buildings

MDPI

Article

Enabling High-Level Worker-Centric Semantic Understanding of Onsite Images Using Visual Language Models with Attention Mechanism and Beam Search Strategy

Hui Deng [1], Kejie Fu [1], Binglin Yu [1], Huimin Li [1], Rui Duan [1], Yichuan Deng [1,2,*] and Jia-rui Lin [3]

[1] School of Civil Engineering and Transportation, South China University of Technology, Guangzhou 510641, China; hdeng@scut.edu.cn (H.D.); 202321008383@mail.scut.edu.cn (K.F.); 202220107333@mail.scut.edu.cn (B.Y.); 202321008393@mail.scut.edu.cn (H.L.); 202021008728@mail.scut.edu.cn (R.D.)
[2] State Key Laboratory of Subtropical Building and Urban Science, Guangzhou 510641, China
[3] Department of Civil Engineering, Tsinghua University, Beijing 100084, China; lin611@tsinghua.edu.cn
* Correspondence: ctycdeng@scut.edu.cn

Abstract: Visual information is becoming increasingly essential in construction management. However, a significant portion of this information remains underutilized by construction managers due to the limitations of existing image processing algorithms. These algorithms primarily rely on low-level visual features and struggle to capture high-order semantic information, leading to a gap between computer-generated image semantics and human interpretation. However, current research lacks a comprehensive justification for the necessity of employing scene understanding algorithms to address this issue. Moreover, the absence of large-scale, high-quality open-source datasets remains a major obstacle, hindering further research progress and algorithmic optimization in this field. To address this issue, this paper proposes a construction scene visual language model based on attention mechanism and encoder–decoder architecture, with the encoder built using ResNet101 and the decoder built using LSTM (long short-term memory). The addition of the attention mechanism and beam search strategy improves the model, making it more accurate and generalizable. To verify the effectiveness of the proposed method, a publicly available construction scene visual-language dataset containing 16 common construction scenes, SODA-ktsh, is built and verified. The experimental results demonstrate that the proposed model achieves a BLEU-4 score of 0.7464, a CIDEr score of 5.0255, and a ROUGE_L score of 0.8106 on the validation set. These results indicate that the model effectively captures and accurately describes the complex semantic information present in construction images. Moreover, the model exhibits strong generalization, perceptual, and recognition capabilities, making it well suited for interpreting and analyzing intricate construction scenes.

Keywords: visual language model; construction scene; image scene understanding; image captioning; attention mechanism

Academic Editor: Yasser Mohamed

Received: 7 February 2025
Revised: 12 March 2025
Accepted: 14 March 2025
Published: 18 March 2025

Citation: Deng, H.; Fu, K.; Yu, B.; Li, H.; Duan, R.; Deng, Y.; Lin, J.-r. Enabling High-Level Worker-Centric Semantic Understanding of Onsite Images Using Visual Language Models with Attention Mechanism and Beam Search Strategy. *Buildings* **2025**, *15*, 959. https://doi.org/10.3390/buildings15060959

1. Introduction

With the widespread adoption of camera-equipped devices, including smartphones, head-mounted cameras, and unmanned aerial vehicles (UAVs), construction project managers increasingly rely on video equipment to efficiently, intelligently, and conveniently capture real-time construction site conditions [1]. The recorded visual data can be analyzed and interpreted to assist project managers in identifying environmental risks [2], overseeing worker activities [3], and monitoring equipment utilization [4]. This enables more effective

management and control of the construction process, enhancing overall project efficiency and safety. However, manually analyzing and interpreting the vast amounts of visual data generated by these devices is time-consuming and challenging [5]. Therefore, automatically capturing visual-semantic information from visual data has become an essential requirement for the advancement of intelligent applications in the construction industry [6].

In recent years, computer vision techniques combined with deep learning algorithms have been widely used to capture visual-semantic information from visual data [7], with four main categories of approaches: object recognition [8,9], motion recognition [10,11], activity recognition [12], and scene analysis [13,14]. While current research in computer vision primarily focuses on single element identification (e.g., objects, relationships, and attributes) from images, there is less attention given to finer-grained semantic information such as the relationships between different objects and contextual information of scenes. However, the construction site environment is highly complex, with limited space and high personnel density, and involves multiple types of construction machinery and tools [15,16]. Furthermore, complex interactions exist between workers, machinery, and tools [17]. There is a need for an automated capture method that treats the construction activity scene as an integral whole by capturing the multiple objects, relationships, and related attributes from visual data and converting them into complete sentences [18].

The visual language model based on computer vision (CV) and natural language processing (NLP) provides a solution to automatically capturing visual-semantic information [19]. By describing these images, text captions can be generated that are fluent in semantic content and language, aiding in the perception and understanding of construction scenes [20]. A complete visual language model consists of a deep learning model and datasets. The most commonly used model is the encoder–decoder, consisting of two independent modules: an encoder based on convolutional neural networks (CNNs) for feature extraction from images and object recognition, and a decoder based on recurrent neural networks (RNNs) for decoding the semantic features extracted by the encoder, recognizing object relationships and attributes, and generating captions that match the image content [21]. In addition, Transformer, as a powerful deep learning model, can effectively capture the complex relationship between images and texts by using its self-attention mechanism. The development of multimodal large language models (MLLMs) has also brought new opportunities for image description tasks. These models can generate high-quality image descriptions with zero or few samples by fusing visual and language information. However, these methods generally have the problems of low flexibility, poor interpretability, and high computing resource requirements. The level of model performance is dependent on the size and quality of the training dataset [22]. However, existing visual-language datasets (e.g., MS-COCO [23], Flickr8k [24], Flickr30K dataset [25]) do not encompass construction-related scenes or are mainly focused on general construction scenes.

In addressing these needs, this research involves two innovations. First, a new image semantic model is built, which combines attention mechanisms and beam search with Resnet101 as encoder and LSTM (long short-term memory) as decoder. The improvement of beam search on model performance is verified by ablation experiments. Second, in order to verify the effectiveness of the method, a visual-language dataset based on the construction scene is constructed by crowdsourcing, and the model is trained and tested on the dataset. The experimental results show that the model has a higher performance score and stronger attribute-learning ability. In order to benefit future research, the constructed dataset, namely SODA-ktsh, is publicly shared with its full images and annotations.

In the following section, a brief outline is given of the extracted visual-semantic information from construction site images (Section 2.1), along with a brief discussion on

the current research state regarding encode–decode (Section 2.2); then, a brief overview is presented of existing image datasets (Section 2.3). Section 3 outlines the framework of the model and developing process of datasets, and the experimental testing is described in Section 4. Section 5 discusses the research results, contribution to knowledge, practical implications, application challenges, and research limitations, while Section 6 briefly summarizes this paper and discusses future work.

2. Related Work

2.1. Visual-Semantic Information Extracting from Construction Site Images

In recent years, with the development of computer vision and deep learning algorithms, more and more research has been devoted to extracting visual-semantic information from construction site images. The existing research areas primarily encompass object detection and recognition, motion recognition and tracking, activity recognition and analysis, and scene understanding and analysis.

Object detection and recognition involve identifying and localizing objects from construction site images. Traditional approaches have relied on feature extraction and machine learning algorithms such as Histograms of Oriented Gradients (HOG) [26], Scale-Invariant Feature Transform (SIFT) [27], and Speeded Up Robust Features (SURF) [28]. These methods demonstrate notable reliability and stability in simple scenes across varying fields of view, lighting conditions, individual poses, and occlusions. However, their accuracy significantly declines when applied to complex scenes. In recent years, deep learning has emerged as the mainstream approach for object recognition in the construction industry. Common deep learning methods for object recognition include Faster R-CNN [29], You Only Look Once (YOLO) [30,31], and Single Shot Multi-Box Detector (SSD) [8].

Object detection and recognition have been applied in various domains within the construction industry, including material tracking [30,32], equipment monitoring [8,9,33], and safety management [31,34]. For example, Li et al. [30] proposed a deep learning approach based on a YOLOv3 detector for automatic steel bars detection and counting through images. Wu et al. [8] proposed a single-stage data-driven CNN method based on the SSD framework for detecting the wearing of safety helmets by workers. Nath et al. [9] proposed three YOLO-based deep learning models that aim to verify if workers are wearing proper personal protective equipment (PPE).

Motion recognition and tracking is another essential field of research for analyzing and comprehending the activities and behaviors of workers on construction sites, through the systematic analysis of camera-captured movement data [35]. Recognizing and tracking workers' postures can be utilized to estimate their health and safety behavior status and identify potential safety hazards and ergonomic risks in the workplace [36,37]. To this end, several studies have proposed motion recognition models for construction workers. Kim et al. [36] proposed an LSTM-based model designed to assess worker safety and productivity by analyzing individual movement patterns. Similarly, Antwi-Afari et al. [38] proposed a wearable insole pressure system and an RNN model for identifying and classifying awkward working postures among different types of construction work. To address the challenge of low accuracy in worker gesture recognition, Wang et al. [38] employed two state-of-the-art 3D convolutional neural networks (CNNs), ResNeXt-101, and Res3D+ConvLSTM+MobileNet, to enhance gesture recognition.

Activity recognition is a research field that involves the automatic detection and recognition of human activities from visual data, such as images or videos. Recent studies have focused on improving activity recognition in the context of the built environment through various techniques and datasets [12,39–41]. For instance, Yang et al. [41] used advanced video description methods with dense trajectories to recognize actions and

developed a new real-world video dataset with 1176 instances. Similarly, Luo et al. [39] used deep learning and Faster R-CNN to identify construction-related objects in static site images and construct an association network for activity recognition. They matched activity patterns defined by relationships between detected objects. Although this method lacks temporal information, the temporal segment network (TSN) has been used to recognize building activities in surveillance videos.

The understanding and analysis of architectural scenes in the field of architecture have received relatively less attention in the academic community compared to other fields. However, in recent years, some research has begun to focus on utilizing image captioning technology to achieve a better understanding of construction scenes. For instance, Bang et al. [14] proposed an image captioning model based on CNN and LSTM networks, which aims to generate textual descriptions encompassing information about the location, status, motion, color, and quantity of architectural resources. By learning the semantic correspondence between images and text, their model is capable of generating accurate descriptions of architectural scenes. Similarly, Liu et al. [17] introduced a CNN-LSTM-based image captioning model specifically designed for describing construction activities within construction scenes. Their model automatically identifies activities taking place in the construction site and generates descriptive textual information associated with these activities, offering a detailed understanding of the construction process.

These studies indicate the potential of utilizing image captioning techniques for addressing the understanding of construction scenes. In the following section, a more detailed discussion will be presented.

2.2. Image Captioning

Image captioning refers to the generation of textual descriptions that capture the content of an image. This task involves analyzing and understanding the visual elements of the image, such as objects, people, and scenes, and generating coherent and relevant descriptions to communicate the image content to the audience. Image captioning is a challenging task that requires the integration of computer vision, natural language processing, and deep learning techniques. Common approaches to image captioning include template-based, retrieval-based, and deep learning-based methods.

Template-based image captioning relies on predefined descriptive templates and extracts relevant information from the image to fill them [42]. While simple to implement, these methods struggle to generate diverse and creative descriptions. The fixed templates limit flexibility and hinder the production of unique and imaginative captions. In template-based image captioning, the image is analyzed to extract visual features using techniques like convolutional neural networks (CNNs) [43]. These features are then mapped to slots in the template representing objects, actions, or attributes. Words or phrases are selected from the extracted features or a predefined vocabulary to fill the template, forming a coherent caption [44]. However, template-based methods lack flexibility, resulting in generic captions that fail to capture the image's nuanced details and characteristics.

Retrieval-based image captioning retrieves captions from a pre-existing database and selects the most relevant description for a given image [45]. This approach relies on an annotated image database, comparing image features using similarity measures like cosine similarity or Euclidean distance. The most relevant captions are chosen based on similarity scores or ranking. However, retrieval-based methods have limited variability as they rely on the available captions in the database. This can lead to repetitive and generic descriptions that do not capture the unique nuances of the specific image being captioned.

To overcome the limitations of both template-based and retrieval-based methods, recent advancements in image captioning have focused on developing more sophisticated

and adaptive approaches. These innovative techniques have revolutionized the way neural networks are trained to generate descriptive captions for a given image [19]. One prevalent and successful approach involves the utilization of a convolutional neural network (CNN) for extracting essential features from the image, coupled with a recurrent neural network (RNN) responsible for generating corresponding descriptions [18]. This powerful combination, known as the encoder–decoder model, has been extensively studied and proven effective in generating coherent and semantically meaningful captions.

In recent years, language models based on the Transformer architecture, such as Bidirectional Encoder Representations from Transformers (BERT), Generative Pre-trained Transformers (GPTs), and the Text-to-Text Transfer Transformer (T5), have become mainstream solutions for natural language processing (NLP) and computer vision (CV) tasks. The Transformer architecture has revolutionized traditional sequence modeling methods through its self-attention mechanism and parallel computing advantages, overcoming the limitations of recurrent neural networks (RNNs) and long short-term memory networks (LSTMs) in terms of long-distance dependencies and computational efficiency.

BERT uses a bidirectional Transformer encoder to better understand contextual information [46]. Zhang et al. [47] proposed an automatic hazard inference method using construction scene graphs and C-BERT networks. GPT is based on an autoregressive generative architecture and uses a unidirectional Transformer decoder to generate text, with powerful natural language generation capabilities [48]. Curtò et al. [49] used a CLIP-based captioning technique and a YOLOv7 detector method to use drones for scene understanding and enhanced the generated text by prompting GPT natural language instructions. T5 unifies all tasks into a text-to-text conversion form, further promoting multi-task learning and model versatility. Tsai et al. [50] proposed a framework for automatically identifying safety violations in construction images through image descriptions. The model generates descriptive captions with multiple embeddings to extract information from images using image-to-text techniques. These models not only perform well in the field of natural language processing but also play an important role in tasks such as image captioning and visual question answering (VQA) through cross-modal extension [51].

