nuanced semantic aspects as effectively as the 'Ours;

Transformer' variant. Nevertheless, the model's strong performance across most metrics, particularly in BLEU and ROUGE, highlights its capability to generate contextually rich and detailed image captions, suggesting its usefulness as a method in the realm of image captioning.

In our empirical investigation of prefix length's influence on model performance for image captioning, we present an in-depth analysis through Figure 3, which comprises Figure 3a,b, corresponding to two different model configurations: MLP and Transformer. These graphs illustrate the effects of prefix length variation on the captioning performance across the AIC-ICC dataset.

Figure 3a delineates the results from the MLP-based model configuration. It is evident that the BLEU@4 and CIDEr scores, which represent the model's linguistic accuracy and the salience of the generated captions, respectively, exhibit a substantial increase as the prefix length progresses from 2.5 to around 10. Beyond this point, the performance plateaus, indicating that the model has reached its optimal capacity for prefix length, after which no significant improvement is observed. The saturation point suggests a limitation in the MLP architecture, where additional length does not equate to better learning or generalization capabilities. This trend is marked by a leveling off in the scores, which remain consistent despite further increments in prefix length.

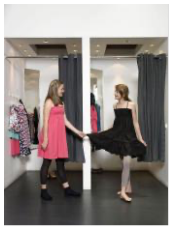
In contrast, Figure 3b shows the performance curve for the Transformer-based model. Here, we observe a sharp increase in both BLEU@4 and CIDEr scores as the prefix length is extended, with a notable improvement in performance continuing up until a prefix length of around 50. Past this juncture, the graph indicates a plateau, echoing the findings from the MLP model but at a substantially extended prefix length, underscoring the Transformer's superior ability to manage longer dependencies without a compromise in performance. The gradual ascent and subsequent stabilization of the scores highlight the Transformer's robustness and its aptitude for scaling with prefix length, up to the memory constraints imposed by the attention mechanism.

The comparative analysis of the MLP and Transformer configurations in Figure 3 reinforces the notion that an optimal prefix length is crucial for maximizing model performance. While both configurations show improvements in captioning performance with increased prefix lengths up to their respective saturation points, the Transformer model distinctly outperforms the MLP. This is attributed to the Transformer's architectural efficiency in handling longer sequences, which is well-aligned with the complex task of image captioning that often requires the integration of extensive contextual information. Thus, these results not only validate the findings of Li and Liang [48] regarding the existence of a beneficial range for prefix lengths but also explicate the differential impact of this range on varied architectural frameworks, contributing to a more nuanced understanding of model optimization in the realm of image captioning.

5.5. Case Studies

To demonstrate the advantages of our model (UVLS) over others like CLIPCAP and Oscar, we present example images with corresponding captions generated under the following setup: all the models ran on the AIC-ICC test set's 60,000 images with a beam size of 2. We randomly selected four examples of text generated from images to demonstrate the significant advantages of UVLS over the other models. The examples in Table 5 illustrate our model's finer and more detailed depiction of relationships between objects compared to the other models. Our model demonstrates quantitative superiority through enhanced precision in object relationship depiction. For instance, in the case of the boxing image in Table 5, UVLS does not just recognize the boxing activity; it goes a step further to accurately describe the action in a detailed manner—identifying the color of the gloves and the specific action of one boxer landing a punch on the other's chin. This level of detail is notably absent in the captions generated by other models such as CLIPCAP and Oscar, which provide more generalized descriptions of the scene.

Table 5. Performance comparison of image captioning methods: our approach vs. baseline methods on the AIC-ICC dataset.

