Image Captioning with Multi-Context Synthetic Data

Feipeng Ma[1*], **Yizhou Zhou**[2], **Fengyun Rao**[2], **Yueyi Zhang**[1,3†], **Xiaoyan Sun**[1,3†]

[1]University of Science and Technology of China
[2]WeChat, Tencent Inc.
[3]Institute of Artificial Intelligence, Hefei Comprehensive National Science Center
mafp@mail.ustc.edu.cn, {harryizzhou, fengyunrao}@tencent.com, {zhyuey, sunxiaoyan}@ustc.edu.cn

Abstract

Image captioning requires numerous annotated image-text pairs, resulting in substantial annotation costs. Recently, large models (e.g. diffusion models and large language models) have excelled in producing high-quality images and text. This potential can be harnessed to create synthetic image-text pairs for training captioning models. Synthetic data can improve cost and time efficiency in data collection, allow for customization to specific domains, bootstrap generalization capability for zero-shot performance, and circumvent privacy concerns associated with real-world data. However, existing methods struggle to attain satisfactory performance solely through synthetic data. We identify the issue as generated images from simple descriptions mostly capture a solitary perspective with limited context, failing to align with the intricate scenes prevalent in real-world imagery. To tackle this, we present an innovative pipeline that introduces multi-context data generation. Beginning with an initial text corpus, our approach employs a large language model to extract multiple sentences portraying the same scene from diverse viewpoints. These sentences are then condensed into a single sentence with multiple contexts. Subsequently, we generate intricate images using the condensed captions through diffusion models. Our model is exclusively trained on synthetic image-text pairs crafted through this process. The effectiveness of our pipeline is validated through experimental results in both the in-domain and cross-domain settings, where it achieves state-of-the-art performance on well-known datasets such as MSCOCO, Flickr30k, and NoCaps.

Introduction

The realm of image captioning, which aims to craft informative textual descriptions for provided images, has witnessed remarkable progress. The crux of the challenge in image captioning hinges on comprehending the interplay between images and text, a dependency strongly rooted in the image-text pairs. Two approaches emerge to tackle this hurdle: one involves utilizing readily existing paired image-text data, while the other entails creating pairs from independent data sources.

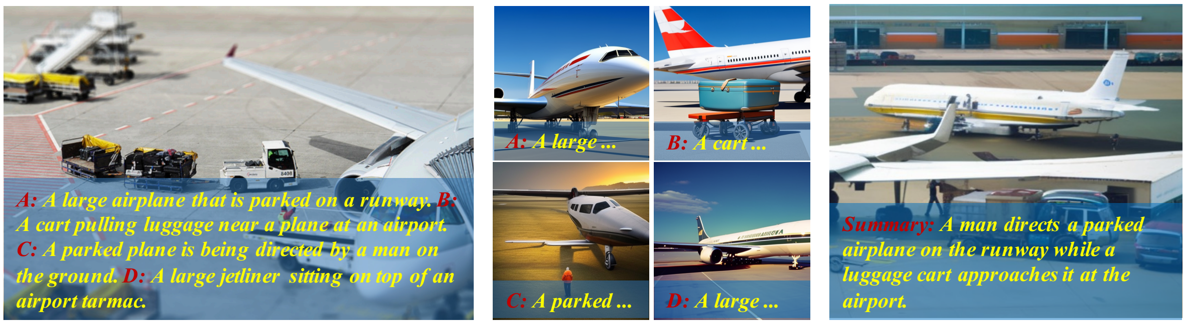
There are two primary sources of existing paired image-text data: human-annotated and web-crawled. Human-annotated data, employed in many studies (Dai and Lin 2017; Anderson et al. 2018), can lead to significant improvements. However, the small size of available data limits its scalability and restricts domain generality. Web-crawled data often suffer from low quality due to inaccurate correlations between images and text. As a result, models (Kang et al. 2023) trained on web-crawled data can exhibit limited performance in zero-shot settings. Moreover, large-scale crawling of image-text pairs possibly involves privacy and copyright issues.

In the absence of paired image-text data, unsupervised strategies often forge noisy image-text pairs from independent datasets for model bootstrapping (Feng et al. 2019) or domain alignment (Laina, Rupprecht, and Navab 2019), leveraging a pretrained object detector. Moreover, Meng et al. (2022) suggest establishing object-text pairs by associating objects with sentences, rather than seeking candidates within the image collection. These techniques operate under the assumption that a pretrained detector can consistently discern visual concepts, thus establishing connections between disparate images and text. However, this assumption might not hold universally.