However, these methods generally have the problems of low flexibility, low interpretability, and high computing resource requirements. The language models based on the Transformer architecture (such as BERT, GPT, and T5) have complex frameworks, deep neural network layers, and are not easy to change and have low flexibility. The traditional encoder–decoder model can flexibly select the network architecture suitable for the encoder and decoder parts. Secondly, the high complexity and depth of modern network architectures further exacerbate the issue of low model interpretability. In contrast, traditional encoder–decoder frameworks provide greater interpretability in the transformation process from image input to text output, offering a more transparent understanding of the underlying mechanisms. In addition, the general LLM model often does not work well when facing problems in specific fields [52]. At this time, the model needs to be fine-tuned, and model fine-tuning usually requires the support of a large amount of high-quality data and high-performance equipment. However, in most actual projects, the conditions for model fine-tuning cannot be configured, so the problem of high computing resource requirements of language models based on the Transformer architecture cannot be ignored.

While the encoder–decoder model has shown remarkable performance in image captioning tasks, researchers have explored further enhancements to improve its capabilities. One notable extension involves the incorporation of attention mechanisms into the encoder–decoder framework [7]. Attention mechanisms enable the model to selectively focus on different regions or aspects of the image when generating captions, mimicking the human ability to attend to specific visual details [53–56].

The attention-based image captioning architecture proposed by Zhai et al. [56] exemplifies this approach. Their model combines the power of CNNs for image feature extraction with long short-term memory (LSTM) networks in the decoder to encode word positions. By incorporating attention mechanisms, the model gains the ability to dynamically attend to relevant image regions while generating captions, effectively aligning the visual content with the textual descriptions. This attention-driven approach has demonstrated substantial improvements in caption quality and has gained significant traction in recent research.

After applying a linear layer to transform the decoder's output into scores for each word in the vocabulary, a greedy algorithm is typically used to select the highest-scoring word at each time step and predict the next word. However, this approach often fails to yield the optimal final output. To address this limitation, this study employs beam search, which has been widely demonstrated in previous research to significantly outperform the greedy algorithm in generating more accurate and coherent sequences. Huang [57] proposed an optimal beam search algorithm for neural text generation, addressing the challenge of determining the termination point of beam search to ensure optimality. The study demonstrated that beam search effectively enhances the quality of generated text across various neural text generation tasks, including neural machine translation, text summarization, and image captioning. Similarly, Freitag [58] examined the performance of beam search in machine translation, comparing it with greedy decoding. The findings indicated that beam search enhances translation performance when applied to a given parallel corpus.

Building upon these existing foundations, we aim to further advance the performance of our image captioning model. By leveraging the power of deep learning, attention mechanisms, and beam search, we seek to enhance the model's ability to generate accurate, descriptive, and contextually relevant captions. The integration of attention mechanisms allows the model to selectively attend to salient visual features, thereby enriching the captions with more detailed and informative descriptions. The incorporation of beam search enables the model to generate more accurate sentences.

2.3. Overview of Existing Image Captioning Datasets

The MS-COCO dataset [23] is a widely used dataset for computer vision tasks, including image captioning. It contains 164,062 images, including 82,783 training images, 40,504 validation images, and 40,775 test images, with at least five captions per image, except for the test set. The training set has 41,411 captions, and the validation set has 202,654 captions, all recorded in a JSON file.

In 2013, researchers publicly released the Flickr8K dataset [24], which consists of 8000 images sourced from the photo and image sharing website Flickr. Compared to MS-COCO, the dataset has a smaller scale and primarily focuses on images featuring people and animals. The image labels in the dataset were manually annotated through Amazon's crowdsourcing platform. Each image in the dataset is accompanied by five descriptive sentences.

The Flickr30K dataset [25] is an extension of Flickr8K, consisting of around 31,000 image data and 158,000 manually annotated image captions sourced mainly from Yahoo's Flickr website. MS-COCO and Flickr30K are general image captioning datasets and do not include construction scene image captions.

In addition to the aforementioned publicly available datasets, some researchers have developed specialized datasets specifically for construction scenes. For example, Liu et al. [18] created two unique image captioning datasets, which include 7382 images covering five categories of general construction scenes: masonry work, wheelbarrow work, rebar work, plastering work, and tiling work. Bang et al. [14] constructed a dedicated image

caption dataset consisting of 1431 images and 8601 captions describing the image regions captured by the UAVs at six different construction sites. Wang et al. [59] annotated over 6000 images selected from ACID with natural language captions, collecting 2–3 captions for each image, to create the ACID-C dataset, which was used for information mining and scene caption in large construction machinery construction scenes based on object detection and semantic information extraction models.

Construction sites have complex and dense spatial interactions, and the datasets mentioned above lack construction-related scenes or are mainly focused on general daily life scenes. The research on construction scenes is relatively limited, the fine-grained information of construction scenes is difficult to excavate, and there is limited open-source sharing of such datasets. In our previous work, we have already established a dataset called Site Object Detection Dataset (SODA), which includes 15 object classes categorized into workers, materials, machines, and layouts. We collected over 20,000 images from multiple construction sites, covering different perspectives, angles, situations, weather conditions, and construction stages [60]. Building upon this dataset, we will construct a new dataset specifically for training, validation, and testing of image captioning models in the context of construction scenes.

3. Methodology

Figure 1 shows the proposed framework of this study. The approach described here involves the development of the visual language model and the datasets.

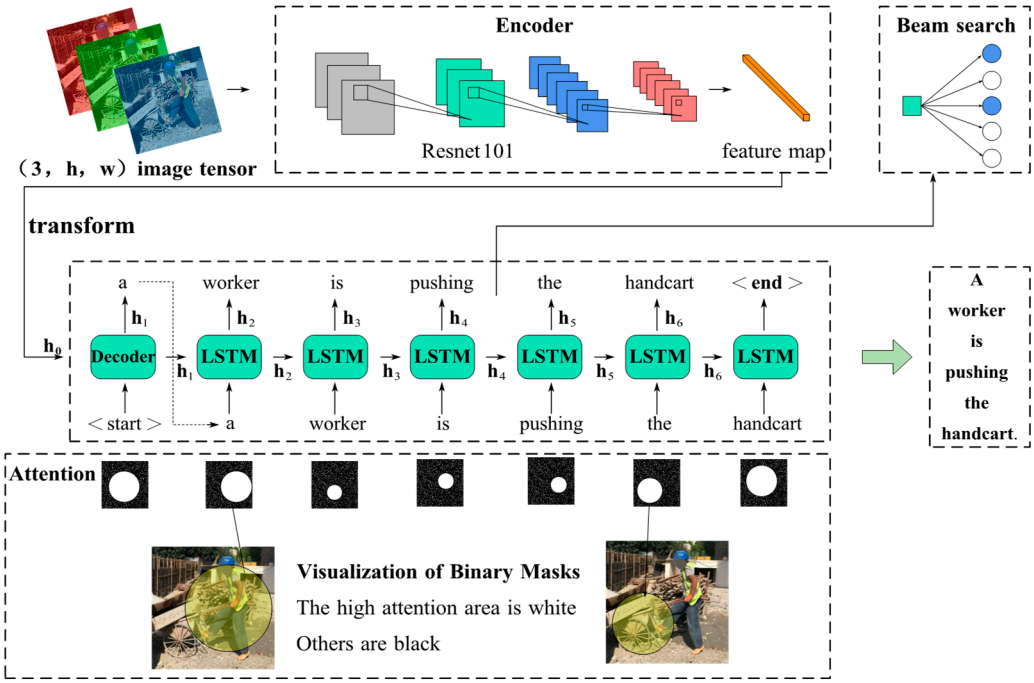


Figure 1. Flow diagram of the proposed framework in this study.

3.1. Visual Language Model

1. Encoder and decoder

The encoder–decoder architecture is the most commonly used model for the visual language model, which consists of two independent modules. Firstly, the CNN is used to perceive visual information from the image, and the output of the last layer of the CNN is used as input to the RNN. Based on the output of the RNN, a semantic caption of the current image is generated.

In this paper, Resnet101 is used as an encoder to extract features from construction images. The traditional CNN is stacked with a series of convolutional and down-sampling layers. However, when the network depth reaches a certain level, training becomes very difficult, the accuracy of the model will begin to saturate, and there will be gradient vanishing and exploding and degradation problems. As shown in Figure 2, Resnet101 uses a residual network to enable the construction of deep networks and uses batch normalization to accelerate training. The residual network alleviates the degradation problem and performs better as the network depth increases. In this paper, pre-trained ResNet-101 weights are used, and the last pooling layer and fully connected layer in the network are discarded, with the output of the last convolutional layer in the model taken as the input to the decoder.

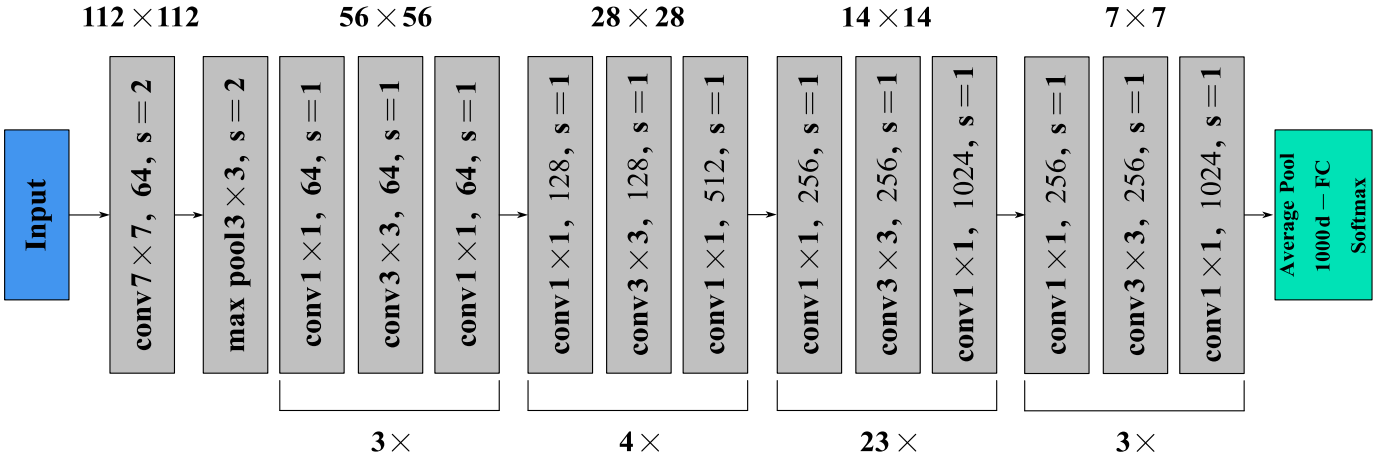


Figure 2. The structure of Resnet101.

As shown in Figure 3, the RNN neural unit in the traditional RNN chain structure is relatively simple, with only one information propagation layer containing a tanh activation function, which cannot guarantee long-term memory characteristics. Therefore, in this paper, the decoder adopts LSTM in the RNN to dynamically integrate image semantic feature information and generate caption sentences with temporal consistency. LSTM can generate the current state of the word at each time state based on the previous context information and selectively retain or discard the information.

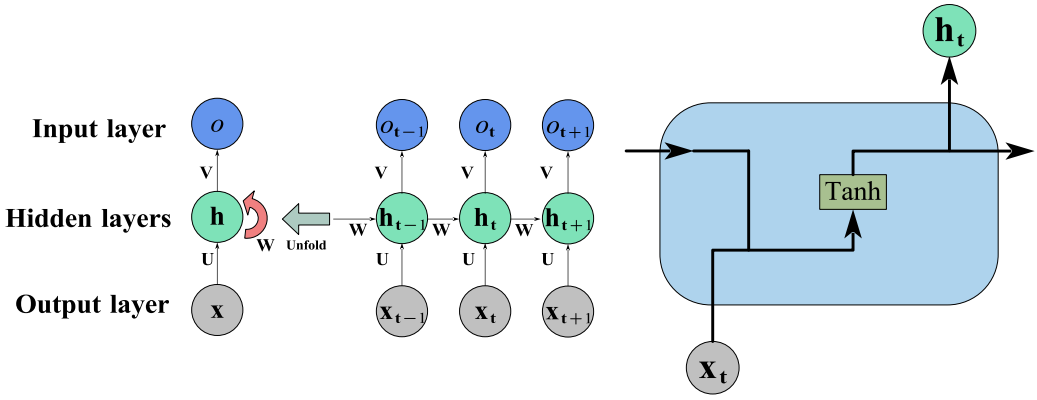


Figure 3. The chain structure of RNN.

As shown in Figure 4, the LSTM neuron structure is mainly composed of an input gate, forget gate, output gate, and memory cell. x_t represents the input at the current time step, h_t represents the output of the forget gate at the current time step, C_t represents the state of

the memory cell at the current time step, o_t represents the information of the output gate at the current time step, W represents the weights, b represents the bias parameters, and σ represents the Sigmoid activation function. The symbol "$\otimes$" denotes the vector product, and the symbol "$\oplus$" denotes the vector sum.

$$ft = \sigma(Wf[h_{t-1}, x_t] + b_f) \tag{1}$$

$$it = \sigma(Wi[h_{t-1}, x_t] + b_i) \tag{2}$$

$$\widetilde{C}t = tanh(Wc[h_{t-1}, x_t] + b_c) \tag{3}$$

$$C_t = f_t \times C_{t-1} + i_t \times \widetilde{C}_t \tag{4}$$

$$o_t = \sigma(W\sigma[h_{t-1}, x_t] + b_\sigma) \tag{5}$$

$$h_t = o_t \times tanh(C_t) \tag{6}$$

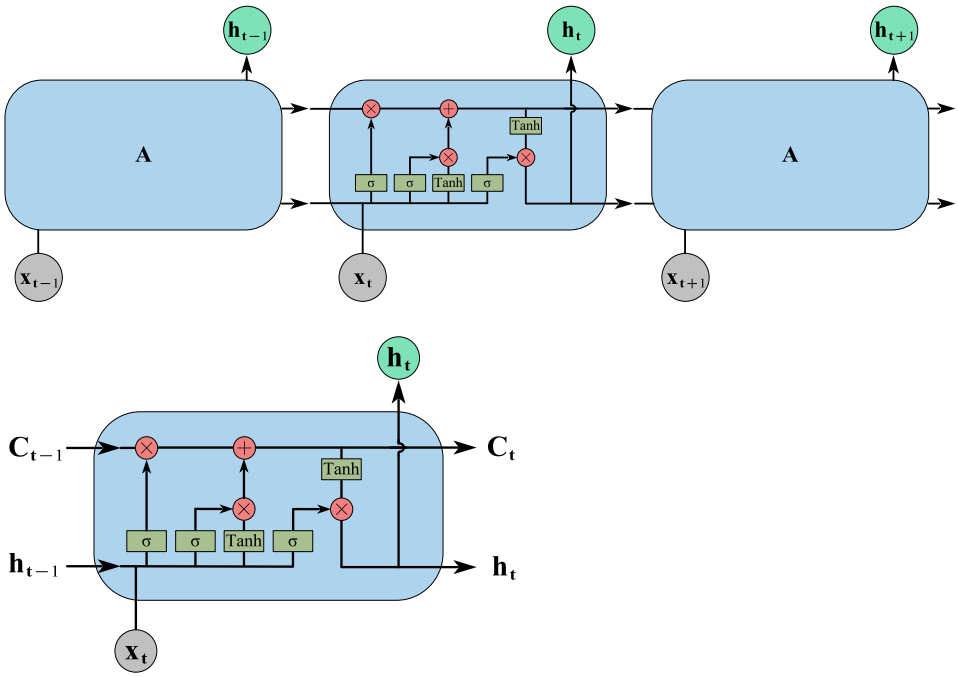


Figure 4. The neuron structure of LSTM.