Image				
CS-NIC	一个女人在喂羊	运动员在打网球	两个小女孩在房间玩耍	两个拳击手在打拳击
in English	A woman is feeding sheep.	The athlete is playing tennis.	Two little girls are playing in the room.	Two boxers are boxing.
CIC	女人在喂养羊	运动员在网球场打网球	两个女孩在更衣室玩耍	拳击手在擂台打拳击
in English	The woman is feeding the sheep.	The athlete is playing tennis on a tennis court.	Two girls are playing in the dressing room.	The boxer is fighting in the ring.
SC	穿着蓝色衣服的女人在喂养小羊	带着头巾的运动员在网球场打网球	两个穿着裙子的女孩在更衣室玩耍	两个拳击手在擂台打拳击
in English	woman dressed in blue is feeding a lamb.	The athlete wearing a headband is playing tennis on a tennis court.	Two girls wearing dresses are playing in the dressing room.	Two boxers are fighting in the ring.
Oscar	身穿蓝色衣服的女子在和小羊玩耍	穿着蓝色衣服的运动员在打网球	两个穿着裙子的小女孩在房间玩	两个拳击手在擂台比赛
in English	The woman in blue clothes is playing with a lamb.	The athlete dressed in blue is playing tennis.	Two little girls wearing dresses are playing in the room.	Two boxers are competing in the ring.
CLIPCAP	身穿蓝色衣服的女人在和一头白色的羊一起玩耍	穿着蓝色短袖的运动员在打网球	两个穿着裙子的小女孩在更衣室换衣服	两个拳击运动员在八角笼打拳击
in English	A woman in blue clothes is playing with a white sheep.	The athlete wearing a blue short-sleeved shirt is playing tennis.	Two little girls wearing dresses are changing clothes in the dressing room.	Two boxing athletes are fighting in an octagon cage.
Ours	一个身穿蓝色衣服的女人在和一头白色的羊一起玩耍	穿着蓝色短袖的运动员在网球场打网球	穿红色裙子的小女孩拉着另一个穿黑色裙子的女孩的衣服	带着红色拳套的拳击运动员一拳打在了另一个拳击运动员的下巴
in English	A woman wearing blue clothes is playing with a white sheep.	The athlete wearing a blue short-sleeved shirt is playing tennis on a tennis court.	The little girl wearing a red dress is pulling on the clothes of another girl wearing a black dress.	The boxer with red gloves landed a punch on the chin of another boxer.

Qualitatively, the UVLS model stands out for its ability to weave intricate details into a coherent narrative. The captions generated are not only contextually relevant but also rich in content, providing a comprehensive understanding of the images. The mentioned boxer image serves as a testament to this, where UVLS captures the dynamic nature of the sport, offering viewers a vivid portrayal of the intensity within the ring.

Finally, the contextual understanding of our model is evident in its ability to discern and articulate complex interactions between objects and their surroundings. The captions generated by UVLS convey a deeper level of interaction and a finer grasp of the activities being depicted, as observed in the boxing image where the precise nature of the interaction—down to the color of the gloves and the impact of the punch—is clearly communicated. This surpasses the capabilities of other models, which tend to provide more surface-level descriptions without delving into the specifics of the depicted events. In summary, the examples from Table 5 underscore our model's advanced ability to identify and describe intricate object relationships with a high degree of accuracy and detail, establishing UVLS as a notably superior model in the domain of image captioning.

6. Conclusions and Future Work

This study introduces the Unified Visual and Linguistic Semantics Method (UVLS), an innovative encoder–decoder framework that addresses key challenges in image captioning. The UVLS model is designed to bridge the gap between visual perception and linguistic interpretation, providing a more integrated and nuanced understanding of image content. It comprises three essential components: an encoder leveraging SegCLIP for efficient feature extraction from images, a mapping network for transforming these features into a consistent format, and a GPT-2-based decoder that generates accurate Chinese language captions. The integration of SegCLIP and SegmentCLIP within this framework enhances the semantic understanding between visual and textual elements, with SegmentCLIP's clustering mechanisms further refining the model's ability to capture complex contextual semantics.

Looking forward, our research in image captioning seeks to delve deeper into the intricacies of visual and linguistic interplay. Future work will focus on enhancing the model's ability to handle diverse and complex visual scenes, improving the granularity of feature extraction and linguistic translation. We aim to explore the potential of integrating additional contextual data sources and advanced natural language processing techniques to enrich the semantic depth of the generated captions. Additionally, efforts will be directed towards improving the model's adaptability and efficiency, ensuring its applicability across various domains and datasets. By continuously refining and advancing the UVLS model, we aspire to set new benchmarks in the field of image captioning, contributing to the development of more intelligent, context-aware, and linguistically sophisticated captioning systems.