Inspired by (He et al. 2023; Zhao et al. 2023), which effectively employ diffusion model-driven synthetic data in image classification and segmentation, we have noticed the progress text-to-image models have achieved in crafting high-quality images from textual descriptions. This opens the door to the use of diffusion models for constructing image-text pairs in the image captioning domain. In comparison to human-labeled and web-crawled datasets, synthetic data offer efficiency in cost and time, enable customization for specific domains, bootstrap generalization capability for zero-shot performance, and sidestep privacy issues linked to real-world data. Customization for specific domains, referring to the in-domain ability, involves generating data tailored to specific domains, such as particular objects, attributes, or scenarios. Generalization capabilities, pertaining to cross-domain capability, entail generating synthetic data encompassing a broader range of scenarios, not limited to a specific objective. However, there exists no prior work that specifically addresses the image captioning task solely using synthetic data generated via diffusion models.

[*] This work was performed while Feipeng Ma was an intern with WeChat, Tencent Inc.

[†] Corresponding authors: {zhyuey, sunxiaoyan}@ustc.edu.cn


A: A large airplane that is parked on a runway. B: A cart pulling luggage near a plane at an airport. C: A parked plane is being directed by a man on the ground. D: A large jetliner sitting on top of an airport tarmac.
A: A large ...
B: A cart ...
C: A parked ...
D: A large ...
Summary: A man directs a parked airplane on the runway while a luggage cart approaches it at the airport.

(1) The natural image and the caption with four sentences *A, B, C, D*.

(2) The uni-context images generated by captions *A, B, C, D*.

(3) The multi-context image generated by a summarized caption.

Figure 1: Examples of the natural image, the uni-context images, and the multi-context image.

The method to train an image captioning model using synthetic data involves two essential steps: (1) generating images along with captions through a diffusion model, utilizing established image captioning datasets. (2) subsequently training the image captioning model with the newly created synthetic dataset. However, a significant limitation arises when utilizing synthetic images, as they often lack the contextual depth necessary for ensuring precise image captioning accuracy. Our analysis highlights that synthetic images, originating from readily available basic captions, tend to exhibit constrained contexts, resulting in the omission of intricate and multifaceted scenes. These images are specifically referred to as “uni-context” images. In contrast, natural images inherently encompass a multi-contextual essence, portraying a diverse array of objects, arrangements, interactions, and encapsulating complex and elaborate scenes. In this context, we provide a collection of examples for a comprehensive comparison between uni-context and multi-context images, depicted in Figure 1. Notably, employing complex captions enables diffusion models to generate multi-context images, aligning with multi-faceted captions. Our focus for the image captioning task centers around the generation of multi-context captions to facilitate the synthesis of images with diverse contextual characteristics.

In this paper, we propose a pipeline for **I**mage **C**aptioning with Multi-Context **S**ynthetic **D**ata (**ICSD**). Our pipeline starts with a text corpus containing accessible simple captions from diverse sources such as datasets, web crawls, and generated content. Comprising two key stages, the pipeline initiates with the generation stage and moves on to the training stage. **The generation stage** begins with obtaining complex captions. To optimally harness the corpus, we suggest selecting simple captions that might collectively depict the same scene rather than focusing on a small subset of direct complex captions. This process, termed selection and summarization, not only taps into the corpus’ potential but also generates varied combinations of simple captions for diverse scenes. However, existing methods face challenges as they lack suitable metrics to determine if captions portray the same scene, and they need to be adaptable across different domains due to varied corpus sources. Leveraging the strengths of Large Language Models (LLMs) with their expansive knowledge and generalization abilities, we employ LLMs to execute selection and summarization tasks through provided instructions. Initially, we cluster captions based on text feature similarity, treating each caption as a query to construct LLM input candidates. The subsequent step instructs LLMs to pick captions from these clusters that could potentially form a complex scene. These chosen captions are then condensed into a single comprehensive caption. Subsequently, we generate multi-context images employing a generative model aided by these summarized captions. Moving into **the training stage**, our approach involves training models based on multi-context images derived from summarized sentences and the captions present in the corpus. Each multi-context image is associated with its corresponding selected sentences, offering multiple related descriptions for every multi-context image. The training of our model relies solely on this synthetic data.

Our main contributions are summarized as follows:
(1) Pioneering the utilization of synthetic data in image captioning through the synergistic application of diffusion models and LLMs, introducing a novel approach in this field.
(2) Addressing the deficiency in complexity found in synthetic images generated from basic captions by analyzing the multi-context nature of natural images. We introduce a multi-context data generation pipeline tailored for enhancing image captioning.
(3) Demonstrating the efficacy of our exclusively synthetic data-driven approach, we attain state-of-the-art performance in in-domain and cross-domain image captioning across three datasets: MSCOCO, Flickr30k and NoCaps.

Related Work

Supervised Image Captioning

Conventional image captioning methods treat the task as a form of translation (Vinyals et al. 2015; Karpathy and Fei-Fei 2015). These methods typically comprise a CNN-based encoder for image encoding and an RNN-based decoder for caption generation. Recent approaches (Wang, Xu, and Sun 2022; Barraco et al. 2022) adopt transformers architecture, yielding promising results. Additionally, research efforts integrate object (Anderson et al. 2018; Song et al. 2021), segmentation (Wu et al. 2022), gaze patterns (Alahmadi and Hahn 2022), and attributes (Fang et al. 2022) to enhance image captioning models. Due to limited annotated data, studies (Li et al. 2022; Hu et al. 2022; Wang et al. 2022) pretrain models on expansive web-crawled datasets, followed by fine-tuning on smaller human-annotated datasets. Although these methods benefit from pretraining, their performance still heavily depends on the fine-tuning phase, which entails human-annotated data.