In the model, x_t and h_{t-1} are passed through the forget gate, which uses the σ function as shown in Equation (1) to output f_t and decide which information to discard by multiplying it with C_{t-1}. Then, x_t and h_{t-1} are passed through the input gate, which uses the σ function as shown in Equation (2) to output it, representing the decision of which information to keep for update. Then, a new candidate value (C_t.) is created using the tanh function as shown in Equation (3). The old state, C_{t-1}, is updated to Ct using the output of the forget gate and the input gate, as shown in Equation (4). The output gate, which also uses the σ function, is used to output o_t based on x_t and h_{t-1}. Then, a new state, C_t, is processed using the *tanh* function and combined with o_t using new control parameters to generate the final output, as shown in Equation (5). Finally, a new h_t is generated as shown in Equation (6). The LSTM network continuously updates its control parameters C and selectively passes and discards information using the memory gate, forget gate, and output gate, generating sequential data.

2. Attention Mechanism

Due to the complexity and strong noise interference of construction sites, using CNN to process images tends to make the network focus on the areas that need attention, so that a limited-length vector can adaptively pay attention to important objects. Therefore, an attention mechanism is added to the decoder. The attention mechanism is a resource allocation scheme that allocates computing resources to more important tasks and solves the problem of information overload when computing capacity is limited [61]. The attention mechanism is commonly classified into two types: soft attention and hard attention [19]. Soft attention is a probability-based mechanism that assigns a weight to each input position, indicating its importance to the model's output. Hard attention, on the other hand, is a sampling-based mechanism that selects a subset of the input data for the model's output calculation. This paper adopts the soft attention mechanism, as it allows for weighted processing of the input image at each time step, enabling dynamic focusing on different regions of the input image when generating each word, resulting in more accurate performance. In addition, the differentiability of the soft attention mechanism enables it to be directly trained using the backpropagation algorithm, which is superior to the discrete and non-differentiable hard attention mechanism.

The soft attention mechanism can be divided into channel attention mechanism and spatial attention mechanism. The structures of the channel attention mechanism and spatial attention mechanism are shown in Figure 5. The channel attention mechanism focuses on the features of the image channels and alters the model's attention distribution by amplifying or reducing certain channel features. This involves global average and max pooling of the input feature map, followed by a shared fully connected layer. The output is then processed with a sigmoid activation function to obtain channel weights (Mc) between 0 and 1. The spatial attention mechanism focuses on the spatial structure of an image by amplifying or reducing certain regions, which alters the model's attention. To do this, the input feature map undergoes global average and max pooling, followed by a convolutional layer. The resulting output is passed through a Sigmoid function, resulting in a spatial weight map (Ms) with values between 0 and 1.

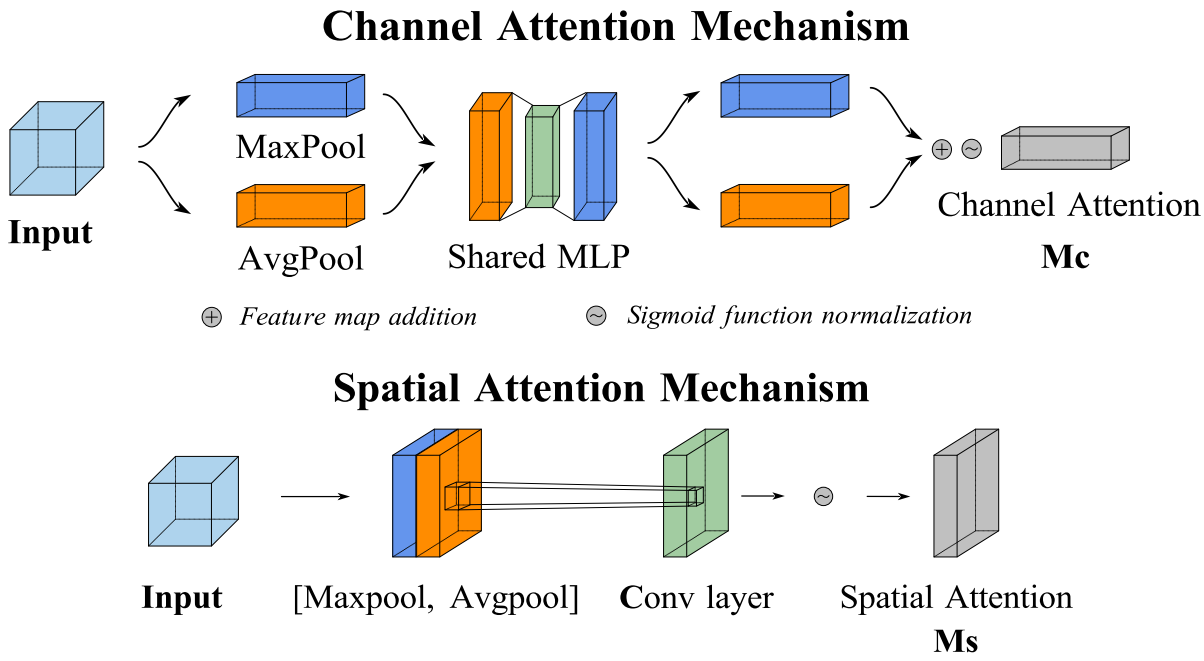


Figure 5. The structure of soft attention mechanism.

By introducing the channel attention mechanism, the neural network can better focus on key features, thereby improving the overall performance. For example, in image classification tasks, the channel attention mechanism can help the network more accurately identify key objects in the image; in target detection tasks, it can help the network better locate the position of the target object. By concentrating on the most salient regions within an image, the spatial attention mechanism effectively enhances both the accuracy and efficiency of the model. By selectively emphasizing critical features while suppressing less relevant information, this mechanism facilitates more precise feature extraction and improves overall model performance. For example, in target detection tasks, the spatial attention mechanism can help the model locate the target object more accurately and reduce the impact of background noise.

Since the CNN is used to process images, in previous studies ("SODA: A large-scale open site object detection dataset for deep learning in construction") [60], the image detection accuracy has reached a high level, and in the image description task of this study, more attention is paid to the recognition of the spatial position relationship of the target, so this paper uses the spatial attention mechanism.

In this paper, the implementation of the attention mechanism is spatial attention, which calculates weights for each position in an image and then uses these weights to compute a weighted average overall feature. As shown in Figure 6, the attention mechanism visualizes where attention is focused during the recognition process, highlighting the worker and the hook to enhance perception and understanding of the image.

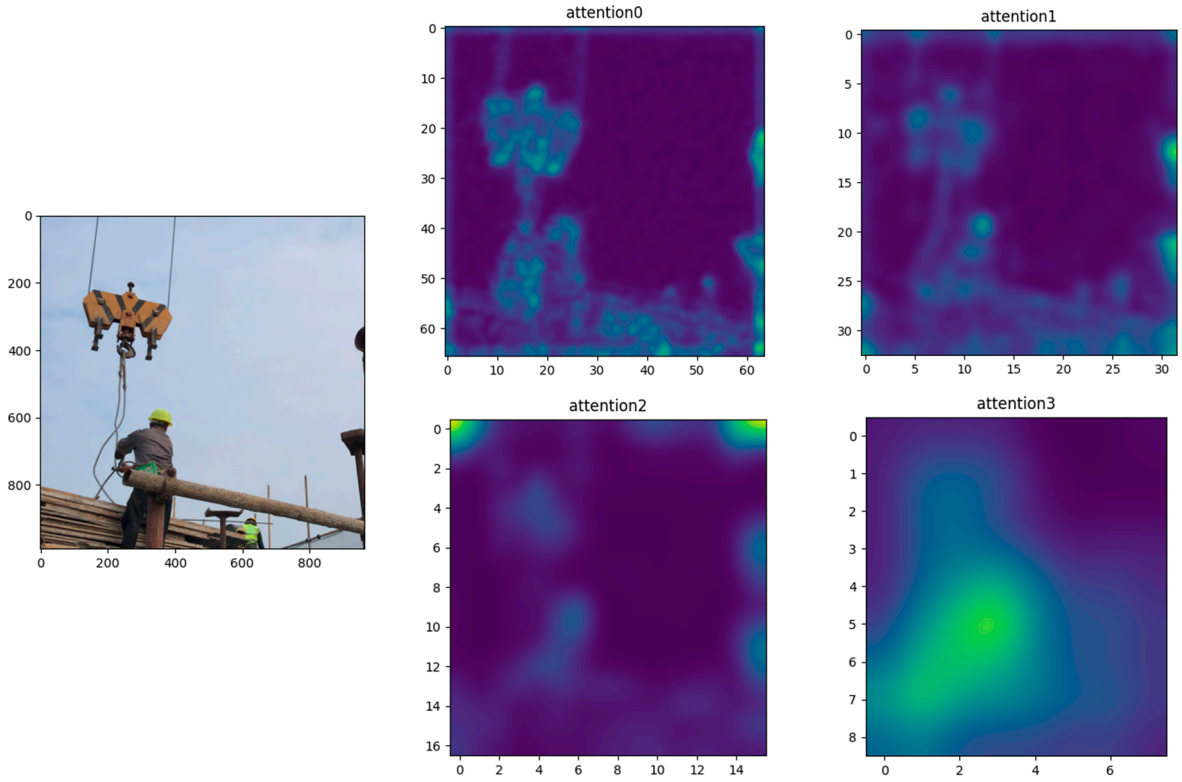


Figure 6. The visualization of RNN attention mechanism (The part that the attention mechanism focuses on appears bright green, while the rest appears dark blue).

3. Beam search

After using a linear layer to convert the output of the decoder into scores for each word in the vocabulary, a greedy algorithm is generally used to select the word with the highest score at each time step and use it to predict the next word. However, the sequence obtained by greedy search is often not the optimal sequence because the first

word affects the prediction of the following words. As shown in Figure 7a,b, the output sequence obtained by the greedy algorithm may not be the best sequence. Exhaustive search enumerates all possible output sequences and their conditional probabilities before selecting the best one, but this is computationally infeasible. Therefore, this paper adopts a beam search strategy [62].

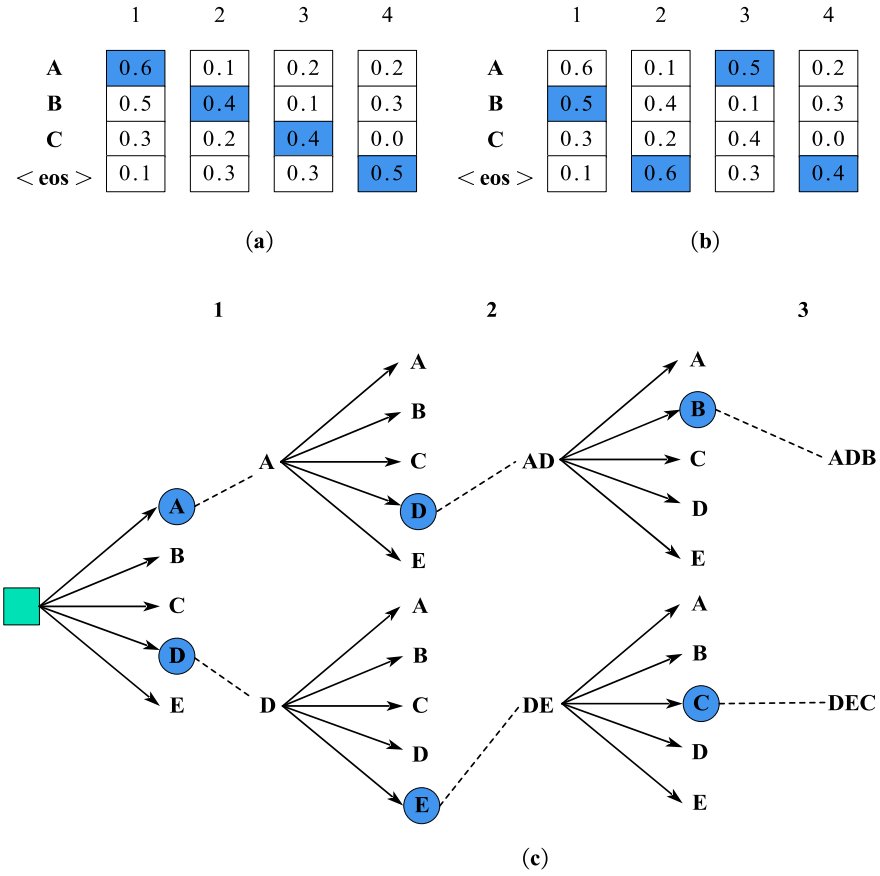


Figure 7. Beam search algorithm. (**a**) Score per word of greedy algorithm, (**b**) Theoretical best score for each word, (**c**) Simplified process example of beam search.