Author Contributions: Conceptualization, J.P. and T.T.; methodology, J.P.; validation, J.P. and T.T.; formal analysis, J.P.; resources, J.P.; data curation, J.P.; writing—original draft preparation, J.P.; writing—review and editing, J.P.; supervision, T.T.; project administration, T.T.; funding acquisition, T.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number 62266004, Research on Explainable Deep Classification Models and Methods with Coherent Granule Embedding in Multimodal Data. Also, this research was funded by the National Natural Science Foundation of China, grant number 61762009, Efficient Reduction and Fusion of Heterogeneous Multimodal Data Based on Collaborative Granulation and Its Application.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

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

References

1. Derkar, S.B.; Biranje, D.; Thakare, L.P.; Paraskar, S.; Agrawal, R. Captiongenx: Advancements in deep learning for automated image captioning. In Proceedings of the 2023 3rd Asian Conference on Innovation in Technology, Pune, India, 25–27 August 2023; pp. 1–8.
2. Hossain, M.Z.; Sohel, F.; Shiratuddin, M.F.; Laga, H. A comprehensive survey of deep learning for image captioning. *ACM Comput. Surv.* **2019**, *51*, 1–36. [CrossRef]
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, Long Beach, CA, USA, 15–20 June 2019; pp. 4125–4134.
4. Zeng, A.; Attarian, M.; Ichter, B.; Choromanski, K.; Wong, A.; Welker, S.; Florence, P. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv* **2022**, arXiv:2204.00598.
5. Ghanem, F.A.; Padma, M.C.; Alkhatib, R. Automatic short text summarization techniques in social media platforms. *Future Internet* **2023**, *15*, 311. [CrossRef]
6. Can, Y.S.; Mahesh, B.; André, E. Approaches, applications, and challenges in physiological emotion recognition—A tutorial overview. *Proc. IEEE* **2023**, *111*, 1287–1313. [CrossRef]
7. Stefanini, M.; Cornia, M.; Baraldi, L.; Cascianelli, S.; Fiameni, G.; Cucchiara, R. From show to tell: A survey on deep learning-based image captioning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 539–559. [CrossRef] [PubMed]