Unsupervised Image Captioning

Unsupervised image captioning seeks to train captioning models without the need for human-annotated data. Prior work utilizes independent image sources and text corpora for training, often leveraging object detectors to establish an initial link between the two modalities. Feng et al. (2019) pioneer this field by introducing policy gradient to reward generated captions aligned with correct visual concepts. Subsequently, Laina, Rupprecht, and Navab (2019) propose a shared multi-modal space constructed through visual concepts to align images and text. Meng et al. (2022) suggest harvesting objects corresponding to given sentences instead of finding candidate images. Nonetheless, these approaches depend heavily on object detectors, overlooking object attributes and relationships, constrained by detector generalization. Recent text-only training methods focus on training text decoder to reconstruct text from CLIP text encoder-derived features. During inference, they align image features extracted by CLIP's image encoder with text features in the same space. Li et al. (2023) introduce a training-free mechanism using training text features to project visual embeddings into text embedding space at inference. Nukrai, Mokady, and Globerson (2022) and Gu, Clark, and Kembhavi (2023) propose noise injection training to reduce the modality gap during inference. However, these methods rely on CLIP's cross-modality capacity and struggle to transfer to new domains without fine-tuning CLIP.

Applications of Diffusion Models

Diffusion models excel in generative capacities, spanning image creation, video synthesis, and text generation (Ho, Jain, and Abbeel 2020; Dhariwal and Nichol 2021; Villegas et al. 2023; Chen, Zhang, and Hinton 2023). Conditional versions enhance control and produce premium outcomes, extending their usefulness, as seen in text-to-image generation with models like DALL-E 2, Imagen, and Stable Diffusion (Ramesh et al. 2022; Saharia et al. 2022; Rombach et al. 2022). Synthetic data from GLIDE demonstrated efficacy in image classification (Nichol et al. 2022; He et al. 2023), with further improvements achieved through ImageNet finetuning (Azizi et al. 2023). X-Paste (Zhao et al. 2023) leverages Stable Diffusion and CLIP to obtain synthetic images with accurate categories, which are transformed into instances for image segmentation. These tasks need high-quality synthetic images but with less focus on matching meaning exactly. Only the object linked to a single label should appear in the image, without considering the whole scene. The text-to-image diffusion model manages this basic requirement. Unlike these tasks, image captioning requires intricate scenes in synthetic images that can be described from various perspectives. Because diffusion models cannot generate multi-context images from simple sentences, creating suitable training data for image captioning becomes quite a challenge.

Method

Overview

Our pipeline, presented in Figure 2, comprises two stages: the generation stage and the training stage. The generation stage begins with a given text corpus which comes from multiple sources for in-domain or cross-domain settings. For in-domain setting, we utilize human-annotated text to generate in-domain data. For cross-domain setting, we employ web-crawled text or text generated from LLMs with rich knowledge, to produce large-scale cross-domain data. The generation stage consists of three steps: (1) Grouping of simple captions. In this step, for each simple caption acting as a query, we retrieve the most similar captions from the text corpus. These retrieved captions are then combined with the query caption to form a group. The captions in the same group possibly describe the same scene from diverse perspectives. (2) LLM-based selection and summarization. Providing the group of simple captions, which includes captions that potentially describe the same scene from diverse perspectives, we meticulously design prompt. These prompts guide LLMs to select simple captions that coherently align with a particular scene and summarize them into one sentence for image generation. (3) Finally, we employ stable diffusion to generate images with the summarized captions. In the training stage, we train the image captioning model solely on the synthetic multi-context image-text pairs obtained from the generation stage.

Generation Stage

Grouping of Simple Captions. Given a text corpus $T = \{t_1, t_2, ..., t_N\}$ with N captions, directly utilizing the whole text corpus as input is infeasible. This is because the input context length limitation in LLMs prevents the use of the full corpus. To address this issue, we propose to partition the text corpus into multiple groups of simple captions, with each group serving as a candidate set for selection and summarization. Considering the large size of the text corpus, we form a group for each simple caption by retrieving instead of clustering algorithms. Since captions describing the same scene often exhibit substantial semantic similarity with shared visual concepts, we employ CLIP to extract caption features and calculate the cosine similarity between each query caption and others within the corpus:

$$s_{ij} = \frac{f(t_i) \cdot f(t_j)}{||f(t_i)||\,||f(t_j)||} \tag{1}$$

where s_{ij} is the cosine similarity between $f(t_i)$ and $f(t_j)$, t_i is the query caption, t_j is another caption in corpus, $f(\cdot)$ represents the text encoder of CLIP. For query caption t_i, we retrieve the top k similar sentences to form a group G_i including t_i:

$$G_i = \{t_i, t_m, \ldots, t_n\} \tag{2}$$

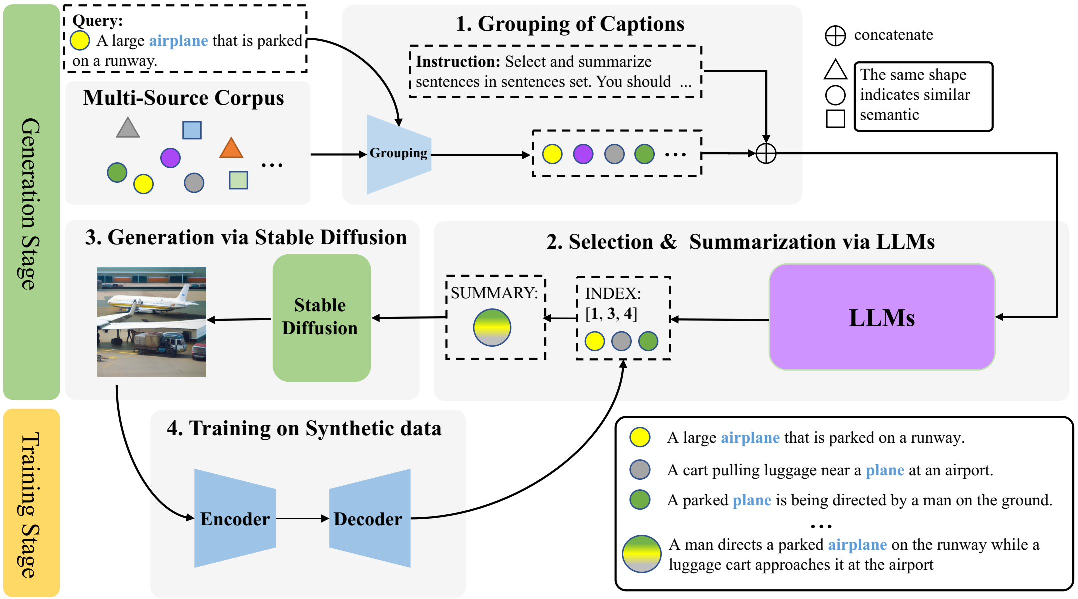

Generation Stage
Training Stage
Query:
A large airplane that is parked on a runway.
Multi-Source Corpus
1. Grouping of Captions
Instruction: Select and summarize sentences in sentences set. You should ...
Grouping
concatenate
The same shape indicates similar semantic
2. Selection & Summarization via LLMs
LLMs
INDEX: [1, 3, 4]
SUMMARY:
3. Generation via Stable Diffusion
Stable Diffusion
4. Training on Synthetic data
Encoder
Decoder
A large airplane that is parked on a runway.
A cart pulling luggage near a plane at an airport.
A parked plane is being directed by a man on the ground.
A man directs a parked airplane on the runway while a luggage cart approaches it at the airport

Figure 2: Overview of our proposed ICSD pipeline. The pipeline comprises two stages: the generation stage and the training stage. In the generation stage, we commence by performing the grouping of simple captions within the corpus. Next, LLMs are employed to select captions that depict the same scene from multiple perspectives, which are extracted from the obtained candidate sets. These selected captions are then condensed into a single sentence through summarization. These condensed sentences play a pivotal role in generating multi-context images using stable diffusion. Finally, in the training stage, we exclusively train the image captioning model on the synthetic multi-context image-text pairs.

where the cardinality of G_i is $k+1$, and $s_{ii} \geq s_{im} \geq ... \geq s_{in}$. This results in N groups, corresponding to the number of captions within the text corpus, with these groups containing overlapping captions. We use a greedy algorithm to minimize redundancy and ensure that a small number of groups cover all corpus captions. The algorithm works as follows: (1) initialize a set C identical to the corpus T. (2) repeatedly find and remove from C the group G_i with the most overlap with C, until C is empty. Finally, we can find groups that can cover the entire corpus.

Selection and Summarization via LLMs. Selection and summarization pose significant challenges for existing technology. Selection aims to select simple captions that are able to describe the same image from various perspectives. The challenges of selection are (1) the demand for common sense: the selection process should incorporate the knowledge of natural scenes to decide what kind of objects should appear together in the scene and the given descriptions should not conflict. (2) the lack of metrics: the current metrics of text similarity are not designed for our target; the similarity metrics can not identify the descriptions that depict the same image. Summarization aims to combine the selected simple captions into one complex caption for image generation. The challenge is that traditional text summarization approaches are ill-suited for our scenario since they aim to extract key information from long documents. And our summarization also demands strong generalization capabilities in open domains, since the corpus is so diverse. Fortunately, the potency of LLMs empowers us to tackle these intricate challenges. LLMs are pretrained on a large scale of data, showing advancement in common knowledge and generalization ability. The instruction-following ability of LLMs makes it possible for us to formulate the selection and summarization task through language. So we employ LLMs to tackle both tasks. We consider each group of captions as a candidate set for describing a specific image. We then formulate a prompt that enables us to accomplish both selection and summarization through LLMs. The prompt template is provided in Appendix C.