Beam search adds a beam width parameter k to greedy search. At time step 1, the k tokens with the highest probabilities are selected. In the following time steps, k candidate output sequences are selected based on the possible choices of the previous k candidates with the highest conditional probabilities. After k sequences terminate, the sequence with the highest product of conditional probabilities is selected as the output sequence.

$$\frac{1}{L^a} logP|(y_1, \cdots, y_L|c) = \frac{1}{L^\alpha} \sum_{t'=1}^{L} logP(y_{t\prime}|y_1, \cdots, y_{t'-1}, c) \tag{7}$$

where L is the length of the final candidate sequence and α is set to 0.75. In beam search, the smaller the beam width, the fewer exhaustive searches and the faster the prediction, but the lower the accuracy. Conversely, the larger the beam width, the more exhaustive searches and the slower the prediction, but the higher the accuracy. When k = 1, it is equivalent to greedy search, and when k = n, it is equivalent to exhaustive search. Beam search can find a balance between accuracy and computational cost by flexibly choosing the beam width k. Figure 7c shows a simplified process of beam search when k = 2 and the output sequence length is 3.

Beam search is widely used in natural language processing tasks such as machine translation, text generation, and speech recognition. By retaining multiple candidate

sequences, beam search can explore as much output space as possible while ensuring certain computing resources, thereby obtaining better translation or generation results. For different projects and different amounts of training data, beam search makes it easier for project personnel to balance computing resources and output result accuracy, which further reflects the high flexibility of this research method.

3.2. Development of Visual-Language Dataset

This paper creates a special data set using a common JSON format for the visual language model of the construction process related to tower crane hoisting.

1. Image collection

Based on this research, a portion of the construction image data is selected from the SODA dataset [60]. Additionally, some videos are crawled from the internet, and frame extraction is used to obtain some image data. In total, nearly 20,000 on-site construction images are obtained and screened. The screened images must meet at least the following requirements: (1) contain one or more construction activities; (2) have sufficient clarity; (3) have clear image semantics. Through screening and processing, 14,000 construction image data are obtained.

2. Image annotation

Due to the inherent complexity of object relationships in construction sites, the visual language model often requires personnel with a certain level of engineering-related knowledge to annotate images based on specific grammatical rules. To this end, we employed a crowdsourcing approach, reaching out to 35 professionals in the field of construction for the visual language model. Crowdsourcing is a multi-disciplinary collaborative approach that shifts problem-solving from the traditional "individual" approach to a "distributed, diverse, and collaborative" approach, whereby tasks previously performed by specific organizations or individuals are outsourced to large groups of people in an open manner [63].

Before launching the crowdsourcing initiative, a webpage and vocabulary library were created for online image annotation. As shown in Figure 8, annotators can log in using individual account passwords and click buttons to provide annotations for each image. They can replace or delete incorrect annotations by clicking "reset". The image annotation process is shown in Figure 9. As shown in Figure 10, annotators are required to annotate according to the basic linguistic paradigm of quantifier + subject–predicate–object + attribute + adverb.

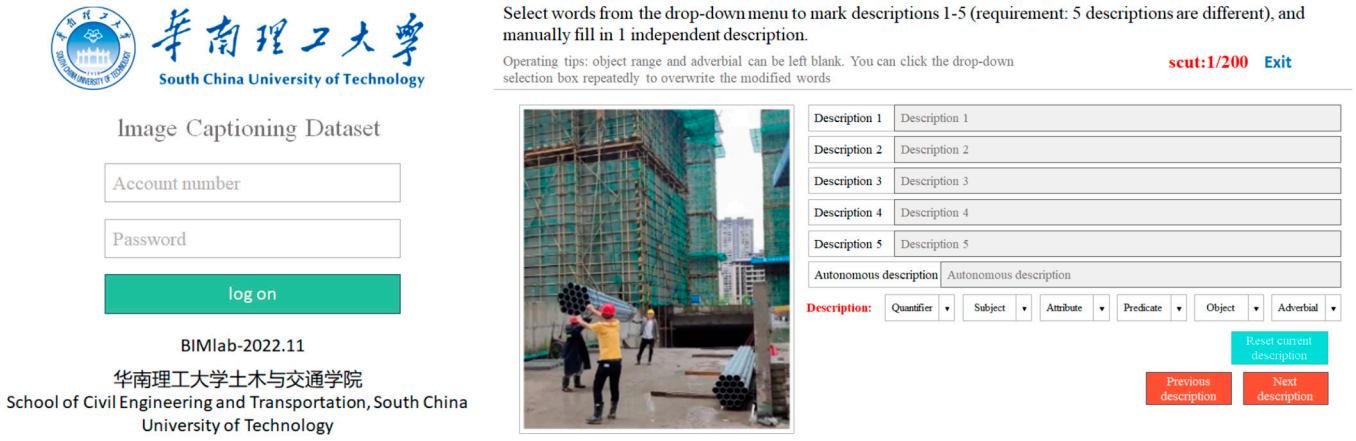


Figure 8. Annotation webpage.

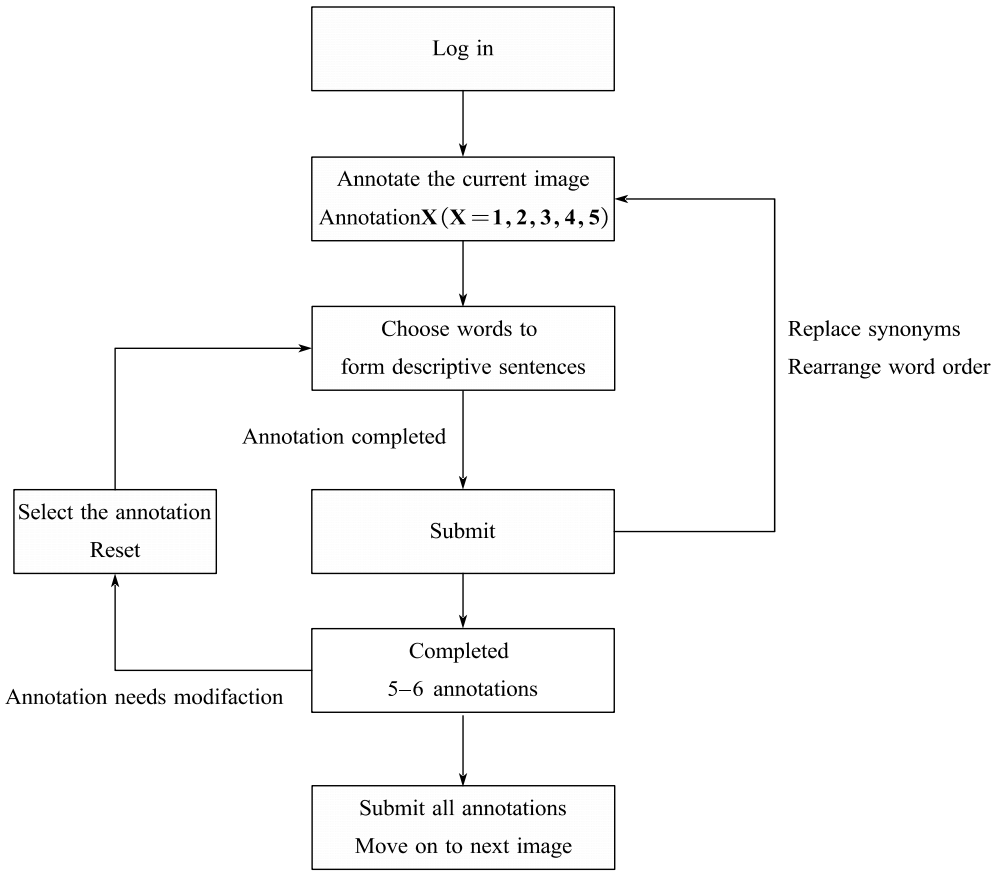


Figure 9. Flowchart of image annotation process.

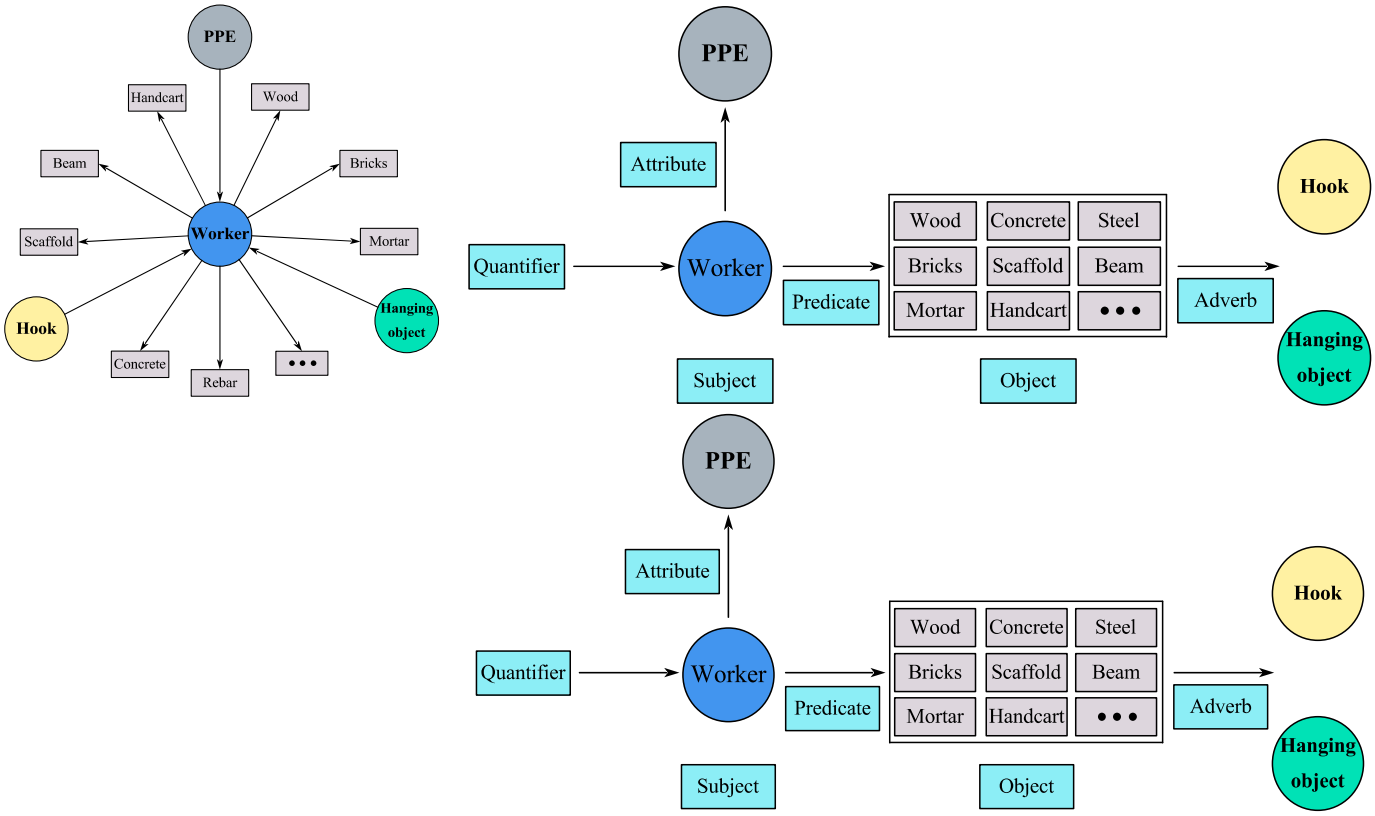


Figure 10. Paradigms of basic languages.

Creating the vocabulary library and the basic linguistic paradigm not only facilitates application to crowdsourcing tasks but also offers many advantages for subsequent research. (1) It ensures that all crowdsourcing annotators use the same terminology to annotate the image content, avoiding confusing engineering terminology, and also considers synonym substitution for some words, thus improving the accuracy and diversity of semantic descriptions. (2) It enables crowdsourced annotators to write descriptions that conform to the specifications and guide crowdsourced annotators to describe around core workers, work scenarios, and operational processes, thus ensuring readability and usability of the dataset and avoiding ambiguity in semantic descriptions due to different individual concerns. (3) The error rate and incompleteness in the dataset are reduced, so that the constructed visual-language data contain complete relationship and attribute annotations of common objects centered on construction workers, thus effectively reducing the cost and workload of subsequent data processing.

In addition, we also allow the crowdsourced annotators to make their own annotation for the sixth sentence, and after manual review by the background, their high-quality annotations can be recorded in the dataset. Some examples of the datasets are illustrated in Figure 11. After the annotation is completed, the annotation is stored in JSON format. The data are later manually checked to remove invalid data and supplement valid data.

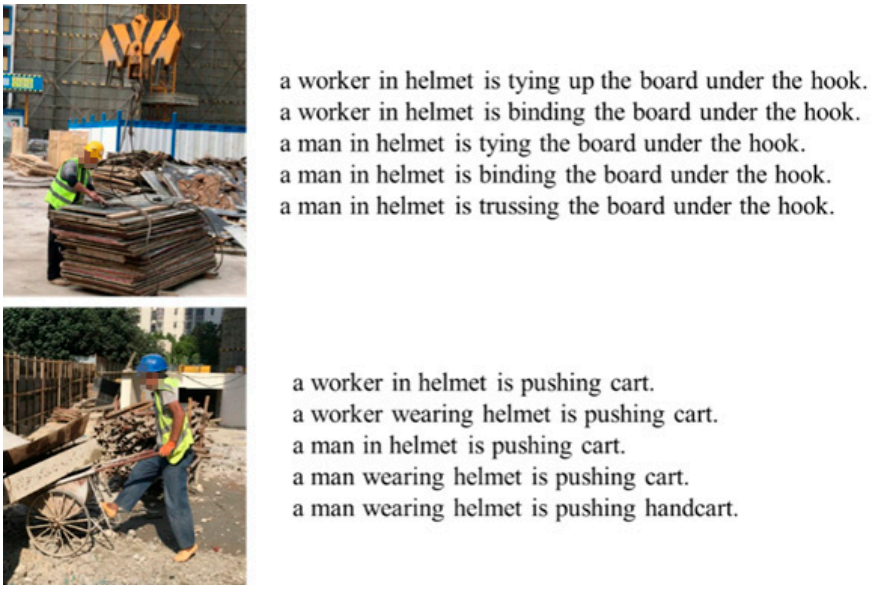


Figure 11. Some examples of the datasets.

