Displays 82 (2024) 102627

Contents lists available at ScienceDirect

Displays

journal homepage: www.elsevier.com/locate/displa

CODP-1200: An AIGC based benchmark for assisting in child language acquisition☆,☆☆

Guannan Leng[1], Guowei Zhang[1], Yu-Jie Xiong*, Jue Chen

School of Electric and Electronic Engineering, Shanghai University of Engineering Science, ShangHai, 201600, China

ARTICLE INFO

Keywords:
Diffusion model
Image captioning
AI generated content
Child language acquisition
Benchmark

ABSTRACT

AIGC (Artificial Intelligence Generated Content) is a novel AI technology that encompasses tasks such as text-to-image generation, text-to-text generation, and image-to-text generation. In the process of child language acquisition, some children may face challenges, exhibiting symptoms such as delayed language development, limited vocabulary, and poor expressive ability. To address this issue, the "Look and Speak" method can be employed, which allows children to learn and express language by observing images. In our paper, we build a dataset, named CODP-1200, benchmark for assisting in children language acquisition, which is curated and augmented using AIGC techniques. The dataset consists of 1200 children cartoon images paired with 6000 corresponding sentences that are used to describe them. Initially, we carefully curated and selected twelve Chinese language textbooks, ranging from the first to the sixth grade, as part of the primary compulsory education curriculum, to construct the foundational corpus. Based on the original data, two famous large language models ChatGPT and SparkDesk are employed for data augmentation, subsequently. Finally, the ERNIE-ViLG is utilized to generate children's style images corresponding to the textual descriptions. In addition, based on our proposed dataset, we propose a benchmark approach called DDMXCap, which is a diffusion-based model for image captioning, specifically from image to text. Experimental results demonstrate that our method achieves promising performance in children's image captioning tasks and provides a standardized learning process for child language acquisition. The implementation codes for our approach and build dataset are available at https://github.com/Leng-bingo/Chinese-Child-Captions.

1. Introduction

During the language acquisition stage for the child, the oral description of pictures plays a crucial role [1]. Although most children can naturally acquire language within a certain age range, there are still some children who experience delays in language development or face language barriers. Through comprehension of visual information, children utilize their acquired vocabulary to describe the content of the images they perceive, thereby enhancing their linguistic skills. Unfortunately, children with visual impairments encounter challenges in effectively receiving visual elements due to their limited visual acuity [2]. With the rapid development of artificial intelligence, new possibilities have emerged for children's language education, such as the use of image captioning methods to support language acquisition. Image captioning has emerged as a technology that converts visually extracted features from images by computers into high-level semantic information. It can assist children with visual impairments in learning to describe visual materials orally [3]. During the application of AIGC assisted technology, a key issue is how to evaluate the effectiveness of outputs generated by artificial intelligence. An obvious potential answer for this problem is a public dataset. It can serve as a basis for related studies. In particular, evaluating the actual performance of different AIGC assisted techniques is valuable for selecting a more suitable solution.

In recent years, the Flickr8K and Flickr30K datasets [4] are compiled from Flickr, a photo-sharing website owned by Yahoo. These datasets comprise a wide range of images, encompassing various subjects such as people and landscapes, with 8000 and 30,000 images respectively. The MS COCO Caption dataset is an extension of MS COCO (Microsoft Common Objects in Context) dataset. [5]. It uses Amazon's

☆ This work is supported by the National Natural Science Foundation of China under Grant No. 62006150, and the Science and Technology Commission of Shanghai Municipality, China under Grant No. 21DZ2203100.

☆☆ This paper was recommended for publication by Prof Guangtao Zhai.

* Corresponding author.
E-mail addresses: 805477481@qq.com (G. Leng), 2510047926@qq.com (G. Zhang), xiong@sues.edu.cn (Y.-J. Xiong), jadeschen@sues.edu.cn (J. Chen).

[1] The authors contributed equally to this work and should be considered the co-first authors.

https://doi.org/10.1016/j.displa.2023.102627
Received 16 July 2023; Received in revised form 22 November 2023; Accepted 25 December 2023
Available online 29 December 2023

Mechanical Turk service to generate descriptive captions for images. It has gained significant popularity as a large-scale image captioning benchmark, including 330,000 images accompanied by diverse and comprehensive annotations.

However, there is still a gap in the availability of dedicated datasets specifically designed for children's image captioning. For instance, the current datasets are primarily sourced from the internet and mainly consist of categories such as animals, flowers, trees, buildings, and street scenes. These datasets have few image related to children daily life or preschool knowledge. Furthermore, the image style of these datasets are realistic style, rather than a cartoon style that is more conducive to children's understanding.

To address these issues, we build a multi-scene, multi-semantics dataset specifically designed for children, named CODP-1200, it contains 1200 cartoon images. The dataset consists of sentences extracted from twelve Chinese language textbooks from the compulsory education curriculum spanning grades one to six. The corresponding images are generated in a children's cartoon painting style using the ERNIE-ViLG [6]. The CODP-1200 dataset is specifically designed for young children, incorporating a combination of volunteer annotation, large language models (ChatGPT, SparkDesk), and augmented translation techniques. This approach ensures the generation of age-appropriate text, aligning with the cognitive abilities of children.

Additionally, we also propose a diffusion-based image captioning approach, called **D**iscrete **D**iffusion **M**odel with **X**-Linear Attention for Image Captioning(DDMXCap). By integrating the X-Linear attention module into the diffusion model, it facilitates enhanced focus on regions of interest, consequently aiding visually impaired children in acquiring more precise information from the image. The scores of BLEU-1, BLEU-4, METEOR, ROUGL-L and CIDEr of DDMXCap on our proposed CODP-1200 are 54.16, 26.62, 24.63, 47.36, and 151.12. The main contributions of this study are summarized as follows:

- We build a new dataset for **C**hildren **O**ral **D**escription of **P**icture (CODP-1200). Each unit in the dataset consists of two similar cartoon images and five corresponding image description sentences. The creation of this dataset addresses the existing gap in child image captioning. And it offers valuable resources for studying visual restoration in children. By conducting experiments on our dataset, we have verified its quality and effectiveness. The code for these datasets can be found at https://github.com/Lengbingo/Chinese-Child-Captions.
- We propose a novel approach, Discrete Diffusion Model with X-Linear Attention for Image Captioning (DDMXCap), which incorporates an X-Linear attention module to capture image features that are highly correlated with the corresponding textual descriptions. By leveraging this approach, we are able to generate more accurate and precise image captions.
- The proposed method achieved a BLEU-4 score of 26.62 and a CIDEr score of 151.12 on the CODP-1200 dataset. The scores are improved by 0.77 and 31.02 compared to the baseline methods.

2. Related work

2.1. Generative model

Recently, Jonathan et al. [7] proposed the powerful Denoising Diffusion Probabilistic Models (DDPM), followed by DDIM [8], Semantic Guidance Diffusion [9], GLIDE [10], and DALL-E-2 [11]. These models made significant contributions to the field of image generation and gradually replaced the generative models previously represented by GAN [12]. Karras et al. [13] proposed Style-GAN, which can automatically learned, unsupervised separation of high-level attributes and stochastic variation in the generated images. Liu et al. [9] proposed a novel image generation method that incorporates semantic diffusion guidance to enhance the quality and controllability of image generation by leveraging semantic information. In this approach, the semantic vector is combined with random noise, and the diffusion process is systematically applied to propagate the vector while calculating the similarity between the generated image and the semantic vector. This similarity measure guides the subsequent iterations of the process.