To avoid the hallucination problem of LLMs, we incorporate the chain of thought technique (Wei et al. 2022) into our design. We instruct the LLMs to first select sentences and subsequently summarize them into a single sentence. Additionally, we specifically prompt LLMs to provide the index of the chosen sentences, instead of generating these sentences anew, which may lead to hallucination problems.

Image Generation with Stable Diffusion. In this crucial step, we harness the power of stable diffusion to generate synthetic images based on the summarized sentences derived from the text corpus. We refrain from using any prompt engineering on the captions that are inputted to stable diffusion, aiming to minimize the influence of human intervention during large-scale generation. Consequently, we can acquire a substantial volume of multi-context images.

Training Stage

Architecture and data. We follow BLIP (Li et al. 2022) to adopt the encoder-decoder architecture (Vaswani et al. 2017). The encoder is initialized from the weights of ViT-B/32, while the decoder is initialized from the weights of BERT-base (Devlin et al. 2019). We exclusively utilize synthetic data for training, specifically focusing on multi-context image-text pairs.

Objective. We utilize synthetic data to train our model using Cross Entropy Loss:

$$\mathcal{L} = -\sum_{i=1}^{n} \log(P(y_i|y_{1:i-1}, v)) \tag{3}$$

where P denotes the probability distribution from the language decoder, y_i denotes the ground-truth word at time step i, $y_{1:i-1}$ refers to the prior words, n stands for the length of the ground-truth sentence, and v is the synthetic image.

Experiments

We conduct experiments in two settings: in-domain and cross-domain image captioning. In the in-domain setting, the training and test data are derived from the same dataset. In the cross-domain setting, the training and test data are sampled from different datasets, requiring the model to effectively generalize across diverse data sources.

Settings

Datasets. For in-domain image captioning, we utilize MSCOCO (Lin et al. 2014) and Flickr30k (Young et al. 2014) datasets. MSCOCO contains 123,287 images, each annotated with five captions. Following (Karpathy and Fei-Fei 2015), we split MSCOCO into 118,287 for training, 4,000 for validation, and 1,000 for testing. Flickr30k contains 31,783 images, with each image accompanied by five captions. Regarding cross-domain image captioning, we train our model on SS1M (Feng et al. 2019) dataset and evaluate its performance using MSCOCO and NoCaps (Agrawal et al. 2019). SS1M is a web-scraped text corpus and contains 2,322,628 image descriptions from Shutterstock using eighty object class names in MSCOCO as keywords. In line with DeCap (Li et al. 2023), we exclude sentences containing more than fifteen words on SS1M. We use the validation set of NoCaps to evaluate performance in three settings: in-domain, near-domain, and out-of-domain. For all datasets, we solely utilize their text during training and do not acquire any natural images.

To measure the quality of generated captions, we follow prior studies (Dai et al. 2017; Meng et al. 2022) and employ metrics such as BLEU (Papineni et al. 2002), METEOR (Banerjee and Lavie 2005), ROUGE (Lin 2004), and CIDEr-D (Vedantam, Lawrence Zitnick, and Parikh 2015).

Implementation Details. During the generation stage, we use different group sizes for various datasets: 30 for MSCOCO, 20 for Flickr30k, and 10 for SS1M. We employ the GPT-3.5-turbo model for selecting and summarizing captions via API access. For image generation, we utilize Stable Diffusion v1.4 at a 512×512 resolution with 20 sampling steps, and we speed up the diffusion model's sampling process using DPM-Solver (Lu et al. 2022). We train the model for 30 epochs using Adam (Kingma and Ba 2015) and a batch size of 36. The learning rate is 1e-5, and a warm-up strategy is applied during training. Additionally, the input synthetic images are resized to 384×384. For inference, we follow the BLIP to use beam search with a beam size of 3. All experiments are conducted using eight NVIDIA A100 GPUs.

Comparisons with State-of-the-art Models

In-domain Image Captioning. We perform in-domain image captioning on MSCOCO and Flickr30k datasets, comparing our ICSD with state-of-the-art unsupervised methods: Feng et al. (2019) and Laina, Rupprecht, and Navab (2019) train models on independent image and text data, using visual concepts to establish connections between images and text. ZeroCap (Tewel et al. 2022), Magic (Su et al. 2022), and ESPER-Style (Yu et al. 2022) incorporate GPT-2 (Radford et al. 2019) as the language decoder. CLIPRe (Su et al. 2022) is a CLIP-based method for retrieving captions. CapDec (Nukrai, Mokady, and Globerson 2022), DeCap (Li et al. 2023) and CLOSE (Gu, Clark, and Kembhavi 2023) conduct text-only training, leveraging the powerful cross-modal capability of CLIP.