After determining the crowdsourcing annotation tasks, we initiated the webpage crowdsourcing annotation process with 35 undergraduate students majoring in civil engineering and engineering management. Upon completion of the annotation, the data were stored in JSON format. To ensure the quality of the annotation, we employed a cross-checking method between annotators and conducted additional verification by 4 experts. The data are checked to remove invalid data and supplement valid data. Finally, the dataset of the visual language model is randomly divided into 80% (11,200 images) for the training set, 10% (1400 images) for the test set, and 10% (1400 images) for the validation set.

3. Preprocessing for datasets

Before training, text needs to be preprocessed. First, remove the tag symbols, convert all uppercase letters to lowercase, calculate the frequency of each word, and generate the wordmap as shown in Figure 12. The wordmap also includes the words “start”, “end”, and “pad” that are not present in the captions. This is because before predicting and generating

the first word, the zeroth word <start> is needed to initiate the prediction process. At the same time, the decoder also needs the end word <end> to mark the end of decoding. In addition, to pass the caption sentences as fixed-size tensors, <pad> is needed to expand captions of different lengths into the same length. Each word is assigned a unique index in the wordmap, and vectors for each word can be generated based on the index. Figure 12 shows the process of padding sentence vectors of different lengths. Each word is given an independent index in s, and the vector of each word can be generated by the index.

{"three": 1, "workers": 2, "wearing": 3, "helmets": 4, "are": 5, "lifting": 6, "the": 7, "hanging": 8, "scaffold": 9, "in": 10, "hoisting": 11, "with": 12, "construction": 13, "four": 14, "casting": 15, "concrete": 16, "pouring": 17, "binding": 18, "rebar": 19, "on": 20, "steel": 21, "tying": 22, "up": 23, "two": 24, "without": 25, "observing": 26, "object": 27, "away": 28, "from": 29, "hook": 30, "load": 31, "five": 32, "talking": 33, "a": 34, "is": 35, "moving": 36, "tower": 37, "crane": 38, "worker": 39, "helmet": 40, "slab": 41, "one": 42, "carrying": 43, "shoveling": 44, "sand": 45, "installing": 46, "bending": 47, "to": 48, "work": 49, "next": 50, "many": 51, "plastering": 52, "group": 53, "of": 54, "mortar": 55, "lot": 56, "beam": 57, "working": 58, "board": 59, "column": 60, "under": 61, "drilling": 62, "wall": 63, "": 64, "cement": 65, "operating": 66, "machine": 67, "twisting": 68, "wire": 69, "squatting": 70, "lots": 71, "commanding": 72, "pulling": 73, "bricks": 74, "climbing": 75, "lying": 76, "wood": 77, "timber": 78, "member": 79, "hopper": 80, "measuring": 81, "walking": 82, "far": 83, "pushing": 84, "standing": 85, "cable": 86, "formwork": 87, "ladder": 88, "stepladder": 89, "welding": 90, "washing": 91, "cart": 92, "leveling": 93, "cutting": 94, "sitting": 95, "rest": 96, "beating": 97, "handcart": 98, "building": 99, "floor": 100, "ground": 101, "sweeping": 102, "elevator": 103, "container": 104, "hooper": 105, "pipe": 106, "crash": 107, "resting": 108, "tools": 109, "painting": 110, "heating": 111, "stand": 112, "by": 113, "using": 114, "bucket": 115, "wear": 116, "<unk>": 117, "<start>": 118, "<end>": 119, "<pad>": 0}

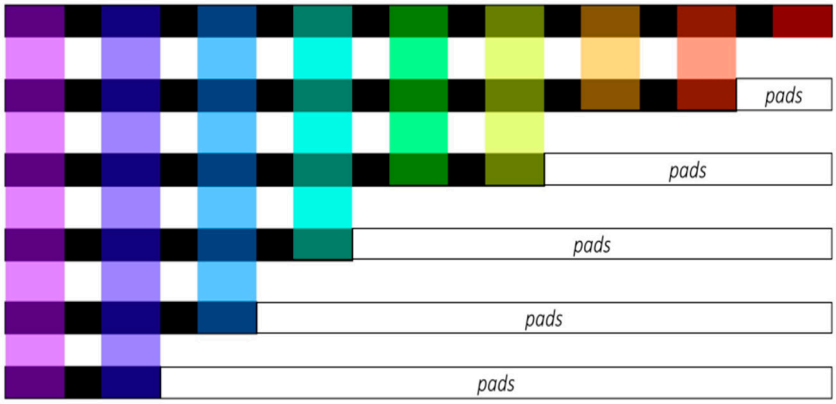


Figure 12. The process of padding and wordmap of SODA-ktsh.

In the end, the collected number of images and captions are as follows: 56,000 captions for 11,200 training images, 7000 captions for 1400 testing images, and 7000 captions for 1400 validation images. The distribution of construction scenes and the number of images and captions in the lifting site are shown in Table 1 and Figure 13. The constructed SODA-ktsh dataset is rich in data volume and covers a variety of construction tasks, providing comprehensive coverage of lifting construction scenes. Table 2 compares the SODA-ktsh dataset with existing visual-language datasets in the construction industry. The number of images, captions, and covered scenes in the SODA-ktsh dataset are currently leading among visual-language datasets in the construction industry.

Table 1. Distribution of SODA-ktsh construction scene quantities.

Scenes	Image Count	Caption Count
Reinforcement work	380	1900
Scaffold work	785	3925
Concrete casting work	359	1795
Formwork	399	1995
Handcart work	129	645
Machine work	184	920
Bricklaying and plastering work	399	1995
Ladder work	77	385
Excavation and earthmoving work	279	1395
Leveling ground work	459	2295
Transport work	265	1325
Mechanical hoisting work	338	1690
Personnel hoisting work	2193	10,965
Commanding work	167	835
Surveying work	109	545
Stand, rest, walk	478	2390

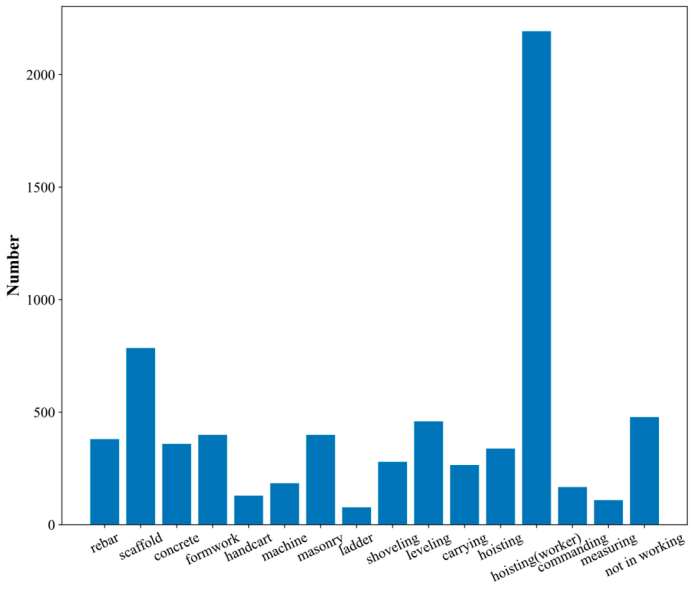


Figure 13. Distribution of SODA-ktsh construction scene quantities.

Table 2. Comparison of SODA-ktsh with existing datasets.

Dataset	Image Count	Caption Count	Scenes	Date
SODA-ktsh	14,000	70,000	16	2023
Huan [13]	7382	36,910	5	2020
Bang [14]	1431	8601	1	2020
ACID-C [15]	6000	18,000	1	2022

4. Experiments and Results

4.1. Model Training

The experiment was conducted on a computer with the following configuration: HP Computer, NVIDIA Geforce RTX 3080Ti, Intel (R) Core (TM) i710750H CPU @ 2.60 GHz 2.60 GHz, 64 GB RAM, from Guangzhou, China. All algorithms are built using the Py-Torch framework. The Adam optimizer is used, with a training epoch of 120 times and a batch size of 32.

The objective of the visual language model is to generate a semantic caption. At each time step, a predicted word is generated, which can be viewed as a word classification process. The word with the highest classification probability is selected as the predicted output. Therefore, the loss function is set as the following cross-entropy loss function:

$$Loss(\theta) = -\sum_{t=1}^{T} log\left(p_\theta\left(y_t^* \middle| y_{1:t-1}^*\right)\right) \quad (8)$$

where $y_{1:T}^*$ represents the ground truth word sequence and θ represents the model training parameters.

The initial learning rates for the encoder and decoder are 1×10^{-3} and 4×10^{-3}, respectively, while the hidden layer and word embedding dimensions for the LSTM are set to 64. In "Show, Attend and Tell" [19], it is suggested that the correlation between the loss and BLEU (Bilingual Evaluation Understudy) score will disappear as the model is trained to a certain extent, and it is recommended to stop training before the BLEU score begins to decrease. To mitigate overfitting, an early stopping mechanism was introduced during training, monitoring the BLEU score on the validation set. Training is terminated if the BLEU score does not improve for ten consecutive epochs. This approach not only prevents overfitting but also provides a degree of flexibility, avoiding premature termination due

to short-term fluctuations. The curve of the loss parameter is shown in Figure 14, and the loss parameter continuously decreases and stabilizes as the training iterations increase. It reaches a small state at around 120 rounds, and the subsequent rounds show a little decrease in loss.

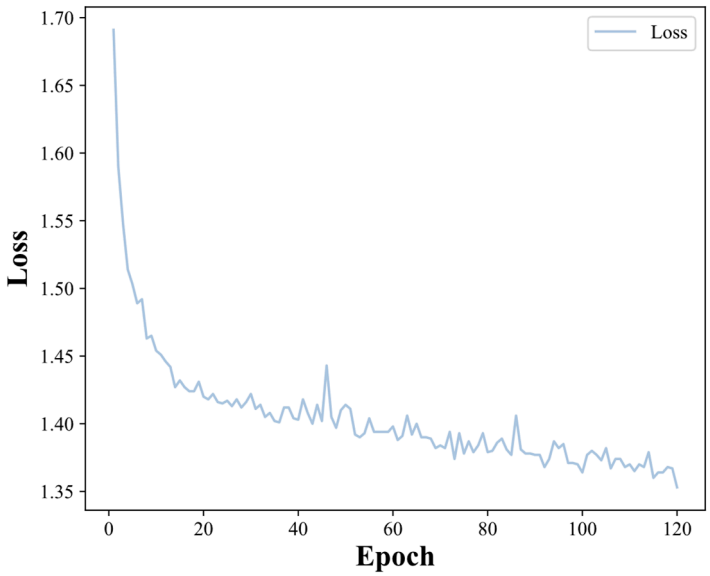


Figure 14. The total loss curve on the SODA-ktsh training set.

4.2. Evaluation Metrics

BLEU (Bilingual Evaluation Understudy) is an evaluation metric proposed by Papineni for measuring the accuracy of machine translation [64]. BLEU calculates the difference between the sentences generated by a model and the actual sentences annotated by humans, using the relatedness of n-grams (n-tuple models) to match them on a scale from 0.0 to 1.0. BLEU has the advantages of being computationally light, easy to understand, free of linguistic differences, and highly correlated with human evaluation results. It is divided into multiple evaluation metrics based on n-grams, such as BLEU-1, BLEU-2, etc. Higher-order BLEU can measure the fluency of sentences. The higher the evaluation metric score, the closer the generated language caption is to the human-annotated caption, indicating a higher quality of language caption.

$$CP_n(C,R) = \frac{\sum_i \sum_k \min(h_k(c_i), \max_{j \in m} h_k(r_{i,j}))}{\sum_i \sum_k h_k(c_i)} \tag{9}$$

where $h_k(c_i)$ represents the frequency of the possible n-gram in the generated sentence, and $h_k(r_{i,j})$ represents the frequency of the possible n-gram in the human-annotated sentence. In the second step, a penalty factor is added for sentence length to penalize the generation of shorter sentences, so that the constructed $CP_n(C,S)$ is not biased towards generating shorter sentences. The penalty factor for balancing sentence length is calculated as follows:

$$b(C,R) = \begin{cases} 1, & if\ lc > lr \\ e^{\frac{1-lr}{lc}}, & if\ lc < lr \end{cases} \tag{10}$$

The calculation of the final score in the third step is as follows:

$$BLEU_N(C,R) = b(C,R) \exp\left(\sum_{n=1}^{N} \omega_n \log CP_n(C,R)\right) \tag{11}$$

where the values of n-grams range from 1 to 4, and $\omega_n = \frac{1}{N}$.

Top-5 Accuracy is one of the metrics used to evaluate the performance of a visual language model. It measures the number of the top 5 predictions with the highest probabilities output by the model that match the actual annotation for a given image. If at least one prediction matches the annotation, the Top-5 Accuracy is 1; otherwise, it is 0. Top-5 Accuracy can provide a more flexible evaluation of the model's performance, avoiding small prediction errors being considered as incorrect predictions. It is less affected by factors such as noise, blurriness, and ambiguity, as the model has more opportunities to make choices. In actual training, Top-5 Accuracy can help us understand the diversity and robustness of the model's output. A high Top-5 Accuracy indicates that the model can generate multiple correct captions, while a low Top-5 Accuracy indicates that the model may have overfitting or underfitting issues.

During the evaluation phase, we introduced another two metrics, CIDEr [65] and ROUGE_L [66], to assess the semantic quality of the generated caption. CIDEr and ROUGE_L are also widely used metrics for visual language model evaluation, with higher scores indicating better quality of the generated semantic descriptions.

Table 3 shows the results of various metrics obtained during the training process with the main generations of iterations. Figure 15 shows the curves of BLEU-4 and Top-5 Accuracy during the training process. It can be seen that both BLEU-4 and Top-5 Accuracy continuously increase to a stable state as the LOSS decreases during the training process. This demonstrates good model training fitting.