Sutskever et al. [14] proposed an end-to-end sequence training method that makes minimal assumptions about the sequence structure. They employed multiple layers of LSTM to map the input sequence into a fixed-dimensional vector, and then used another deep LSTM to decode the target sequence from the vector. BERT [15] is designed to pretrain deep bidirectional representations of unlabeled text by jointly conditioning on left and right context at all layers. This distinguishes BERT from previous language representation models. Brown et al. [16] proposed GPT-3. For all tasks, GPT-3 is applied without any gradient updates or fine-tuning, with tasks and few-shot demonstrations specified purely via text interaction with the model. As a result, fine-tuning the pretrained BERT model by adding an output layer alone allows the development of state-of-the-art models for various tasks. It includes question answering and language inference, without requiring significant modifications to the task-specific architecture.

2.2. Image captioning dataset

Currently, the field of image captioning benefits from the availability of several datasets, including the Flickr8K and Flickr30K datasets [4], as well as the MS COCO Captions dataset [5]. These datasets consist of diverse images capturing complex scenes featuring people, animals, and everyday objects. The MS COCO Captions images were gathered by searching for pairs of 80 object categories and various scene types on Flickr. The goal of the MS COCO image collection process was to gather images containing multiple objects in their natural context. Given the visual complexity of most images in the dataset, they posed an interesting and difficult challenge for image captioning. The text generation process involved leveraging Amazon's Mechanical Turk service, where human annotators generate a minimum of five captions for each image, resulting in a comprehensive collection of approximately 1.5 million captions. On the other hand, Flickr serves as an image search website that retrieves relevant images based on the provided input information. The Flickr datasets consisted of everyday activities, events and scenes. Followed annotation guidelines and implemented quality controls to correct spelling mistakes, remove ungrammatical or non-descriptive sentences, as well as eliminate inconsistencies in the annotations, as used in Hodosh Hodosh et al. [17].

VizWiz [18] is a mobile application project designed to facilitate real-time assistance for visually impaired individuals by recruiting sighted workers. Through this application, users can capture images using their smartphones, ask questions, and receive multiple answers. By leveraging the power of the internet, individuals with visual impairments gain access to information that would otherwise be inaccessible to them, such as reading menus, identifying canned food, or determining the availability of free benches in a park. The VizWiz-Captions dataset [19] is a dedicated dataset specifically designed to address the needs of visually impaired individuals. All images in this dataset are collected from actual blind users utilizing the VizWiz, ensuring its relevance to the users' real-world requirements and encompassing various practical and intricate issues they encounter. Trained personnel annotate the image captionings, and metadata for each image is also collected, indicating the presence of text and the severity of image quality issues. This comprehensive dataset enables systematic analysis based on these factors.

The Conceptual Captions (CC) dataset [20] is a collection of image URL, caption pairs used for training and evaluating machine learning image captioning systems. The dataset consists of two versions: CC3M with approximately 3.3 million images and CC12M with approximately 12 million images. It is automatically gathered from the web using a simple filtering process to collect weakly related descriptions. In

contrast to the images in the MS COCO Captions dataset, the images in the Conceptual Captions dataset are sourced from the web along with their original descriptions, representing a wider range of styles. However, it should be noted that the availability of images is not guaranteed since only image URLs are provided, and the quality of the accompanying text cannot be guaranteed.

2.3. Image captioning

Image captioning serves as a crucial link between images and text, making it a prominent research area in the field of artificial intelligence. It provides great assistance to visually impaired individuals, enabling them to observe information in the images. By image captioning, visually impaired individuals can enhance their understanding of images and enrich their visual experience.

Initially, Vinyals et al. [21] drew inspiration from machine translation and adopted an encoder–decoder architecture for image captioning. Later, Kelvin et al. [22] further enhanced the model by utilizing a combination of CNN and long short-term memory (LSTM) as the encoder and decoder, respectively. Anderson et al. [23] made significant contributions by integrating object detection techniques into the field of natural image captioning. They proposed a novel approach that combined Bottom-Up visual feed-forward attention with Top-Down non-visual or task-specific contextual attention. Qin et al. [24] introduced a novel approach called Look Back and Predict Forward (LBPF) consisting of two main components. The two probabilities generated by the model are combined together to predict the current word.

2.3.1. Transformer-based image captioning

Due to the inherent limitations of RNN models' sequential nature, there is a drawback in effectively retaining distant past information within the sequence. While LSTM models can partially mitigate the long-range dependency problem in RNN, they have a limited memory capacity for information storage. Ashish et al. [25] proposed self-attention mechanisms as a replacement for recursion and convolution, addressing the limitation of RNN and its variants in parallel computation. Lu et al. [26] proposed an adaptive attention module capable of dynamically determining the focus of each decoding stage on specific regions of an image. Chen et al. [27] proposed a spatial and channel attention mechanism that integrates features from various spatial regions and channels in an image. Pan et al. [28] proposed an optimization technique for the attention module utilizing bilinear pooling. Cornia et al. [29] proposed a natural image captioning model based on the Transformer architecture, integrating the Meshed-Memory structure. It incorporates image regions and textual features, utilizing memory slots for keys and values in self-attention to add high-level information and prior knowledge.

2.3.2. Diffusion-based image captioning

The Denoising Diffusion Probabilistic Models (DDPM) are initially proposed by Jonathan et al. [7]. The DDPM consists of two main phases: the forward process, also referred to as the diffusion process, which gradually transforms the original image into a fully noisy image, and the inverse process, known as the denoising process, which progressively restores the noisy image to its original state. Regardless of the direction (forward or backward), the process is modeled as a parameterized Markov chain. To accelerate the generation process, Song et al. [8] proposed the Denoising Diffusion Implicit Models (DDIM). DDIM shares the same training objective as the DDPM but does not impose the Markov chain constraint on the diffusion process. It allows the smaller sampling steps during generation.

Xu [30] proposed an image captioning approach that combines the diffusion model with CLIP. By introducing the diffusion model, it becomes possible to generate image captionings without the need for explicit alignment between images and texts. It leverages the output of the CLIP model as the initial state, which undergoes gradual diffusion through multiple iterations. In each iteration, the diffusion model randomly generates text and updates the state based on this text and the current state. Austin et al. [31] proposed a structured denoising diffusion model based on a discrete-space formulation. Li et al. [32] proposed a text generation model based on the diffusion model to achieve controllability. To tackle the issues of minimizing irrelevant content and enhancing fluency, it takes a random noise vector as input and utilizes a diffusion process to progressively generate text with targeted semantics.

3. CODP-1200 dataset

The deficiency of existing benchmark datasets in the number of images and scene types and diversity of descriptions hinders the advancement of novel captioning approaches. Therefore, we build the CODP-1200 dataset, some samples in CODP-1200 are shown in Fig. 1. It is a Chinese dataset that consists of images in a children's cartoon style along with their corresponding content descriptions. The CODP-1200 dataset contains 600 units, consisting of 1200 images and 3000 sentences. Each unit consists of five semantically similar sentences and two images generated based on the corresponding semantics. The resolution of the images is 1024×1024. All text annotations in the dataset are generated through a voluntary annotation process involving three experienced volunteers. These annotations are subsequently refined through rewriting and data augmentation using large language models, including ChatGPT and SparkDesk. We conduct this process to ensure the dataset's textual data exhibited both quality and diversity. The images in the dataset are generated using the ERNIE-ViLG from Baidu's Wenxin AI model suite [6]. This model, given relevant prompts, is capable of generating cartoon-style images that match the provided descriptions. The CODP-1200 dataset offers several advantages compared to currently available benchmark datasets. Firstly, our texts are sourced from twelve Chinese language textbooks from the compulsory education curriculum spanning grades one to six, which gives the dataset a sense of authority, universality, and makes it more appealing to children. Secondly, it covers a broader range of age-appropriate scenarios, catering to the needs of children across different age groups. The CODP-1200 dataset includes a series of images with related descriptions, aimed at allowing children to describe images through observation. This process not only promotes children's visual perception ability but also helps children enhance vocabulary accumulation, improve semantic understanding and expression, and foster creative thinking. Additionally, the dataset exhibits a rich and diverse vocabulary in its descriptions, providing a wide range of textual representations. The methodology for constructing the dataset is illustrated in Fig. 2. It can be accessed through the following link: https://github.com/Leng-bingo/Chinese-Child-Captions. In the following sections, we will provide a detailed overview of the dataset's creation process and analysis results.