8. Yan, C.; Hao, Y.; Li, L.; Yin, J.; Liu, A.; Mao, Z.; Chen, Z.; Gao, X. Task-adaptive attention for image captioning. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 43–51. [CrossRef]
9. Herdade, S.; Kappeler, A.; Boakye, K.; Soares, J. Image captioning: Transforming objects into words. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; p. 32.
10. Yao, T.; Pan, Y.; Li, Y.; Mei, T. Hierarchy parsing for image captioning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 2621–2629.
11. He, S.; Liao, W.; Tavakoli, H.R.; Yang, M.; Rosenhahn, B.; Pugeault, N. Image captioning through image transformer. In Proceedings of the Asian Conference on Computer Vision (ACCV), Kyoto, Japan, 30 November–4 December 2020.
12. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 2048–2057.
13. 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 (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 2048–2057.
14. Mokady, R.; Hertz, A.; Bermano, A.H. Clipcap: Clip prefix for image captioning. *arXiv* **2021**, arXiv:2111.09734.
15. Xu, J.; De Mello, S.; Liu, S.; Byeon, W.; Breuel, T.; Kautz, J.; Wang, X. Groupvit: Semantic segmentation emerges from text supervision. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June 2022; pp. 18134–18144.
16. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language models are unsupervised multitask learners. *OpenAI Blog* **2019**, *1*, 9.
17. Liu, X.; Xu, Q.; Wang, N. A survey on deep neural network-based image captioning. *Vis. Comput.* **2019**, *35*, 445–470. [CrossRef]
18. Geetha, G.; Kirthigadevi, T.; Ponsam, G.G.; Karthik, T.; Safa, M. Image captioning using deep convolutional neural networks. *Proc. J. Phys. Conf. Ser.* **2020**, *1712*, 012015. [CrossRef]
19. Liu, S.; Bai, L.; Hu, Y.; Wang, H. Image captioning based on deep neural networks. *Proc. Matec Web Conf.* **2018**, *232*, 01052. [CrossRef]
20. Yang, L.; Wang, H.; Tang, P.; Li, Q. CaptionNet: A tailor-made recurrent neural network for generating image descriptions. *IEEE Trans. Multimed.* **2020**, *23*, 835–845. [CrossRef]
21. Chen, X.; Ma, L.; Jiang, W.; Yao, J.; Liu, W. Regularizing rnns for caption generation by reconstructing the past with the present. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 8–22 June 2018; pp. 7995–8003.
22. Wang, J.; Wang, W.; Wang, L.; Wang, Z.; Feng, D.D.; Tan, T. Learning visual relationship and context-aware attention for image captioning. *Pattern Recognit.* **2020**, *98*, 107075. [CrossRef]
23. Wang, C.; Shen, Y.; Ji, L. Geometry attention transformer with position-aware LSTMs for image captioning. *Expert Syst. Appl.* **2022**, *201*, 117174. [CrossRef]
24. Zohourianshahzadi, Z.; Kalita, J.K. Neural attention for image captioning: Review of outstanding methods. *Artif. Intell. Rev.* **2022**, *7*, 3833–3862. [CrossRef]
25. Huang, L.; Wang, W.; Chen, J.; Wei, X.Y. Attention on attention for image captioning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 4634–4643.
26. Pedersoli, M.; Lucas, T.; Schmid, C.; Verbeek, J. Areas of attention for image captioning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1242–1250.
27. Wang, W.; Chen, Z.; Hu, H. Hierarchical attention network for image captioning. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 8957–8964.
28. 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*; Association for the Advancement of Artificial Intelligence: New York, NY, USA, 2020; Volume 34, pp. 13041–13049.
29. Wang, W.; Yang, Z.; Xu, B.; Li, J.; Sun, Y. ViLTA: Enhancing vision-language pre-training through textual augmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 3158–3169.
30. Li, Y.; Fan, J.; Pan, Y.; Yao, T.; Lin, W.; Mei, T. Uni-EDEN: Universal encoder-decoder network by multi-granular vision-language pre-training. *ACM Trans. Multimed. Comput. Commun. Appl.* **2022**, *18*, 48. [CrossRef]
31. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning, Virtual Event, 18–24 July 2021; pp. 8748–8763.
32. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
33. Wu, J.; Zheng, H.; Zhao, B.; Li, Y.; Yan, B.; Liang, R.; Wang, W.; Zhou, S.; Lin, G.; Fu, Y. AI challenger: A large-scale dataset for going deeper in image understanding. *arXiv* **2017**, arXiv:1711.06475.
34. Li, X.; Lan, W.; Dong, J.; Liu, H. Adding chinese captions to images. In Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval, New York, NY, USA, 6–9 June 2016; pp. 271–275.

35. Li, X.; Xu, C.; Wang, X.; Lan, W.; Jia, Z.; Yang, G.; Xu, J. COCO-CN for cross-lingual image tagging, captioning, and retrieval. *IEEE Trans. Multimed.* **2019**, *21*, 2347–2360. [CrossRef]
36. 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, Philadelphia, PA, USA, 6–12 July 2002; pp. 311–318.
37. Banerjee, S.; Lavie, A. Meteor: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Ann Arbor, MI, USA, 29 June 2005; pp. 65–72.
38. Lin, C.Y. Rouge: A package for automatic evaluation of summaries. In Proceedings of the Text Summarization Branches Out 2004, Barcelona, Spain, 25–26 July 2004; pp. 74–81.
39. Vedantam, R.; Lawrence Zitnick, C.; 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.
40. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jégou, H. Training data-efficient image transformers & distillation through attention. In Proceedings of the International Conference on Machine Learning, Virtual Event, 18–24 July 2021; pp. 10347–10357.
41. Tolstikhin, I.O.; Houlsby, N.; Kolesnikov, A.; Beyer, L.; Zhai, X.; Unterthiner, T.; Dosovitskiy, A. Mlp-mixer: An all-mlp architecture for vision. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 24261–24272.
42. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
43. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. *arXiv* **2017**, arXiv:1711.05101.
44. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Rush, A.M. Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Virtual Event, 16–20 November 2020; pp. 38–45.
45. Aneja, J.; Deshpande, A.; Schwing, A.G. Convolutional image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 5561–5570.
46. Gu, J.; Cai, J.; Wang, G.; Chen, T. Stack-captioning: Coarse-to-fine learning for image captioning. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; p. 32.
47. 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 Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; pp. 121–137.
48. Li, X.L.; Liang, P. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv* **2019**, arXiv:2101.00190.