Table 3. Training parameters and evaluation metrics on the SODA-ktsh training set.

Epoch	Batch Time	LOSS	Top-5 Accuracy	BLEU-4
10	0.308	1.523	97.277%	0.639
20	0.308	1.490	97.769%	0.656
30	0.310	1.487	97.753%	0.657
40	0.308	1.482	97.775%	0.660
50	0.277	1.468	97.957%	0.670
60	0.309	1.455	98.010%	0.682
70	0.306	1.452	98.190%	0.695
80	0.301	1.438	98.305%	0.702
90	0.298	1.439	98.233%	0.702
100	0.302	1.430	98.260%	0.709
110	0.309	1.438	98.365%	0.698
120	0.311	1.418	98.531%	0.720

One major difference between this study and previous research based on encoder–decoder architecture is the incorporation of an attention mechanism and beam search strategy in the prediction process. Existing research [19] demonstrated that the attention mechanism enables the model to focus on salient features, suppress weak features, discard redundant and noisy features, enhance the perception and understanding of construction scenes, and reduce the loss of key information. To validate the effectiveness of beam search, this paper conducted ablation experiments to explore the impact of beam search on model performance. As shown in Table 4, the experimental results indicate that the model with beam search outperforms the model without beam search in 6 evaluation metrics of BLEU-1, BLEU-2, BLEU-3, BLEU-4, CIDEr, and ROUGE_L. Moreover, as the beam width parameter k increases, these four metrics show an upward trend, which fully demonstrates the effectiveness of the beam search strategy in encoder–decoder models in the visual language model.

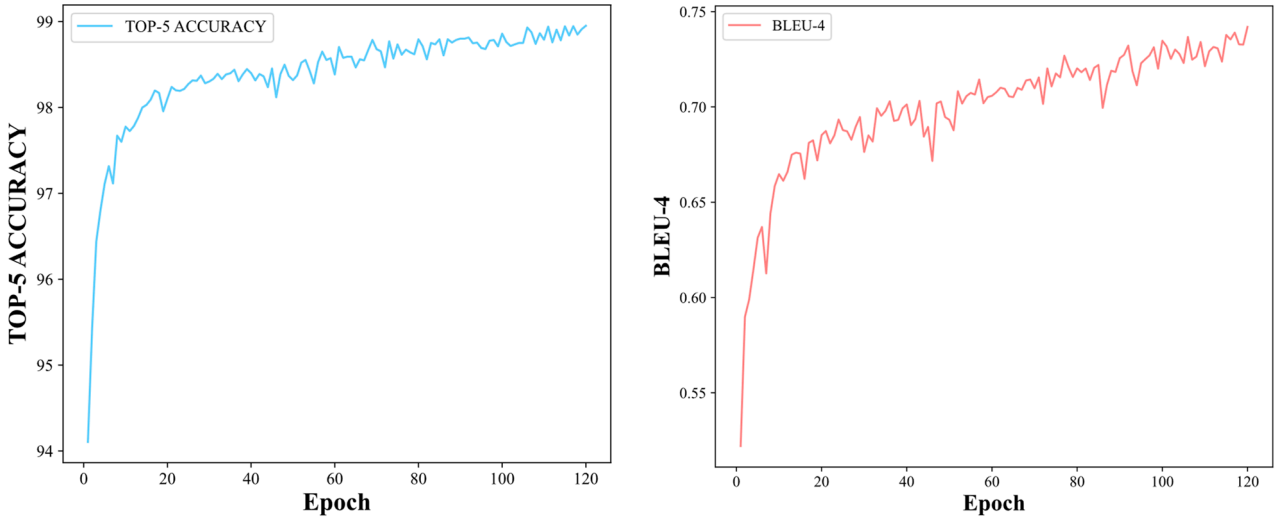


Figure 15. Training accuracy and BLEU-4 curve of SODA-ktsh validation set.

Table 4. Results of ablation experiment on the SODA-ktsh validation set.

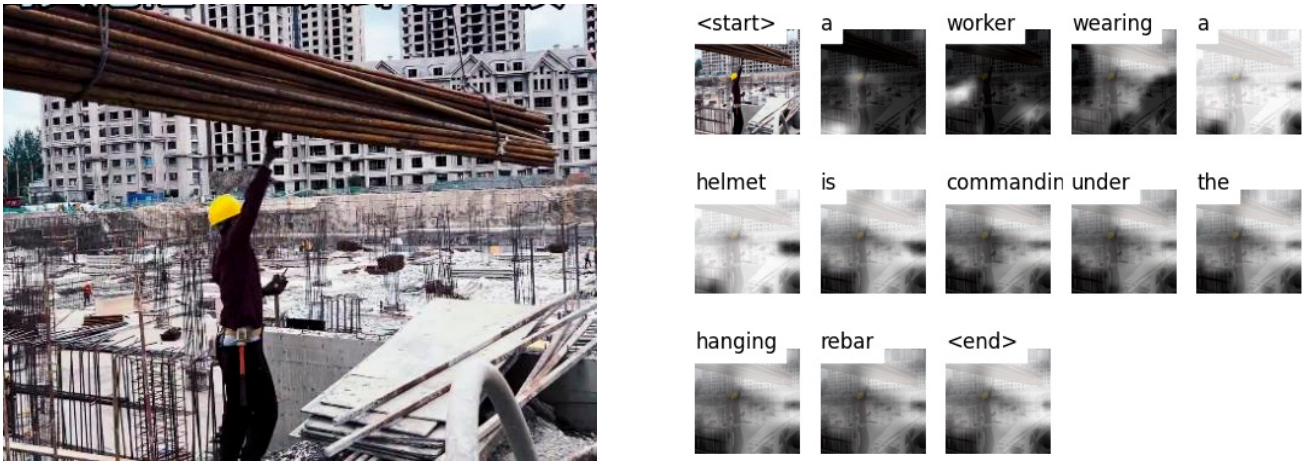
Process	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr	ROUGE_L
beam size = 1	0.8087	0.7788	0.7569	0.7365	4.9398	0.8075
beam size = 2	0.8117	0.7817	0.7593	0.7383	4.9398	0.8075
beam size = 3	0.8155	0.7865	0.765	0.7448	4.9719	0.8093
beam size = 4	0.8156	0.7869	0.7654	0.7452	4.9908	0.8101
beam size = 5	0.8161	0.7872	0.7661	0.7464	5.0255	0.8106

4.3. Analysis of Results

In order to visually demonstrate the performance of the model, this paper also visualizes the caption generation results of the validation set as shown in Figure 16a–d. The detection results show that the model can accurately describe the complex semantics of construction images and has satisfactory generalization performance and perception and recognition ability for information in complex construction scenes. Moreover, it can accurately distinguish whether workers are wearing safety helmets and whether they are working under hanging hooks, as shown in the detection results in Figure 16c,d. This indicates that the model can also recognize relevant attributes such as safety, quantity, and spatial relationships accurately. In addition, the generated semantic captions strictly conform to the linguistic paradigm of quantifier + subject–predicate–object + adjective + adverb, which provides great convenience for subsequent research on the format of input for determining the degree of danger. Furthermore, to further test the model's generalization ability, real construction images outside of the dataset were used for detection, as shown in Figure 16e,f.

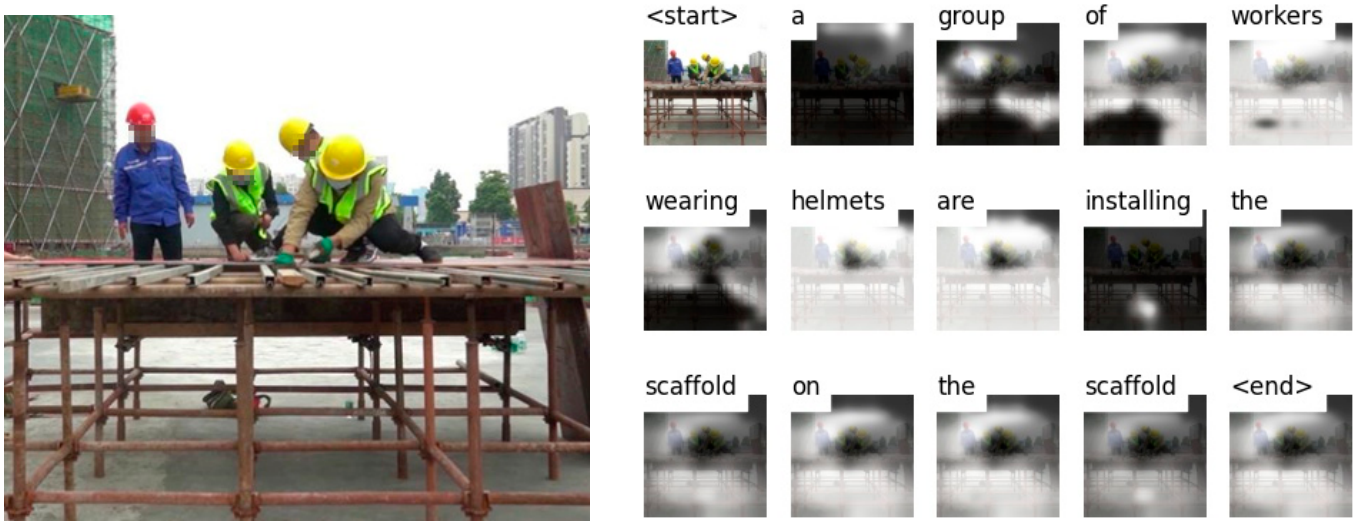
The results demonstrate that our model can still generate accurate captions that align with image engineering conventions beyond the training dataset. However, we have identified certain failure cases, particularly in recognizing helmet usage among workers. Specifically, the model occasionally misclassifies a worker as not wearing a helmet or fails to determine helmet usage for individuals within a group of workers. For instance, in Figure 16e, our model incorrectly identified a worker as not wearing a helmet. Upon analysis, we attribute this misclassification to the influence of image perspective and lighting conditions. Our dataset contains relatively few images captured under low-light conditions or from low angles, which may limit the model's ability to effectively handle such scenarios. In Figure 16f, some workers were not wearing helmets, yet the model erroneously described them as having helmets. This issue likely arises from our annotation

process, which prioritizes overall scene understanding rather than explicitly labeling helmet presence for each individual worker.