3.1. Sentence collection

Considering the target audience of the dataset as children, we selected Chinese textbooks related to them as the source for text collection. To ensure that the selected sentences are understandable for children and guarantee the richness and diversity of the sentences, we manually screened and selected twelve Chinese language textbooks from the first to the sixth grade of primary education to constitute the basic corpus. A total of 600 sentences are collected, with an average of 50 sentences per textbook. The details are shown in Table 1. These sentences cover different scenarios, characters, and plots, ensuring they are suitable for all children. In the selection process, we carefully eliminated redundant, meaningless words and potentially confusing interference terms to ensure that each sentence accurately conveys the intended information. Additionally, we moderately modified and polished the sentences to make them more concise, clear, and easy

Line 1 (Origin text): 在沙滩上，我们垒起城堡。
Line 2 (Volunteer A): 一个男孩在沙滩上堆堡垒。
Line 3 (Volunteer B): 在沙滩上有一个男孩正在用沙子堆堡垒。
Line 4 (Translating): 一个男孩正在海滩上建造一座堡垒。
Line 5 (ChatGPT): 一个男孩在沙滩上堆沙堡。
Line 6 (SparkDesk): 一个男孩在沙滩上认真的堆堡垒。

Line 1 (Origin text): 她坐在鸟儿的背上。
Line 2 (Volunteer A): 一个小女孩坐在一只飞翔的鸟儿身上。
Line 3 (Volunteer B): 一位小女孩骑在一只翱翔的鸟身上。
Line 4 (Translating): 一个小女孩坐在一只会飞的鸟上。
Line 5 (ChatGPT): 一个小女孩坐在一只飞翔的鸟儿背上。
Line 6 (SparkDesk): 小女孩骑鸟飞翔于空中。

Line 1 (Origin text): 邻家的小猫小狗小鸡小鸭。
Line 2 (Volunteer A): 石头旁边有一只小狗。
Line 3 (Volunteer B): 有一只小狗在石头旁边玩耍。
Line 4 (Translating): 石头旁边有一条小狗。
Line 5 (ChatGPT): 石头旁边伴着一只小狗。
Line 6 (SparkDesk): 石头旁坐着一只小狗。

Fig. 1. Examples of the CODP-1200. Line 1 presents the original text extracted from twelve Chinese language textbooks from the compulsory education curriculum spanning grades one to six. Line 2 and line 3, volunteers A and B respectively provided rewritten versions of the original text. Line 4 represents the result of translating the text from Chinese to English using the Baidu Translation API, followed by translating it back to Chinese. Line 5 and line 6 consist of rewrites generated by the large language models ChatGPT and SparkDesk, respectively.

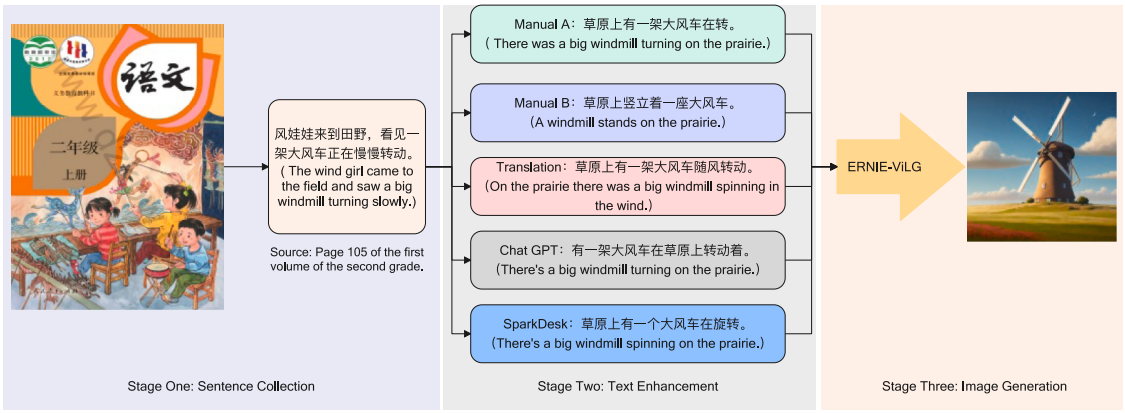


Fig. 2. Process of dataset creation.

to understand. This laid a solid foundation for the construction of subsequent datasets. Through this text selection and polishing process, we ensured the collection of high-quality sentences from the textbooks. These sentences not only hold educational significance but also engage children's interest and curiosity. They establish a robust basis for constructing the dataset and provide substantial support for further tasks and research endeavors.

3.2. Sentence augmentation with AIGC

After the first step of sentence collection, we successfully gathered 600 eligible sentences from twelve Chinese language textbooks from the compulsory education curriculum spanning grades one to six. To further enrich the dataset, we generate four alternative expressions for each sentence. We employ various methods to achieve this objective.

Firstly, volunteers rewriting the collected sentences from the initial step to obtain new sentences with similar meanings. It involves careful thought and deliberate selection to ensure that the new sentences are semantically similar to the original ones but have differences in expression. Secondly, we use the large language models such as ChatGPT and SparkDesk for sentence augmentation. ChatGPT, based on the GPT (Generative Pre-trained Transformer) architecture, is employed in data augmentation tasks to generate new text data that is grammatically correct and highly relevant. SparkDesk focuses on creating text that is rich in emotion and creativity, and its generated results may be more imaginative and emotional. Utilizing these two methods for text

Table 1
Selection of textual details.

Textbooks	Number of sentences
First grade First semester	49
First grade Second semester	48
Second grade First semester	56
Second grade Second semester	44
Third grade First semester	55
Third grade Second semester	72
Fourth grade First semester	41
Fourth grade Second semester	66
Fifth grade First semester	62
Fifth grade Second semester	35
Sixth grade First semester	48
Sixth grade Second semester	24
Total	600

data augmentation can provide the dataset with richer semantic expressions. Lastly, we employ the translation enhancement feature of Baidu Translation API. Specifically, we translated the Chinese sentences into English and then translated them back into Chinese. This bidirectional translation process helped us generate additional variations of sentences, where each sentence have the same semantics but differed in specific expressions.

We ultimately five semantically equivalent but slightly different expressions for each sentence. This diverse collection of sentences provides broader coverage for our dataset, enabling models to have a more comprehensive and flexible understanding and generation of text.

3.3. Image generation with ERNIE-ViLG

All the images in this dataset are generated by the ERNIE-ViLG [6], which is a large-scale pre-trained generative model designed to handle multimodal tasks, such as understanding textual information and generating images in the corresponding style based on that understanding. Specifically for generating children's style images that correspond to text descriptions, ERNIE-ViLG can leverage its pre-trained capabilities to comprehend the content, style, and emotions described in the text and transform it into colorful, concise, or cartoon-like child-friendly images.