The comparison results for MSCOCO and Flickr30k datasets are presented in Table 1. We generate 150,000 multi-context images for MSCOCO and 140,000 images for Flickr30k, with each synthetic image paired with 5 to 10 captions. Our method significantly outperforms other unsupervised approaches across the majority of metrics. In the B@4 metric, our ICSD surpasses previous state-of-the-art methods by 13.3% on MSCOCO and 26.0% on Flickr30k.

Cross-domain Image Captioning. The captions of MSCOCO and Flickr30k can naturally be grouped, as each image has at least five descriptions in these human-annotated datasets. To evaluate the effectiveness of our ICSD, we train the model on the web-crawled SS1M captions, which are not inherently grouped, and perform cross-domain image captioning on MSCOCO and NoCaps. We create 150,000 multi-context images for SS1M. Due to SS1M's large scale and the API call limitations of GPT-3.5-turbo, we additionally generate uni-context images for each caption. We compare our method with several other approaches in this experimental setting: (1) ZeroCap and ConZIC (Zeng et al. 2023) directly use pretrained vision-language models without fine-tuning; (2) CLIPRe and DeCap are trained on the large CC3M-text corpus (Changpinyo et al. 2021); (3) DeCap and Feng et al. (2019) also employ SS1M to train the models.

The results in Table 2 demonstrate the effectiveness of our method. When evaluating ICSD on MSCOCO, our method achieves obvious improvements across all metrics. Especially on BLEU and CIDEr metrics, improved from 8.9 to 13.6 and 50.6 to 54.2, respectively. This implies that the effectiveness of our method is not limited to in-domain image captioning, it remains efficient even when applied to a wide range of data collected from the web. In the NoCaps evaluation, our method performs well on in-domain and near-domain sets but not as strongly on the out-domain set. This discrepancy arises because SS1M is collected based on the objects of MSCOCO, which is not strictly designed for cross-domain setting, affecting its performance on the out-domain set of NoCaps. However, our method is able to address this limitation by utilizing LLMs to generate a cor-

Methods	Data		MSCOCO				Flickr30k			
	I.	T.	B@4	M	R	C	B@4	M	R	C
Feng et al. (2019)	✓	✓	18.6	17.9	43.1	54.9	-	-	-	-
Laina, Rupprecht, and Navab (2019)	✓	✓	19.3	20.2	45.0	61.8	-	-	-	-
ESPER-Style (Yu et al. 2022)	✓	✓	21.9	21.9	-	78.2	-	-	-	-
ZeroCap (Tewel et al. 2022)		✓	7.0	15.4	31.8	34.5	5.4	11.8	27.3	16.8
Magic (Su et al. 2022)		✓	12.9	17.4	39.9	49.3	6.4	13.1	31.6	20.4
CLIPRe (Su et al. 2022)		✓	4.9	11.4	29.0	13.6	5.2	11.6	27.6	10.0
CapDec (Nukrai, Mokady, and Globerson 2022)		✓	26.4	25.1	51.8	91.8	17.7	20.0	43.9	39.1
DeCap* (Li et al. 2023)		✓	25.3	25.2	51.1	92.9	20.0	**21.7**	46.2	49.3
CLOSE (Gu, Clark, and Kembhavi 2023)		✓	-	-	-	95.3	-	-	-	-
ICSD		✓	**29.9**	**25.4**	**52.7**	**96.6**	**25.2**	20.6	**46.7**	**54.3**

Table 1: The results of in-domain image captioning on MSCOCO and Flickr30k. "I." and "T." denote the need for external image data and text data, respectively. "*" means results reproduced using the provided code. B@4: BLEU@4; M: METEOR; R: ROUGE; C: CIDEr. Numbers in bold and underlined text represent the best and second-best results, respectively.

Methods	Dataset	MSCOCO				NoCaps val (CIDEr)			
		B@4	M	R	C	In	Near	Out	Overall
ZeroCap (Tewel et al. 2022)	-	2.6	11.5	-	14.6	-	-	-	-
ConZIC (Zeng et al. 2023)	-	1.3	11.5	-	12.8	-	-	-	-
CLIPRe (Su et al. 2022)	CC3M-text	4.6	13.3	-	25.6	23.3	26.8	36.5	28.2
DeCap (Li et al. 2023)	CC3M-text	8.8	16.0	-	42.1	34.8	37.7	**49.9**	39.7
DeCap (Li et al. 2023)	SS1M	8.9	17.5	-	50.6	41.9	41.7	46.2	**42.7**
ICSD	SS1M	**13.6**	**18.3**	**38.0**	**54.2**	**42.9**	**44.3**	35.6	**42.7**

Table 2: The results of cross-domain image captioning on MSCOCO and NoCaps.

pus containing diverse objects for generalization and applying our pipeline with this corpus. We report the results of this experiment in Appendix A.