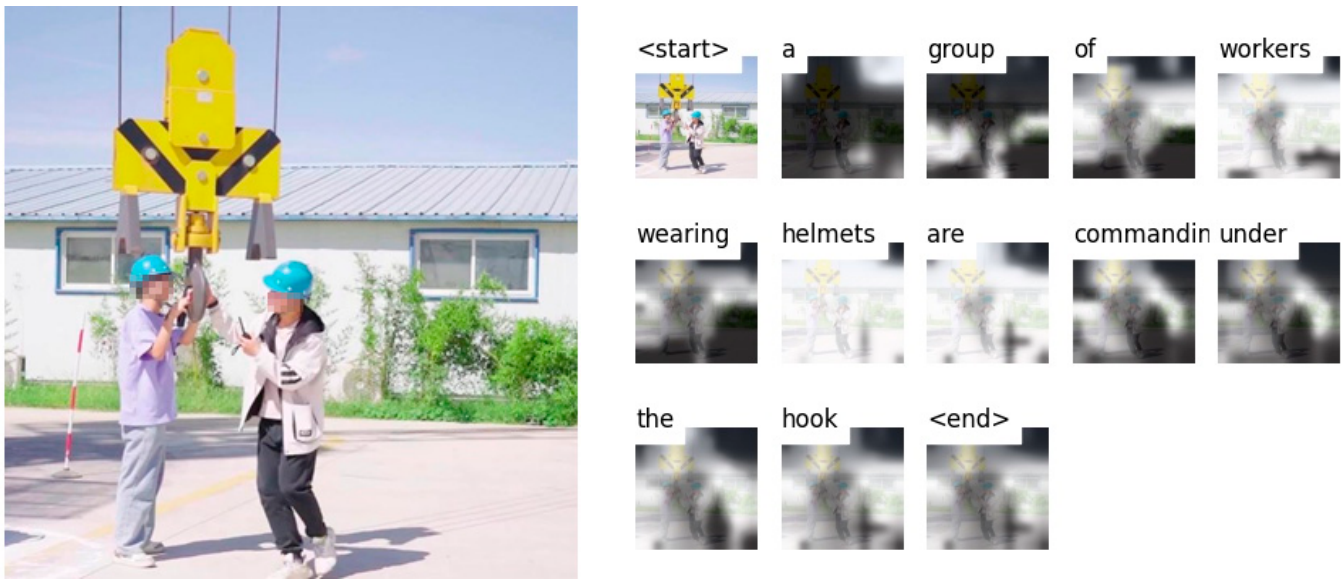
(**a**) Visual display of hoisting

(**b**) Visual display of bricklaying

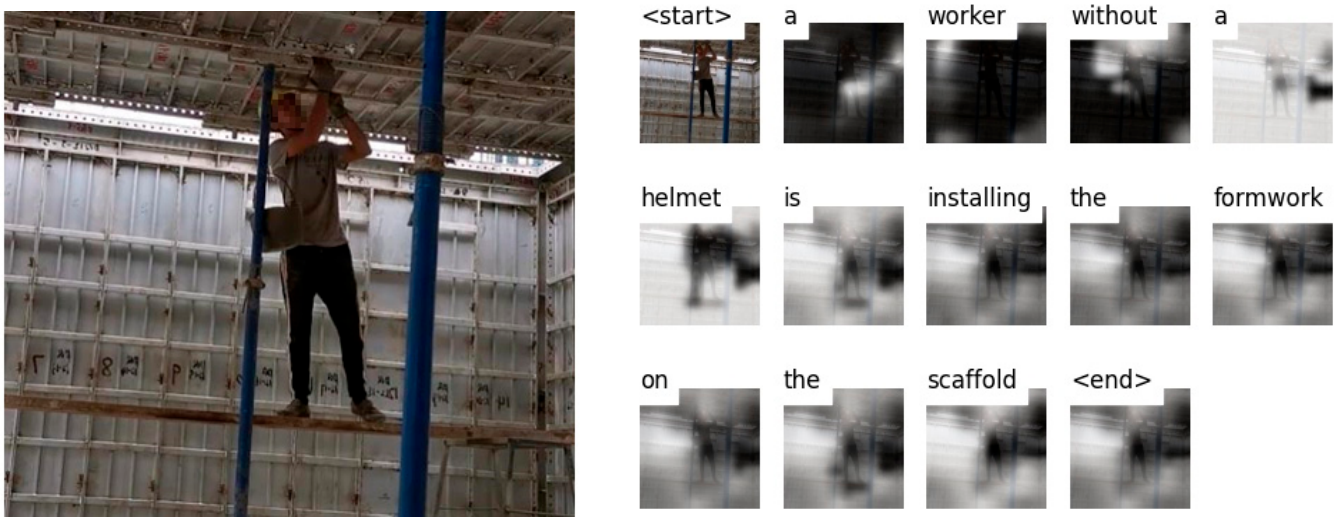


(**c**) Visual display of scaffold installation

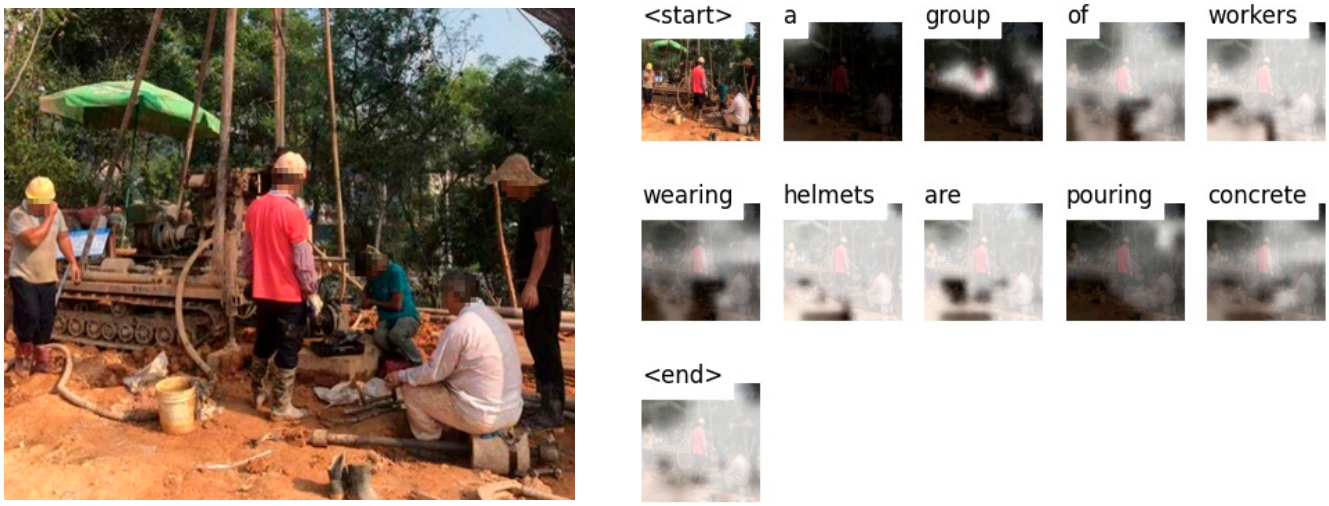
Figure 16. *Cont*.



(**d**) Visual display of command



(**e**) Visual display of formwork installation



(**f**) Visual display of concrete pouring

Figure 16. Some caption generation results of the SODA-ktsh validation set.

5. Discussion

5.1. Contributions to the Body of Knowledge

This research makes a noteworthy contribution to the scholarly literature on construction scene comprehension tasks. Prior research has primarily focused on unit element recognition from videos or images, with limited exploration of deep semantic understanding of the entire construction scene, including objects, relationships, and attributes. This research offers innovation in model design and dataset construction by proposing an automatic visual language model that treats construction scenes holistically with visual information and natural language sentences.

Specifically, an integrated ResNet101 "encoder" and LSTM + attention + beam search "decoder" model is proposed for recognizing semantic information from images. The integration of the attention mechanism enables the model to precisely focus on critical regions within an image, effectively filtering out essential visual elements for the task while mitigating the interference of information overload. This enhances the model's ability to concentrate and improves recognition accuracy. Additionally, the beam search strategy further optimizes the result generation process by selecting from multiple candidate sequences, ensuring that the generated descriptions are not only accurate but also more contextually relevant. The experimental results demonstrate that the proposed model outperforms previous approaches in terms of BLEU-4 and Top-5 Accuracy. Additionally, the ablation experiment showed that incorporating beam search into encoder–decoder architecture in the field of the visual language model can improve six evaluation metrics: BLEU-1, BLEU-2, BLEU-3, BLEU-4, CIDEr, and ROUGE_L. Moreover, with an increase in beam width, these metrics demonstrated a gradually improving trend, which fully validates the effectiveness of the beam search in encoder–decoder architecture for the visual language model. Based on the detection results, the model exhibits good generalization performance and perception and recognition ability in complex construction scenes. It accurately identifies the usage of safety helmets, workers working under hanging hooks, and other relevant attributes such as safety, quantity, and spatial relationships, thereby enhancing its practicality. The generated semantic captions adhere to a specific linguistic paradigm.

Furthermore, this research leverages crowdsourcing techniques to enhance the SODA-ktsh dataset through image annotation. We developed a dedicated online platform and vocabulary library that facilitated workers in annotating images based on a specific linguistic paradigm. This approach ensured the consistency of terminology usage and improved the accuracy, diversity, readability, and practicality of semantic descriptions in the dataset. By combining crowdsourced annotations with manual review, we effectively minimized errors and incompleteness, resulting in comprehensive relationships and attribute annotations centered around construction workers. The application of crowdsourcing significantly reduced the cost and workload of subsequent data processing. Following the removal of invalid data and supplementation with valid data, the final annotations were stored in JSON format. The SODA-ktsh dataset encompasses 16,000 images, 70,000 captions, and covers 16 types of construction scenes, surpassing existing datasets in the construction industry in terms of image and caption quantity, as well as scene-type coverage.

5.2. Practical Implications and Application Challenges

The practical implications and application challenges of the proposed visual language model are discussed in this section.

(1) Practical implications: The proposed approach is able to depict an activity scene of a construction site using full sentences. The model utilizes an attention mechanism to highlight significant elements within the scene, which can assist managers in quickly scanning the generated language information to obtain reliable information regarding

the construction activity status. For instance, by analyzing the frequency of scene words like "helmet" and spatial positional relationships such as "under" and "next to", safety managers can evaluate the safety status of workers.

(2) Application challenges: The proposed approach is confronted with a challenge in its application, as its level of accuracy is dependent on the size and quality of the training dataset, as well as the effectiveness of the model architecture [67]. Given that visual language models are still evolving and there are limitations in generating large datasets in the construction industry, achieving higher levels of accuracy may prove to be difficult.

5.3. Limitations and Further Work

While the proposed model in this article shows great promise, it is important to acknowledge its limitations in practical applications. One such limitation is that training the model requires significant time and labor due to the substantial dataset requirements. Addressing this challenge requires incorporating autonomous learning techniques that combine automated model learning with human sampling, which has the potential to dramatically reduce the amount of labeled training data needed and increase the model's learning efficiency.

Moreover, the evaluation of personnel safety conditions, which remains an important task, still requires expert assessment by safety managers. To streamline this process and reduce human involvement, it is necessary to explore how the model can directly judge personnel safety conditions. This could further enhance the model's efficiency and effectiveness while minimizing the usage of personnel.

Furthermore, to enhance the model's generalization capability, the diversity of the dataset needs to be improved. Specifically, it is essential to incorporate images captured under varying conditions, such as different lighting, occlusions, and camera angles. This would help mitigate issues observed in Figure 16e, where the model failed to recognize the absence of a safety helmet due to lighting and camera angle variations. Additionally, the precision of data annotation warrants further attention. The current dataset lacks detailed annotations for individual safety helmets, preventing the model from accurately classifying workers wearing or not wearing helmets, as seen in Figure 16f. Given the critical role of safety helmet recognition in construction site safety, refining data annotation should be a priority in future work. A potential approach could involve integrating object detection with image captioning techniques to improve the model's ability to identify individual objects within an image.

Finally, the proposed model is designed for detailed analysis of static images. While it is not specifically developed for real-time analysis, its offline capabilities provide a solid foundation for building more efficient real-time systems. The model's high accuracy and detailed performance in static image analysis can serve as the basis for constructing a robust framework, which can be further optimized and adapted for real-time applications. Given the significance of real-time analysis in both industry and academia, future work will explore integrating this framework with real-time processing techniques, such as lightweight model architectures, hardware acceleration, or edge computing. These enhancements aim to improve computational efficiency and response time while maintaining model accuracy, making it more suitable for practical applications such as safety monitoring and hazard prevention.

6. Conclusions

The reliance on visual information in construction management is steadily increasing; however, the vast amount of image data has not been effectively leveraged to support

decision-making processes. The primary challenge lies in the limitations of traditional image processing methods, which predominantly focus on extracting and analyzing low-level visual features (e.g., edges and textures) while lacking the capability for high-level semantic interpretation. This results in a semantic gap between machine vision systems and human cognitive frameworks. Although recent advancements in scene understanding algorithms have led to significant theoretical breakthroughs, their practical implementation remains constrained by the absence of systematically constructed, industry-grade annotated datasets. In particular, the lack of open-source databases with multi-attribute annotations and coverage of complex scenes has severely impeded the engineering application and iterative optimization of these algorithms.

To address the problem, this research presents an attention-based encoder–decoder visual language model and introduces the SODA-ktsh visual-language dataset for training and validating the model's performance. The model utilizes an attention mechanism and beam search strategy to improve its parameters, leading to higher accuracy and generalization in generating semantically relevant captions. The experimental results demonstrate that our model achieves a BLEU-4 score of 0.7464, a CIDEr score of 5.0255, and a ROUGE_L score of 0.8106 on the validation set. The results demonstrate the model's effectiveness in capturing contextual relationships between objects in images and generating captions. The SODA-ktsh dataset was constructed by using crowdsourcing to annotate language information on the original SODA dataset, resulting in a construction site caption dataset. We developed a dedicated online platform and vocabulary to enable professionals to annotate images based on a predefined linguistic paradigm. By integrating crowdsourced annotations with manual review, we effectively minimized errors and inconsistencies, resulting in a comprehensive, worker-centered annotation framework that captures relational and attribute information with high accuracy.

Future research should focus on reducing the reliance on large-scale annotated data by developing a self-supervised collaborative learning mechanism that integrates automated model learning with expert-guided sampling to enhance efficiency. Additionally, establishing an automated safety assessment framework based on multimodal perception, integrating domain knowledge with real-time inference, can progressively replace manual expert review. To further improve real-time applicability, heterogeneous data augmentation strategies should be implemented, incorporating extreme lighting conditions, complex occlusions, and multi-view characteristics, while refining annotation standards and leveraging advanced algorithms to enhance the recognition of personal protective equipment (PPE). Furthermore, exploring lightweight model designs through neural architecture search (NAS), combined with edge computing acceleration and adaptive inference engines, will enable real-time safety monitoring in dynamic environments while maintaining the accuracy of static image analysis.

Author Contributions: H.D.: resources, validation, writing—review and editing, funding acquisition. K.F.: methodology (lead), formal analysis, investigation, writing—original draft. B.Y.: methodology, formal analysis, investigation. H.L.: methodology, formal analysis, investigation. R.D.: methodology, validation, writing—review and editing. Y.D.: conceptualization (lead), methodology, writing—review and editing, funding acquisition, resources, validation. J.-r.L.: writing—review and editing, resources. All authors have read and agreed to the published version of the manuscript.

Funding: The authors would like to acknowledge the support by the National Science Foundation of China (52308314) and the support by Guangdong Basic and Applied Basic Research Foundation (2023A1515030169).

Data Availability Statement: For pictures and annotations of SODA and SODA-ktsh, please visit the following link: https://hkustconnect-my.sharepoint.com/:f:/g/personal/ycdeng_connect_ust_hk/

EiQLht3OhstGnKXrjFXyRZYBIXFjUC43jUUNVBXfM_kkKg?e=jJ2Nhv (accessed on 9 March 2025); the code and data of the study can be reasonably obtained from the corresponding authors.

Acknowledgments: The authors would also like to pay special tribute to students who contribute to the data cleaning and annotation process of SODA at the South China University of Technology. Their names are Yuhang Mo, Fuqiang Shen, Zilin Chen, Yuxuan Li, Deyu Zhong, Pengfei Hu, Weixuan Zheng, Jieheng Zhao, Lei Guo, Wanqi Liao, Lin Li, Jiaxin Li, Xiaopeng Yuan, Lihua Long, Sizhe Zheng, Ruyu Xiang, Jinhui Liang, Xuanqian Huang, Yanting Lin, Jin Xiong, Zhetao Fan, Xinyue Zhang, Chenrun Dong, Yilin Huang, Yi Zou, Rui Li, Jiawei Zeng, Jiexin Yao, Jiaxue Cen, Bingying Wu, Debielige Ming, Yating Wei, Shuijiao Liang, Yongtong Liu, Mingjin Zhou, Zhihong Lin, and Hengyun Zhang.