ERNIE-ViLG adopts an autoregressive generation mode, which enables unified modeling of image and text generation tasks, thereby capturing semantic alignment between modalities and improving the effectiveness of bidirectional image–text generation tasks. It possesses powerful capabilities to generate images based on natural language intelligently. Users can freely input descriptive text without content restrictions, and ERNIE-ViLG can accurately understand the descriptions and support image generation optimization through the configuration of hyperparameters, thus achieving stable and controllable image generation quality.

Firstly, we feed the first volunteerly rewritten sentence from each data unit into the ERNIE-ViLG model. The model generated four corresponding images based on the descriptive text. Subsequently, we meticulous volunteer selection and chose the two images that best matched the description to be saved. Through this construction process, we ensured a high level of consistency and similarity between the images and the descriptive text in the dataset. By combining the generative capacity of the ERNIE-ViLG model with the expertise of volunteer selection, we are able to filter out the highest quality and most appropriate images, thereby ensuring the quality and accuracy of the dataset.

By utilizing the ERNIE-ViLG and volunteer selection, we ensure a close correspondence between the images and descriptive text in the dataset. The construction of this dataset not only offers high-quality image data for model training and research but also fosters advancements in the domains of image generation and text generation.

Table 2
Characterization of our CODP-1200 dataset. The table presented below showcases the average counts of image-to-text descriptions for each image in the dataset, along with the corresponding number of Chinese characters(words), nouns(n), verbs(v), adjectives(adj), and numerals(num) contained in the descriptions. The first six rows pertain to the statistics derived from individual Chinese language textbooks, while the final row provides an overview of the dataset's overall statistics.

	Average count per image				
	words	n	v	adj	num
First grade	11.6	2.3	1.0	0.7	0.8
Second grade	11.1	2.3	1.1	0.7	0.8
Third grade	11.1	2.6	1.1	0.7	0.7
Fourth grade	10.8	2.3	1.0	0.6	0.8
Fifth grade	10.7	2.2	1.1	0.6	0.9
Sixth grade	11.2	2.4	1.1	0.8	0.8
CODP-1200	11.3	2.3	1.1	0.7	0.8
	All count for all images				
	words	n	v	adj	num
First grade	5,609	1,113	491	317	393
Second grade	5,584	1,141	534	347	388
Third grade	7,758	1,667	705	510	423
Fourth grade	5,760	1,236	560	311	412
Fifth grade	5,181	1,048	521	299	426
Sixth grade	4,047	849	404	286	277
CODP-1200	33,939	7,054	3,215	2,070	2,319

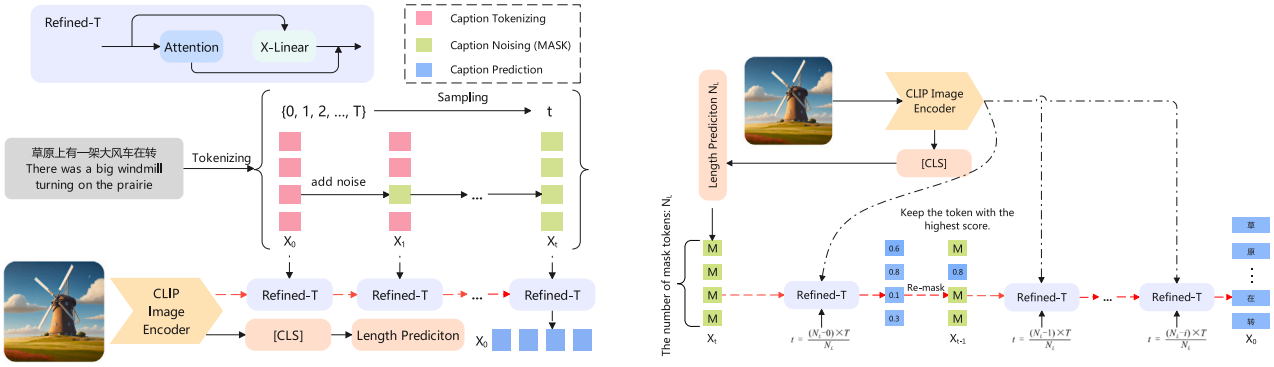
3.4. Dataset analysis

Quality of Images. We initially assessed the quality of all descriptive texts by examining their semantic similarity, clarity, and non-repetitiveness within each unit. Subsequently, we evaluated the images to ensure they effectively conveyed the content described in the texts. It is crucial to consider that the images are generated by artificial intelligence and may possess certain flaws. Additionally, as the dataset is specifically curated for children, it is imperative that the generated image content aligns with their perspective. Some images may exhibit poor quality due to inappropriate colors (e.g., predominantly black or dark colors) or contain elements beyond children's comprehension. we conducted a thorough evaluation of image quality with the assistance of five volunteers who assessed both content and artistic style. The majority of images are deemed of sufficient quality, featuring content and artistic style suitable for children. Only a negligible number of images are considered to have insufficient quality to match the descriptive texts. Nevertheless, humans are still able to identify the key elements described in the texts within the images.

CODP-1200 Dataset Characterization. We quantified various characteristics of the descriptive texts, including their length as well as the quantities of objects, descriptors, actions, and relationships they encompass. To accomplish this, we employed the Jieba Chinese word segmentation tool to calculate the average number of words per description, as well as the average and total numbers of nouns, adjectives, verbs, numerals, and spatial relational words present in each sentence. The characterization of our CODP-1200 dataset is presented in Table 2. It reveals that the typical sentence consists of approximately 11 Chinese characters and includes two to three objects (nouns), one action (verb), one descriptor (adjective), one quantifier (numeral), and one relationship (spatial relational word). Illustrative examples with similar structures are "A boy is earnestly reading under the lamp" and "A sturdy tree is adorned with an ancient bronze bell". These findings demonstrate the dataset's capacity to encompass a wide range of concepts, incorporating over 33,000 unique words. As a result, it provides an effective solution for child image captioning tasks.

4. Discrete Diffusion Model with X-Linear Attention for Image Captioning

In this paper, we propose a novel image captioning approach called DDMXCap (Discrete Diffusion Model with X-Linear Attention for Image



(a) The process of training. (b) The process of inference.

Fig. 3. Overview of the proposed DDMXCap, consisting of a DDM sub-network, an image encoder such as CLIP-based ones, and a transformer-based sub-network. (a) During training, the caption is tokenized and gradually converted to $[mask]$ by adding noise that depends on the sampled step t. Then, these noisy tokens are fed into a transformer model for clean text token prediction, together with image features. The predicted tokens are used for loss calculating together with GT. The $[CLS]$ token of the CLIP model is used to predict the length N_L of the caption. (b) During inference, all $[mask]$ tokens X_T is the input and the caption length N_L is first predicted. For each step t, we have three inputs for the transformer's Adaptive LayerNorm layer: t depending on N_L and the total noise length T, image features, and previous text tokens X_{t-1}. We retain the token with the highest score each time and gradually infer the initial caption X_0.

Captioning). The CLIP pre-trained model is used in our approach for image encoding, and the alignment between image and text features is improved by incorporating the X-Linear attention module into the discrete diffusion model for image captioning. The application of diffusion models in children's image captioning is a novel approach. The CLIP image encoder allows the model to learn from a robust and general foundation of image–text alignment. The use of the X-Linear attention module is more advanced and efficient than traditional attention mechanisms when processing and analyzing complex, non-linear relationships in images. Combined with bilinear attention to capture higher-order interactions, it facilitates the inference of joint representations between image features and hidden states, promoting sentence generation. Fig. 3 provides an overview of the overall architecture of our proposed DDMXCap.