Ablation Study

The effect of components of ICSD. Table 3 presents the results of an ablation study evaluating each component on MSCOCO. The *Baseline* indicates that we train the model using only uni-context image-text pairs. We then investigate three variations of selection and summarization: (1) selection without summarization (*Sel. w.o. Sum.*), where we employ LLMs to select captions for training but do not merge them into a single sentence for multi-context image generation. The captions selected from the same group are associated with the uni-context image, which is generated with query caption of the group. (2) summarization without selection (*Sum. w.o. Sel.*), in which LLMs directly summarize the top five similar captions in the group, potentially leading to errors since the most similar captions are not necessarily conflict-free. (3) both selection and summarization (*Sel. & Sum.*), representing our ICSD approach. To further emphasize the importance of selection in collecting a high-quality caption set for summarization, we conduct experiments using the ground-truth group (*w. GTG*) as a substitute for selection. MSCOCO is a human-annotated dataset where each image has at least five captions, inherently containing captions written for the same image by different annotators. In the *w. GTG* experiment, we use these captions, written for the same image, for summarization.

The results of *Baseline* exhibit limited performance across most metrics. In comparison with previous state-of-the-art methods, the Meteor and CIDEr scores are lower than DeCap by 0.5 and 2.1, respectively, while the ROUGE-L score is lower than CapDec by 0.4. The limited performance indicates that the uni-context images, generated by single captions, are insufficient for image captioning.

The *Sel. w.o. Sum.* approach results in improved performance, particularly in the BLEU and CIDEr metrics. This improvement can be attributed to the fact that *Sel. w.o. Sum.* employs LLMs to select captions that may describe the same scene and associate them with the image generated using the query caption. This process constructs pseudo multi-context pairs in which the uni-context image is associated with multiple captions, albeit with less accurate correlations.

The *Sum. w.o. Sel.* presents another type of pseudo multi-context image-text pair. By directly summarizing similar captions into a single caption through LLMs and generating multi-context images, this approach may introduce errors, as the similarity of captions cannot guarantee that they are conflict-free when describing the same scene. Consequently, the summarized caption may not be fully compatible with the original simple captions. However, the generated images are multi-context images, resulting in improved performance compared to *Sel. w.o. Sum.* method. This observation demonstrates the importance of multi-context images for image captioning, but the inaccurate correlations lead to inferior performance in this approach.

The performance of our proposed ICSD approach, which

CLIPCap:	*A group of people standing next to a red double decker bus.*	*A woman walking down a street next to a fire hydrant.*	*A dog wearing a tie and a dress shirt.*
CapDec:	*A red double decker bus parked in front of a man.*	*A young man riding a skateboard down the side of a metal rail.*	*A dog that is standing up wearing a tie.*
Ours:	*A group of people standing in front of a red double decker bus*	*A black and white photo of a bus driving down a street.*	*A brown and white dog wearing a tie.*

Figure 3: Comparisons of captions generated by CLIPCap (Mokady, Hertz, and Bermano 2021), CapDec (Nukrai, Mokady, and Globerson 2022), and our proposed ICSD, using exemplary images from MSCOCO dataset.

Method	B@4	M	R	C
Baseline	26.9	24.7	51.4	90.8
Sel. w.o. Sum.	29.5	24.7	52.3	94.2
Sum. w.o. Sel.	29.6	25.0	52.9	95.8
Sel. & Sum. (ICSD)	29.9	25.4	52.7	96.6
ICSD *w. GTG*	**30.2**	**25.7**	**53.0**	**97.3**

Table 3: The effect of components of ICSD on MSCOCO.

Number	B@4	M	R	C
10,000	28.0	23.8	51.1	87.9
50,000	29.3	24.7	52.3	93.5
100,000	29.4	25.0	52.3	95.1
150,000	29.9	**25.4**	**52.7**	**96.6**
200,000	**30.1**	25.1	**52.7**	96.5

Table 4: The impact of the number of multi-context images.

incorporates both *Sel. & Sum.*, is significantly improved. As the selection process can gather compatible captions for summarization, which is absent in the *Sum. w.o. Sel.* method, summarization can create multi-context captions for multi-context image generation, an aspect that is lacking in the *Sel. w.o. Sum.* approach. The results indicate that by combining selection and summarization, our ICSD can generate multi-context data that are highly beneficial for image captioning. To further explore our method, we use the ground-truth grouping of MSCOCO and summarize captions. This strategy notably enhances performance across all metrics, underscoring the importance of effective grouping.

The impact of the number of multi-context images. We conduct experiments using different numbers of multi-context images for training and evaluate the model on the test split of MSCOCO. Table 4 presents the results, where we increase the number of multi-context images from 10,000 to 200,000. The results demonstrate that incorporating more multi-context images during training can improve the performance of in-domain image captioning. Particularly, we observe significant gains in the B@4 and CIDEr metrics when increasing the number of multi-context images from 10,000 to 50,000. The best performance achieved thus far is obtained when using 150,000 multi-context images. However, expanding the number of multi-context images to 200,000 yields only marginal improvements, primarily due to the constrained diversity of the text corpus.