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

References

1. Ham, Y.; Kamari, M. Automated content-based filtering for enhanced vision-based documentation in construction toward exploiting big visual data from drones. *Autom. Constr.* **2019**, *105*, 102831. [CrossRef]
2. Xiong, R.; Song, Y.; Li, H.; Wang, Y. Onsite video mining for construction hazards identification with visual relationships. *Adv. Eng. Inform.* **2019**, *42*, 100966. [CrossRef]
3. Slaton, T.; Hernandez, C.; Akhavian, R. Construction activity recognition with convolutional recurrent networks. *Autom. Constr.* **2020**, *113*, 103138. [CrossRef]
4. Harichandran, A.; Raphael, B.; Mukherjee, A. Equipment activity recognition and early fault detection in automated construction through a hybrid machine learning framework. *Comput.-Aided Civ. Infrastruct. Eng.* **2023**, *38*, 253–268. [CrossRef]
5. Zhang, C.; Zhao, Y.; Li, T.; Zhang, X.; Adnouni, M. Generic visual data mining-based framework for revealing abnormal operation patterns in building energy systems. *Autom. Constr.* **2021**, *125*, 103624. [CrossRef]
6. Zhong, B.; Shen, L.; Pan, X.; Lei, L. Visual attention framework for identifying semantic information from construction monitoring video. *Saf. Sci.* **2023**, *163*, 106122. [CrossRef]
7. Hu, N.; Fan, C.; Ming, Y.; Feng, F. MAENet: A novel multi-head association attention enhancement network for completing intra-modal interaction in image captioning. *Neurocomputing* **2023**, *519*, 69–81. [CrossRef]
8. Wu, J.; Cai, N.; Chen, W.; Wang, H.; Wang, G. Automatic detection of hardhats worn by construction personnel: A deep learning approach and benchmark dataset. *Autom. Constr.* **2019**, *106*, 102894. [CrossRef]
9. Nath, N.D.; Behzadan, A.H.; Paal, S.G. Deep learning for site safety: Real-time detection of personal protective equipment. *Autom. Constr.* **2020**, *112*, 103085. [CrossRef]
10. Abdullahi, I.; Chukwuma, N.; Mostafa, N.; Amanda, K.; Ulises, T. Investigating the impact of physical fatigue on construction workers' situational awareness. *Saf. Sci.* **2023**, *163*, 106103.
11. Yu, Y.; Li, H.; Yang, X.; Kong, L.; Luo, X.; Wong, A.Y.L. An automatic and non-invasive physical fatigue assessment method for construction workers. *Autom. Constr.* **2019**, *103*, 1–12. [CrossRef]
12. Chen, C.; Zhu, Z.; Hammad, A. Automated excavators activity recognition and productivity analysis from construction site surveillance videos. *Autom. Constr.* **2020**, *110*, 103045. [CrossRef]
13. Zhu, C.; Zhu, J.; Bu, T.; Gao, X. Monitoring and Identification of Road Construction Safety Factors via UAV. *Sensors* **2022**, *22*, 8797. [CrossRef]
14. Bang, S.; Kim, H. Context-based information generation for managing UAV-acquired data using image captioning. *Autom. Constr.* **2020**, *112*, 103116. [CrossRef]
15. Olanrewaju, A.; AbdulAziz, A.; Preece, C.N.; Shobowale, K. Evaluation of measures to prevent the spread of COVID-19 on the construction sites. *Clean. Eng. Technol.* **2021**, *5*, 100277. [CrossRef] [PubMed]
16. Essam, N.; Khodeir, L.; Fathy, F. Approaches for BIM-based multi-objective optimization in construction scheduling. *Ain Shams Eng. J.* **2023**, *14*, 102114. [CrossRef]
17. Golovina, O.; Teizer, J.; Johansen, K.W.; König, M. Towards autonomous cloud-based close call data management for construction equipment safety. *Autom. Constr.* **2021**, *132*, 103962. [CrossRef]
18. Liu, H.; Wang, G.; Huang, T.; He, P.; Skitmore, M.; Luo, X. Manifesting construction activity scenes via image captioning. *Autom. Constr.* **2020**, *119*, 103334. [CrossRef]
19. 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, ICML 2015, Lille, France, 6–11 July 2015; Volume 37, pp. 2048–2057.
20. Dash, S.K.; Acharya, S.; Pakray, P.; Das, R.; Gelbukh, A. Topic-Based Image Caption Generation. *Arab. J. Sci. Eng.* **2020**, *45*, 3025–3034. [CrossRef]

21. Suresh, K.R.; Jarapala, A.; Sudeep, P.V. Image Captioning Encoder–Decoder Models Using CNN-RNN Architectures: A Comparative Study. *Circuits Syst. Signal Process.* **2022**, *41*, 5719–5742. [CrossRef]
22. Alsakka, F.; El-Chami, I.; Yu, H.; Al-Hussein, M. Computer vision-based process time data acquisition for offsite construction. *Autom. Constr.* **2023**, *149*, 104803. [CrossRef]
23. Lin, T.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014, Proceedings, Part V*; Springer: Cham, Switzerland, 2014; pp. 740–755.
24. Wang, C.; Gu, X. Learning Double-Level Relationship Networks for image captioning. *Inf. Process. Manag.* **2023**, *60*, 103288. [CrossRef]
25. Young, P.; Lai, A.; Hodosh, M.; Hockenmaier, J. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans. Assoc. Comput. Linguist.* **2014**, *2*, 67–78. [CrossRef]
26. Li, B.; Cheng, K.; Yu, Z. Histogram of Oriented Gradient Based Gist Feature for Building Recognition. *Comput. Intell. Neurosci.* **2016**, *2016*, 6749325. [CrossRef] [PubMed]
27. Eo, Y.D.; Pyeon, M.W.; Kim, S.W.; Kim, J.R.; Han, D.Y. Coregistration of terrestrial lidar points by adaptive scale-invariant feature transformation with constrained geometry. *Autom. Constr.* **2012**, *25*, 49–58. [CrossRef]
28. Li, J.; Wang, Y.; Wang, Y. Visual tracking and learning using speeded up robust features. *Pattern Recognit. Lett.* **2012**, *33*, 2094–2101. [CrossRef]
29. Zhou, Y.; Guo, H.; Ma, L.; Zhang, Z.; Skitmore, M. Image-based onsite object recognition for automatic crane lifting tasks. *Autom. Constr.* **2021**, *123*, 103527. [CrossRef]
30. Li, Y.; Lu, Y.; Chen, J. A deep learning approach for real-time rebar counting on the construction site based on YOLOv3 detector. *Autom. Constr.* **2021**, *124*, 103602. [CrossRef]
31. Kim, D.; Liu, M.; Lee, S.; Kamat, V.R. Remote proximity monitoring between mobile construction resources using camera-mounted UAVs. *Autom. Constr.* **2019**, *99*, 168–182. [CrossRef]
32. Kardovskyi, Y.; Moon, S. Artificial intelligence quality inspection of steel bars installation by integrating mask R-CNN and stereo vision. *Autom. Constr.* **2021**, *130*, 103850. [CrossRef]
33. Chen, S.; Demachi, K. Towards on-site hazards identification of improper use of personal protective equipment using deep learning-based geometric relationships and hierarchical scene graph. *Autom. Constr.* **2021**, *125*, 103619. [CrossRef]
34. Hwang, J.; Lee, K.; Ei Zan, M.M.; Jang, M.; Shin, D.H. Improved Discriminative Object Localization Algorithm for Safety Management of Indoor Construction. *Sensors* **2023**, *23*, 3870. [CrossRef]
35. Wang, X.; Zhu, Z. Vision-based hand signal recognition in construction: A feasibility study. *Autom. Constr.* **2021**, *125*, 103625. [CrossRef]
36. Kim, K.; Cho, Y.K. Effective inertial sensor quantity and locations on a body for deep learning-based worker's motion recognition. *Autom. Constr.* **2020**, *113*, 103126. [CrossRef]
37. Cheng, M.; Khitam, A.F.K.; Tanto, H.H. Construction worker productivity evaluation using action recognition for foreign labor training and education: A case study of Taiwan. *Autom. Constr.* **2023**, *150*, 104809. [CrossRef]
38. Antwi-Afari, M.F.; Qarout, Y.; Herzallah, R.; Anwer, S.; Umer, W.; Zhang, Y.; Manu, P. Deep learning-based networks for automated recognition and classification of awkward working postures in construction using wearable insole sensor data. *Autom. Constr.* **2022**, *136*, 104181. [CrossRef]
39. Luo, X.; Li, H.; Yang, X.; Yu, Y.; Cao, D. Capturing and Understanding Workers' Activities in Far-Field Surveillance Videos with Deep Action Recognition and Bayesian Nonparametric Learning. *Comput.-Aided Civ. Infrastruct. Eng.* **2019**, *34*, 333–351. [CrossRef]
40. Luo, X.; Li, H.; Cao, D.; Yu, Y.; Yang, X.; Huang, T. Towards efficient and objective work sampling: Recognizing workers' activities in site surveillance videos with two-stream convolutional networks. *Autom. Constr.* **2018**, *94*, 360–370. [CrossRef]
41. Yang, J.; Shi, Z.; Wu, Z. Vision-based action recognition of construction workers using dense trajectories. *Adv. Eng. Inform.* **2016**, *30*, 327–336. [CrossRef]
42. Fang, H.; Gupta, S.; Iandola, F.; Srivastava, R.K.; Deng, L.; Dollár, P.; Gao, J.; He, X.; Mitchell, M.; Platt, J.C.; et al. From captions to visual concepts and back. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2015, Boston, MA, USA, 7–12 June 2015; pp. 1473–1482.
43. Yatskar, M.; Zettlemoyer, L.; Farhadi, A. Situation recognition: Visual semantic role labeling for image understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2016, Las Vegas, NV, USA, 27–30 June 2016; pp. 5534–5542.
44. Ushiku, Y.; Yamaguchi, M.; Mukuta, Y.; Harada, T. Common subspace for model and similarity: Phrase learning for caption generation from images. In Proceedings of the IEEE International Conference on Computer Vision 2015, Santiago, Chile, 7–13 December 2015; pp. 2668–2676.

45. Xu, C.; Yang, M.; Ao, X.; Shen, Y.; Xu, R.; Tian, J. Retrieval-enhanced adversarial training with dynamic memory-augmented attention for image paragraph captioning. *Knowl.-Based Syst.* **2021**, *214*, 106730. [CrossRef]
46. Du, Y.; Liu, Z.; Li, J.; Zhao, W.X. A survey of vision-language pre-trained models. In Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22), Vienna, Austria, 23–29 July 2022. [CrossRef]
47. Zhang, L.; Wang, J.; Wang, Y.; Sun, H.; Zhao, X. Automatic construction site hazard identification integrating construction scene graphs with BERT based domain knowledge. *Autom. Constr.* **2022**, *142*, 104535. [CrossRef]
48. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
49. De Curtò, J.; De Zarzà, I.; Calafate, C.T. Semantic Scene Understanding with Large Language Models on Unmanned Aerial Vehicles. *Drones* **2023**, *7*, 114. [CrossRef]
50. Tsai, W.-L.; Le, P.-L.; Ho, W.-F.; Chi, N.-W.; Lin, J.J.; Tang, S.; Hsieh, S.-H. Construction Safety Inspection with Contrastive Language-Image Pre-Training (CLIP) Image Captioning and Attention. *Autom. Constr.* **2025**, *169*, 105863. [CrossRef]
51. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.* **2020**, *21*, 1–67.
52. Dinh, N.N.H.; Shin, H.; Ahn, Y.; Oo, B.L.; Lim, B.T.H. Attention-Based Image Captioning for Structural Health Assessment of Apartment Buildings. *Autom. Constr.* **2024**, *167*, 105677. [CrossRef]
53. Tu, Y.; Zhou, C.; Guo, J.; Li, H.; Gao, S.; Yu, Z. Relation-aware attention for video captioning via graph learning. *Pattern Recognit.* **2023**, *136*, 109204. [CrossRef]
54. Li, P.; Gai, S. Single image deraining using multi-scales context information and attention network. *J. Vis. Commun. Image Represent.* **2023**, *90*, 103695. [CrossRef]
55. Dubey, S.; Olimov, F.; Rafique, M.A.; Kim, J.; Jeon, M. Label-attention transformer with geometrically coherent objects for image captioning. *Inf. Sci.* **2023**, *623*, 812–831. [CrossRef]
56. Zhai, P.C.; Wang, J.J.; Zhang, L.T. Extracting Worker Unsafe Behaviors from Construction Images Using Image Captioning with Deep Learning-Based Attention Mechanism. *J. Constr. Eng. Manag.* **2023**, *149*, 04022164. [CrossRef]
57. Huang, L.; Zhao, K.; Ma, M. When to Finish? Optimal Beam Search for Neural Text Generation (modulo Beam Size). In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 9–11 September 2017; Association for Computational Linguistics: Stroudsburg, PA, USA, 2017; pp. 2134–2139.
58. Freitag, M.; Al-Onaizan, Y. Beam Search Strategies for Neural Machine Translation. In Proceedings of the First Workshop on Neural Machine Translation, Vancouver, BC, Canada, 3–4 July 2017; Association for Computational Linguistics: Stroudsburg, PA, USA, 2017; pp. 56–60.
59. Wang, Y.; Xiao, B.; Bouferguene, A.; Al-Hussein, M.; Li, H. Vision-based method for semantic information extraction in construction by integrating deep learning object detection and image captioning. *Adv. Eng. Inform.* **2022**, *53*, 101699. [CrossRef]
60. Duan, R.; Deng, H.; Tian, M.; Deng, Y.; Lin, J. SODA: A large-scale open site object detection dataset for deep learning in construction. *Autom. Constr.* **2022**, *142*, 104499. [CrossRef]
61. Niu, Z.; Zhong, G.; Yu, H. A review on the attention mechanism of deep learning. *Neurocomputing* **2021**, *452*, 48–62. [CrossRef]
62. Li, J.; Monroe, W.; Jurafsky, D. A simple, fast diverse decoding algorithm for neural generation. *arXiv* **2016**, arXiv:1611.08562.
63. Bhatti, S.S.; Gao, X.; Chen, G. General framework, opportunities and challenges for crowdsourcing techniques: A comprehensive survey. *J. Syst. Softw.* **2020**, *167*, 110611. [CrossRef]
64. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W. Bleu: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics 2002, Philadelphia, PA, USA, 6–12 July 2002; pp. 311–318.
65. 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 2015, Boston, MA, USA, 7–12 June 2015; pp. 4566–4575.
66. Lin, C. Rouge: A package for automatic evaluation of summaries. In Proceedings of the Workshop on Text Summarization Branches Out, Barcelona, Spain, 25–26 July 2004; pp. 74–81.
67. Sun, Y.; Gu, Z. Using computer vision to recognize construction material: A Trustworthy Dataset Perspective. *Resour. Conserv. Recycl.* **2022**, *183*, 106362. [CrossRef]