In this approach, each encoded token undergoes a gradual probabilistic transformation into a mask token, allowing for the addition of noise. To enhance effective attention between image features and text, we employ X-Linear attention. The approach is trained using both image features and noisy text. Image features are extracted using the pre-trained CLIP model, where the $[CLS]$ token is used as a feature for predicting the length of the corresponding text. To predict the token length, a simple MLP is employed. Based on the predicted text length, a sequence of masks is generated, and words with the highest confidence scores are selected. Noise is iteratively removed from the text, with the text length serving as the diffusion step length T.

4.1. Noising and denoising process

The diffusion process involves the gradual addition of Gaussian noise to the initial data until it becomes entirely noise. The traditional continuous diffusion model is a parameterized Markov chain that progressively increases noise to generate training samples. In contrast, the discrete diffusion model applies noise processing at the text level by utilizing a mask token. Through a series of T steps with a certain probability, the text is transformed into noise consisting entirely of mask tokens.

Each word in the caption is represented as a discrete state denoted by x. The noise corresponding to the step with a step size of t is denoted as x_t. The diffusion process consists of a total of T_t steps, and in each step, noise is added to the data x_{t-1} obtained in the previous step. Specifically, each token has a probability of ϵ_t to transition to a special state $[mask]$. If x_{t-1} is not a $[mask]$ token, the transformation probability from Step $t-1$ to Step t is defined as follows:

$$p\left(x_t \mid x_{t-1}\right) = \begin{cases} \eta_t, x_t = x_{t-1} \\ \epsilon_t, x_t = [\text{ mask }] \\ 1\text{-}\epsilon_t - \eta_t, \text{ otherwise} \end{cases} \tag{1}$$

In other words, each token x_{t-1} is assigned probabilities ϵ_t for being replaced by a special marker $[mask]$, η_t for remaining unchanged, and $\theta_t = 1 - \epsilon_t - \eta_t$ for being replaced by any other token from the vocabulary except for $[mask]$. If x_{t-1} is the $[mask]$ token, the transformation probability from Step $t-1$ to Step t is defined as:

$$p\left(x_t \mid x_{t-1}\right) = \begin{cases} 1, x_t = [\text{ mask }] \\ 0, \text{ otherwise} \end{cases} \tag{2}$$

By employing the aforementioned noise addition method, when the diffusion step size T is sufficiently large, all encoded word tokens are replaced with a special token, denoted as $[mask]$.

The denoising process refers to the iterative removal of noise from a random noise sequence in order to restore the original signal. We initiated the process by using a noise sequence composed of $[mask]$ elements. Then, we applied a Transformer network and the X-Linear attention mechanism to perform reverse projection denoted as $p(x_{t-1} \mid x_t, y)$, where y represents the image features extracted from the fine-tuned CLIP pre-trained model's image encoder. The utilization of X-Linear attention helped identify regions that exhibited strong correspondence between the text and image features, thereby enhancing the accuracy of caption generation. Additionally, we employed a sine function to encode the position of each time step t.

$$p = t * \text{step}_{\text{scale}}/T, \tag{3}$$

$$\text{PE}_i = \begin{cases} \sin\left(p/10000^{2i/d_{\text{model}}}\right), i < d_{\text{model}}/2 \\ \cos\left(p/10000^{2i/d_{\text{model}}}\right), i \geq d_{\text{model}}/2 \end{cases} \tag{4}$$

where $\text{step}_{\text{scale}}$ is the wavelength, i.e., 8000 in our experiments, and d_{model} is the hidden dimension.

4.2. Training and inference

The traditional continuous diffusion model recovers initial information by predicting the noise distribution, whereas the discrete diffusion model performs denoising at the sentence level using masking. Taking inspiration from DDSM [31] and DDCap [33], we modify the step-by-step process from Step t to $t-1$ to directly compute the initial text at Step 0, combined with image features for training. An overview of

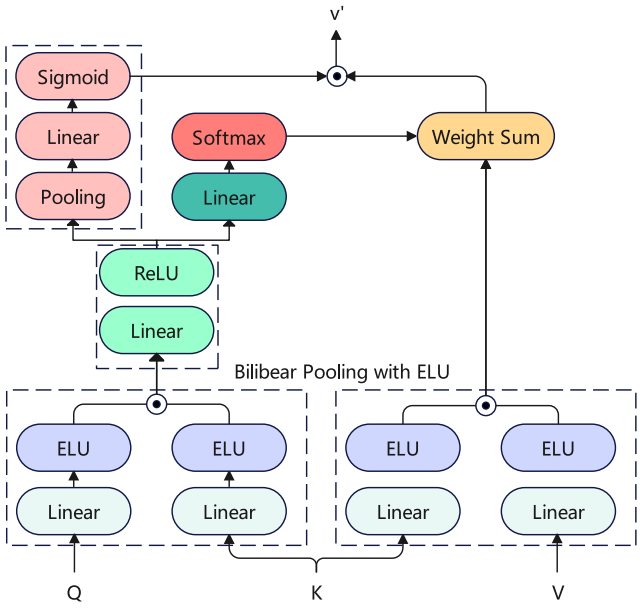


Fig. 4. A schematic diagram of X-Linear attention block plus ELU to capture infinity order feature interactions.

the overall process is depicted in Fig. 3. In the experiments, we set the maximum length of each caption to 20, which corresponds to the maximum step of 20. The process from Step $t-1$ to t is as follows:

$$q(x_t \mid x_{t-1}) = \mathrm{Cat}(x_t; p = x_{t-1}\mathbf{Q}_t) \tag{5}$$

Where $\mathrm{Cat}(x; p)$ represents a categorical distribution over the one-hot row vector x. The transition vector $\mathbf{Q}$ is defined as follows: $[\mathbf{Q}_t]_{ij} = q(x_t = j \mid xt-1 = i)$. The process from Step 0 to t is as follows:

$$q\left(x_t \mid x_0\right) = \mathrm{Cat}\left(x_t; p = x_0\overline{\mathbf{Q}}_t\right), \quad \overline{\mathbf{Q}}_t = \mathbf{Q}_1 \ldots \mathbf{Q}_t \tag{6}$$

In summary, we can train the model by introducing noise to the initial text and combining it with image features.

$$\begin{aligned} q\left(x_{t-1} \mid x_t, x_0, y\right) &= \frac{q\left(x_t \mid x_{t-1}, x_0, y\right) q\left(x_{t-1} \mid x_0, y\right)}{q\left(x_t \mid x_0, y\right)} \\ &= \mathrm{Cat}\left(x_{t-1}; p = \frac{x_t\mathbf{Q}_t^{\top} \odot x_0\overline{\mathbf{Q}}_t}{x_0\overline{\mathbf{Q}}_t x_t^{\top}}\right) \end{aligned} \tag{7}$$

During the inference stage, the $[CLS]$ token of the CLIP model is first utilized as a feature to predict the length of the corresponding image caption. A simple MLP structure is used to predict the length T of the caption. Then, starting from the $[mask]$ token x_T of length T, the trained denoising network is applied, and the image feature $p_\theta(x_0 \mid x_t, y)$ is combined to directly predict x_0. Next, by adding the noise from Step $t-1$ to the predicted original text x_0 through a Markov chain, x_{t-1} is obtained:

$$\begin{aligned} q\left(x_{t-1} \mid x_t, y\right) &\approx \mathbb{E}_{x_0 \sim p_\theta(x_0 \mid x_t)} q\left(x_{t-1} \mid x_t, x_0, y\right) \\ &\propto \mathbb{E}_{p_\theta(x_0 \mid x_t)} q\left(x_t \mid x_{t-1}\right) q\left(x_{t-1} \mid x_0\right) \\ &= x_t\mathbf{Q}_t^T \cdot x_0\overline{\mathbf{Q}}_{t-1} \end{aligned} \tag{8}$$

In summary, after T steps, we can gradually restore the original text x_0.

4.3. X-linear attention

Traditional attention modules calculate in calculating interactions between different components, but their limitation lies in relying solely on first-order feature interactions, which hampers their capacity for intricate image caption reasoning. To overcome this limitation, we draw inspiration from the success of bilinear pooling in tasks like fine-grained visual recognition and visual question answering. We propose a unified attention module for image captioning called X-Linear attention, which is constructed using a single X-Linear attention block, following the approach of X-LAN [28]. It enhances the representation power of features by capturing high-order interactions and facilitates the inference of the joint representation of image features and hidden states, as showed in Fig. 4.

Specifically, suppose a query $\mathbf{Q} \in \mathbb{R}^{D_q}$, a set of keys $\mathbf{K} = \{\mathbf{k}_i\}_{i=1}^N$, and a set of values $\mathbf{V} = \{\mathbf{v}_i\}_{i=1}^N$, where $\mathbf{k}_i \in \mathbb{R}^{D_k}$ and $\mathbf{v}_i \in \mathbb{R}^{D_v}$ represent the ith key/value pair. To facilitate the interaction between the query and keys, the X-Linear attention module employs bilinear pooling. This operation produces joint bilinear query–key representations, denoted as $\mathbf{B}_i^k \in \mathbb{R}^{D_B}$, for each query–key pair:

$$\mathbf{B}_i^k = \sigma\left(\mathbf{W}_k\mathbf{k}_i\right) \odot \sigma\left(\mathbf{W}_q^k\mathbf{Q}\right) \tag{9}$$

In the given equation, the embedding matrices $\mathbf{W}_k \in \mathbb{R}^{D_B \times D_k}$ and $\mathbf{W}_q^k \in \mathbb{R}^{D_B \times D_q}$ represent the embedding matrices. The ReLU unit is denoted by σ, and the symbol $\odot$ signifies element-wise multiplication. Consequently, the bilinear query–key representation $\mathbf{B}_i^k$ effectively captures the second-order feature interactions between query and key.

Next, based on all the bilinear query–key representations $\{\mathbf{k}_i\}_{i=1}^N B_i^k{}_{i=1}^N$, we derive two types of bilinear attention distributions to effectively capture spatial and channel information across all values. To accomplish this, each bilinear query–key representation undergoes projection using two embedding layers, resulting in the generation of a spatial bilinear attention distribution. This distribution is then normalized using a softmax layer.

$$\mathbf{B}_i'^k = \sigma\left(\mathbf{W}_B^k\mathbf{B}_i^k\right), b_i^s = \mathbf{W}_b\mathbf{B}_i'^k, \beta^s = \mathrm{softmax}\left(\mathbf{b}^s\right) \tag{10}$$

where $\mathbf{W}_B^k \in \mathbb{R}^{D_C \times D_B}$ and $\mathbf{W}_b \in \mathbb{R}^{1 \times D_c}$ are embedding matrices. The transformed bilinear query–key representation is denoted as $\mathbf{B}_i'^k$, and b_i^s represents the ith element in $\mathbf{b}^s$. Each element β_i^s in β^s represents the normalized spatial attention weight for each key/value pair. Additionally, we incorporate the squeeze-and-excitation operation on all the transformed bilinear query–key representations $\{\mathbf{B}_i'^k\}_{i=1}^N$ to measure channel-wise attention. Specifically, we aggregate the transformed bilinear query–key representations using average pooling, resulting in a global channel descriptor $\overline{\mathbf{B}}$.

$$\overline{\mathbf{B}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{B}_i'^k \tag{11}$$

Afterward, a sigmoid gating mechanism is applied to the global channel descriptor $\overline{\mathbf{B}}$, resulting in the generation of the channel-based attention distribution β^c through the subsequent excitation operation.

$$\mathbf{b}^c = \mathbf{W}_e\overline{\mathbf{B}}, \beta^c = \mathrm{sigmoid}\left(\mathbf{b}^c\right) \tag{12}$$

where $\mathbf{W}_e \in \mathbb{R}^{D_B \times D_c}$ is embedding matrix.

Finally, the X-Linear attention block accumulates the enhanced bilinear values from both spatial and channel bilinear attention to generate the attentional value feature $\hat{\mathbf{v}}$.

$$\begin{aligned} \hat{\mathbf{v}} &= F_{X-\text{ Linear}}(\mathbf{K}, \mathbf{V}, \mathbf{Q}) = \beta^c \odot \sum_{i=1}^{N}\beta_i^s\mathbf{B}_i^v \\ \mathbf{B}_i^v &= \sigma\left(\mathbf{W}_v\mathbf{v}_i\right) \odot \sigma\left(\mathbf{W}_q^v\mathbf{Q}\right) \end{aligned} \tag{13}$$

where $\mathbf{B}_i^v$ represents the enhanced value of the bilinear pool with respect to the query $\mathbf{Q}$, and each value $\mathbf{v}_i$, $\mathbf{W}_v \in \mathbb{R}^{D_B \times D_v}$, $\mathbf{W}_q^v \in \mathbb{R}^{D_B \times D_q}$ are embedding matrices. In contrast to conventional attention modules, X-Linear attention block uses higher-order feature interactions through bilinear pooling, resulting in more expressive attention features.

5. Experiments and analysis

5.1. Performance metrics

The most commonly used evaluation metrics for caption generation include the Bilingual Evaluation Understudy (BLEU) [34], which comprises BLEU-1, BLEU-2, BLEU-3, and BLEU-4, the Metric for Evaluation of Translation with Explicit Ordering (METEOR) [35], the Recall-Oriented Understudy for Gisting Evaluation (ROUGE-L) [36], and the Consensus-based Image Description Evaluation (CIDEr) [37]. These evaluation metrics are employed to assess the quality of the generated captions. The evaluation metric CIDEr has a score range from 0 to 5, while the remaining range is from 0 to 1.

BLEU is an automatic evaluation metric that exhibits a strong correlation with human judgments. It assesses the similarity between generated captions and ground truth captions through n-gram matching rules. BLEU-n, which encompasses BLEU-1, BLEU-2, BLEU-3, and BLEU-4, is derived based on n-gram analysis. The use of low-order n-grams (e.g., 1-g) is suitable for evaluating word translation accuracy, while high-order n-grams (e.g., 4-g) effectively measure the fluency of captions. Notably, the variation in sentence length is taken into account.

METEOR calculates scores by aligning the generated captions with the ground truth captions using precision and recall measures across the entire corpus. To improve evaluation accuracy, METEOR uses language-specific resources, including Snowball Stemmers. This consideration allows for partial matches by accounting for words with the same stem. In addition, METEOR employs chunks to evaluate the fluency of captions, where a chunk represents a sequence of contiguous and ordered matches between two captions.

ROUGE-L computes scores by comparing the longest common subsequence between the generated captions and the ground truth captions. The "L" in ROUGE-L specifically denotes the longest common subsequence. One notable advantage of ROUGE-L is its ability to capture the order matching of word sequences at the sentence level.

CIDEr is an evaluation metric that directly captures human preferences for captions. It focuses on assessing whether the captions effectively convey key information. One of the main features of CIDEr is its ability to assign less weight to non-visual information words that commonly appear in all reference labels. This approach helps mitigate the influence of sentence length and word frequency. Additionally, CIDEr incorporates a Gaussian penalty and count constraints based on the difference in length between generated captions and ground truth captions. These modifications aim to align the evaluation criterion more closely with human preferences.

5.2. Implementation details

The experiments are performed our proposed CODP-1200 dataset. The dataset comprises 1200 images, where 960 images are allocated for training, 120 images for validation, and another 120 images for testing. Each image is associated with five annotated sentences. PyTorch is used as the platform, and the entire experiment was implemented under the Linux operating system on a server equipped with an Intel(R) Xeon(R) Gold 6326 CPU @2.90 GHz and NVIDIA A100 GPU*4 @40GRAM.

The CLIP pre-trained model, which has been trained on a large dataset of image–text pairs, is utilized as the image feature extractor. The backbone of the model is the ViT-B/16 architecture. It can be downloaded from the following website, and the baseline for the diffusion model is DDCap [33].

All input images are resized to a resolution of 256 × 256 pixels, and the model is trained using 4 A100 GPUs. The training process employed a batch size of 256, with 64 images allocated to each of the 4 GPUs. To enable communication of statistical data across the GPUs, synchronized batch normalization is utilized. For the diffusion model, the maximum length of each sentence is limited to 20, which corresponds to a maximum diffusion step of 20. During training, the AdamW optimizer with a weight decay of 0.01 is utilized. The learning rate is initially linearly increased to 2×10^{-4} and then decreased using cosine decay until reaching 0. The training process consisted of 150 epochs in total, with a warm-up period of 5 epochs and the training time is 40 h.

Table 3
The influence of different pictures on experimental results.

DDCap	B1	B4	M	R	CIDEr
P1*+*A*+*B*+*T*+*C*+*S	55.99	24.17	23.65	46.31	117.05
P2*+*A*+*B*+*T*+*C*+*S	54.47	24.36	23.06	46.01	114.5
P1*+*P2*+*A*+*B*+*T*+*C*+*S	**56.31**	25.85	24.01	47.73	120.1
Our	B1	B4	M	R	CIDEr
P1*+*A*+*B*+*T*+*C*+*S	55.15	25.48	23.82	46.61	140.24
P2*+*A*+*B*+*T*+*C*+*S	53.85	23.82	23.88	**47.93**	144.46
P1*+*P2*+*A*+*B*+*T*+*C*+*S	54.16	**26.62**	**24.63**	47.36	**151.12**

B1, B4, M, and R denote BLEU-1, BLEU-4, METEOR, and ROUGE-L respectively.

5.3. Experiments results

We ensure the diversity and comprehensiveness of the dataset, encompassing a broad of scenes, objects, and visual features. Furthermore, rigorous text quality control measures are implemented to guarantee the accuracy and consistency of the textual descriptions. Subsequently, we conducted three sets of experiments on the dataset and compared our results with those of existing benchmark methods. To evaluate the effectiveness of our approach in image captioning tasks, we employed metrics such as BLEU and CIDEr. The experimental findings are presented in Tables 3–5. ***P1*** and ***P2*** respectively denote the two images generated by the text. ***N*** denotes the original text extracted from the textbook. ***A*** and ***B*** respectively denote volunteers rewritten versions of the original text. ***T*** denotes the text translated from ***A*** to English and then back to Chinese. ***C*** and ***S*** represent the text that are generated through text enhancement by ChatGPT and SparkDesk, respectively.

First, we conduct an evaluation to assess the impact of different images depicting the same text on the image captioning task in our dataset. The results of this evaluation are presented in Table 3. The experimental procedure involved generating two visually similar images based on the text from source ***A*** and training the model using the corresponding image captionings for ***P1***, ***P2***, and ***P1*+*P2***. The findings revealed that training the model with two images associated with a single sentence yielded more suitable image captionings, compensating for any potential information gaps in a single image. Specifically, the BLEU-4 score improved from 23.82 to 26.62, and the CIDEr score increased from 140.24 to 151.12. Additionally, our proposed DDMXCap approach outperformed DDCap, achieving higher scores. The BLEU-4 score improved from 25.85 to 26.62 (an increase of 0.77), and the CIDEr score increased from 120.1 to 151.12 (an increase of 31.02).

Next, we conduct an evaluation to assess the impact of text expression accuracy on image captions, as presented in Table 4. The experiments involved comparing directly selected original texts ***N*** with texts generated by ***A*** and ***B***. The findings indicate that using the original text directly resulted in poorer performance, as it failed to fully capture the events depicted in the images. In contrast, the text expression by ***A*** is more accurate and clear. ***A*** supplemented the original text with additional information and optimized the textual expression, leading to improved results. However, when ***B*** performed semantic rewrites of ***A***, there is a subjective loss of some information. Specifically, the BLEU-4 score increased from 25.14 for the original text to 25.96, and the CIDEr score improved from 118.13 to 148.77. Additionally, our proposed DDMXCap method outperformed DDCap. The BLEU-4 score improved from 25.65 to 25.96 (an increase of 0.30), and the CIDEr score increased from 133.04 to 148.77 (an increase of 15.73).

Table 4
The influence of text expression accuracy on experimental results.

DDCap	B1	B4	M	R	CIDEr
P1+P2+N+C+S+T	52.57	22.13	22.98	44.37	97.5
P1+P2+B+C+S+T	54.6	24.23	23.01	45.27	114.11
P1+P2+A+C+S+T	53.79	25.65	24.25	46.5	133.04
Our	B1	B4	M	R	CIDEr
P1+P2+N+C+S+T	52.66	25.14	23.17	45.05	118.13
P1+P2+B+C+S+T	**55.56**	25.41	**24.42**	**46.94**	128.66
P1+P2+A+C+S+T	49.98	**25.96**	23.53	44.74	**148.77**

Table 5
The influence of different generation models on experimental results.

DDCap	B1	B4	M	R	CIDEr
P1+P2+A+B+S	**55.2**	23.89	**23.29**	**45.95**	116.4
P1+P2+A+B+C	51.84	21.7	22.96	44.48	119.44
P1+P2+A+B+T	52.21	22.71	23.7	46.36	125.8
Our	B1	B4	M	R	CIDEr
P1+P2+A+B+S	51.02	22.3	22.14	42.83	121.16
P1+P2+A+B+C	49.67	22.59	21.98	42.1	124.27
P1+P2+A+B+T	50.65	**24.14**	22.99	44.55	**138.48**

Finally, we conduct an evaluation to assess the influence of generation models on image captions, as presented in Table 5. The experiments involved comparing the text rewritten by the ChatGPT and SparkDesk models. The results indicate that the text generated by the ChatGPT model outperformed that generated by SparkDesk. It can be attributed to the larger training data and more scientifically designed model structure of ChatGPT. However, it is important to note that the combination of ***A***, ***B***, and ***T*** yielded the best results. Large language models still have limitations in rewriting Chinese text compared to human capabilities, and further development is necessary. Human expertise cannot be fully replaced by machines. Specifically, the BLEU-4 score improved from 22.3 for the SparkDesk model to 24.14, and the CIDEr score increased from 121.16 to 138.48. Additionally, our proposed DDMXCap method outperformed DDCap. The BLEU-4 score improved from 22.71 to 24.14 (an increase of 1.43), and the CIDEr score increased from 125.8 to 138.48 (an increase of 12.68).

Based on the comprehensive results of the aforementioned experiments, it conclude that the DDMXCap method exhibits exceptional performance and feasibility on our self-created dataset. The most remarkable outcomes are achieved by combining ***P1***, ***P2***, ***A***, ***B***, ***T***, ***C***, and ***S***, resulting in a BLEU-4 score of 26.62 and a CIDEr score of 151.12. These experimental findings not only validate the effectiveness of our proposed method but also offer substantial empirical evidence in support of our dataset. DDMXCap, due to its utilization of diffusion models, which may result in longer training times. In the future, the model will be optimized to enhance its training speed.

6. Conclusions

In the process of language acquisition for children, the oral description of pictures is a common and effective way of fostering language development. However, visually impaired children often face challenges in completing this process without proper guidance. To address this issue, we build a dataset, named CODP-1200, benchmark for assisting in children language acquisition, which is curated and augmented using AIGC techniques. The dataset consists of 1200 children cartoon images paired with 6000 corresponding sentences that are used to describe them. The construction of the dataset begins with volunteerly selected to constitute the basic corpus from twelve Chinese language textbooks from the compulsory education curriculum spanning grades one to six. Based on the original data, two famous large language models ChatGPT and SparkDesk are employed for data augmentation, subsequently. Finally, the ERNIE-ViLG is utilized to generate children's style images corresponding to the textual descriptions.

In addition, based on our proposed dataset, we propose a benchmark approach called DDMXCap, which is a diffusion-based model for image captioning, specifically from image to text. We compared our model with the baseline methods, and the results showed significant improvements across all evaluation metrics. This indicates that our approach has good adaptability and effectiveness in handling the dataset we proposed.

However, there are still areas that can be further improved. Future endeavors should aim to expand the scale and diversity of the dataset, encompassing a broader range of situations and scenarios. Moreover, further investigation into enhancements and optimizations for the proposed method can significantly improve its performance and adaptability. In the future, we will explore the application to other childhood disorders, such as ADHD (Attention-Deficit/Hyperactivity Disorder). Moreover, we can study the impact of multilingual environments on children's language acquisition and explore the best strategies for multilingual education.

CRediT authorship contribution statement

Guannan Leng: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Formal analysis, Writing – original draft. **Guowei Zhang:** Conceptualization, Investigation, Resources, Data curation, Writing – original draft. **Yu-Jie Xiong:** Conceptualization, Investigation, Resources, Formal analysis, Supervision, Writing – original draft, Writing – review & editing. **Jue Chen:** Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

References

[1] Michael Tomasello, Malinda Carpenter, Josep Call, Tanya Behne, Henrike Moll, Understanding and sharing intentions: The origins of cultural cognition, Behav. Brain Sci. 28 (5) (2005) 675–691.
[2] Xinxin Zhang, Menghan Hu, Yudong Zhang, Guangtao Zhai, Xiao-Ping Zhang, Recent progress of optical imaging approaches for noncontact physiological signal measurement: a review, Advanced Intelligent Systems (2023) 2200345.
[3] Ji-Feng Luo, Yun-Zhu Pu, Jie-Yang Yin, Xiaohong Liu, Tao Tan, Yudong Zhang, Menghan Hu, Is there a difference between paper and electronic chinese signatures?, Advanced Intelligent Systems (2023) 2300439.
[4] Peter Young, Alice Lai, Micah Hodosh, Julia Hockenmaier, From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions, Trans. Assoc. Comput. Linguist. 2 (2014) 67–78.
[5] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, C. Lawrence Zitnick, Microsoft coco captions: Data collection and evaluation server, 2015, arXiv preprint arXiv:1504.00325.
[6] Zhida Feng, Zhenyu Zhang, Xintong Yu, Yewei Fang, Lanxin Li, Xuyi Chen, Yuxiang Lu, Jiaxiang Liu, Weichong Yin, Shikun Feng, et al., ERNIE-ViLG 2.0: Improving text-to-image diffusion model with knowledge-enhanced mixture-of-denoising-experts, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 10135–10145.
[7] Jonathan Ho, Ajay Jain, Pieter Abbeel, Denoising diffusion probabilistic models, Adv. Neural Inf. Process. Syst. 33 (2020) 6840–6851.
[8] Jiaming Song, Chenlin Meng, Stefano Ermon, Denoising diffusion implicit models, 2020, arXiv preprint arXiv:2010.02502.
[9] Xihui Liu, Dong Huk Park, Samaneh Azadi, Gong Zhang, Arman Chopikyan, Yuxiao Hu, Humphrey Shi, Anna Rohrbach, Trevor Darrell, More control for free! image synthesis with semantic diffusion guidance, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 289–299.

[10] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, Mark Chen, Glide: Towards photorealistic image generation and editing with text-guided diffusion models, 2021, arXiv preprint arXiv:2112.10741.

[11] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, Mark Chen, Hierarchical text-conditional image generation with clip latents, 2022, arXiv preprint arXiv:2204.06125.

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio, Generative adversarial networks, Commun. ACM 63 (11) (2020) 139–144.

[13] Tero Karras, Samuli Laine, Timo Aila, A style-based generator architecture for generative adversarial networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4401–4410.

[14] Ilya Sutskever, Oriol Vinyals, Quoc V. Le, Sequence to sequence learning with neural networks, Adv. Neural Inf. Process. Syst. 27 (2014).

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018, arXiv preprint arXiv:1810.04805.

[16] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al., Language models are few-shot learners, Adv. Neural Inf. Process. Syst. 33 (2020) 1877–1901.

[17] Micah Hodosh, Peter Young, Julia Hockenmaier, Framing image description as a ranking task: Data, models and evaluation metrics, J. Artificial Intelligence Res. 47 (2013) 853–899.

[18] Jeffrey P. Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C. Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, et al., Vizwiz: nearly real-time answers to visual questions, in: Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology, 2010, pp. 333–342.

[19] Danna Gurari, Yinan Zhao, Meng Zhang, Nilavra Bhattacharya, Captioning images taken by people who are blind, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16, Springer, 2020, pp. 417–434.

[20] Piyush Sharma, Nan Ding, Sebastian Goodman, Radu Soricut, Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, 2018, pp. 2556–2565.

[21] Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan, Show and tell: A neural image caption generator, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3156–3164.

[22] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, Yoshua Bengio, Show, attend and tell: Neural image caption generation with visual attention, in: International Conference on Machine Learning, PMLR, 2015, pp. 2048–2057.

[23] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, Lei Zhang, 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, 2018, pp. 6077–6086.

[24] Yu Qin, Jiajun Du, Yonghua Zhang, Hongtao Lu, Look back and predict forward in image captioning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 8367–8375.

[25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, Illia Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).

[26] Jiasen Lu, Caiming Xiong, Devi Parikh, Richard Socher, Knowing when to look: Adaptive attention via a visual sentinel for image captioning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 375–383.

[27] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, Tat-Seng Chua, Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5659–5667.

[28] Yingwei Pan, Ting Yao, Yehao Li, Tao Mei, X-linear attention networks for image captioning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10971–10980.

[29] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, Rita Cucchiara, Meshed-memory transformer for image captioning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10578–10587.

[30] Shitong Xu, CLIP-diffusion-LM: Apply diffusion model on image captioning, 2022, arXiv preprint arXiv:2210.04559.

[31] Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, Rianne van den Berg, Structured denoising diffusion models in discrete state-spaces, Adv. Neural Inf. Process. Syst. 34 (2021) 17981–17993.

[32] Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, Tatsunori B. Hashimoto, Diffusion-lm improves controllable text generation, 2022, arXiv preprint arXiv:2205.14217.

[33] Zixin Zhu, Yixuan Wei, Jianfeng Wang, Zhe Gan, Zheng Zhang, Le Wang, Gang Hua, Lijuan Wang, Zicheng Liu, Han Hu, Exploring discrete diffusion models for image captioning, 2022, arXiv preprint arXiv:2211.11694.

[34] Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002, pp. 311–318.

[35] Satanjeev Banerjee, Alon Lavie, 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, 2005, pp. 65–72.

[36] Chin-Yew Lin, Rouge: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, 2004, pp. 74–81.

[37] Ramakrishna Vedantam, C. Lawrence Zitnick, Devi Parikh, Cider: Consensus-based image description evaluation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4566–4575.