Visualization. Our model, trained on synthetic data, is capable of generating accurate captions for natural images. In Figure 3, we present examples of generated captions on the test split of MSCOCO, comparing our approach with CLIPCap and CapDec. The incorrect portions of the captions are highlighted in red, while the improvements made by our method are emphasized in green. For the first image, our method accurately describes the location and the number of people. Regarding the second and third images, our approach outperforms the others by capturing more detailed descriptions, such as the colors "black and white" and "brown and white".

In Appendix B, we further explore a range of alternative methods, including retrieving images instead of generating them, as well as creating a detailed caption from a given input caption, as opposed to merging multiple simple captions.

Conclusion

We observe that synthetic images generated from single captions lack the ability to be described from multiple perspectives, unlike real-world images. To address this issue, we propose a pipeline called ICSD that generates multi-context training data by combining LLMs and diffusion models for image captioning. The pipeline has two stages: generation and training. In the generation stage, we group captions in the corpus and select diverse perspectives using LLMs. These perspectives are summarized into a single sentence, which is then used to generate multi-context images through diffusion models. This results in high-quality synthetic multi-context image-text pairs where each image can be described from various perspectives. In the training stage, we train image captioning models using the synthetic data generated in the generation stage. Extensive experiments on in-domain and cross-domain image captioning demonstrate the effectiveness of our ICSD pipeline.

Acknowledgments

This work was in part supported by the National Natural Science Foundation of China under grants 62032006 and 62021001.

References

Agrawal, H.; Desai, K.; Wang, Y.; Chen, X.; Jain, R.; Johnson, M.; Batra, D.; Parikh, D.; Lee, S.; and Anderson, P. 2019. Nocaps: Novel object captioning at scale. In *ICCV*, 8948–8957.

Alahmadi, R.; and Hahn, J. 2022. Improve Image Captioning by Estimating the Gazing Patterns from the Caption. In *CVPR*, 1025–1034.

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 6077–6086.

Azizi, S.; Kornblith, S.; Saharia, C.; Norouzi, M.; and Fleet, D. J. 2023. Synthetic Data from Diffusion Models Improves ImageNet Classification. In *Transactions on Machine Learning Research*.

Banerjee, S.; and Lavie, A. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 65–72.

Barraco, M.; Cornia, M.; Cascianelli, S.; Baraldi, L.; and Cucchiara, R. 2022. The unreasonable effectiveness of CLIP features for image captioning: an experimental analysis. In *CVPR Workshops*, 4662–4670.

Changpinyo, S.; Sharma, P.; Ding, N.; and Soricut, R. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, 3558–3568.

Chen, T.; Zhang, R.; and Hinton, G. 2023. Analog bits: Generating discrete data using diffusion models with self-conditioning. In *ICLR*.

Dai, B.; Fidler, S.; Urtasun, R.; and Lin, D. 2017. Towards diverse and natural image descriptions via a conditional gan. In *ICCV*, 2970–2979.

Dai, B.; and Lin, D. 2017. Contrastive learning for image captioning. *NeurIPS*, 30.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 4171–4186.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *NeurIPS*, 34: 8780–8794.

Fang, Z.; Wang, J.; Hu, X.; Liang, L.; Gan, Z.; Wang, L.; Yang, Y.; and Liu, Z. 2022. Injecting semantic concepts into end-to-end image captioning. In *CVPR*, 18009–18019.

Feng, Y.; Ma, L.; Liu, W.; and Luo, J. 2019. Unsupervised image captioning. In *CVPR*, 4125–4134.

Gu, S.; Clark, C.; and Kembhavi, A. 2023. I Can't Believe There's No Images! Learning Visual Tasks Using only Language Supervision. In *ICCV*, 2672–2683.

He, R.; Sun, S.; Yu, X.; Xue, C.; Zhang, W.; Torr, P.; Bai, S.; and Qi, X. 2023. Is synthetic data from generative models ready for image recognition? In *ICLR*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *NeurIPS*, 33: 6840–6851.

Hu, X.; Gan, Z.; Wang, J.; Yang, Z.; Liu, Z.; Lu, Y.; and Wang, L. 2022. Scaling up vision-language pre-training for image captioning. In *CVPR*, 17980–17989.

Kang, W.; Mun, J.; Lee, S.; and Roh, B. 2023. Noise-aware learning from web-crawled image-text data for image captioning. In *ICCV*, 2942–2952.

Karpathy, A.; and Fei-Fei, L. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 3128–3137.

Kingma, D. P.; and Ba, J. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Laina, I.; Rupprecht, C.; and Navab, N. 2019. Towards unsupervised image captioning with shared multimodal embeddings. In *ICCV*, 7414–7424.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, volume 162, 12888–12900. PMLR.

Li, W.; Zhu, L.; Wen, L.; and Yang, Y. 2023. DeCap: Decoding CLIP Latents for Zero-Shot Captioning via Text-Only Training. In *ICLR*.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *ACL*, 74–81.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *ECCV*, 740–755. Springer.

Lu, C.; Zhou, Y.; Bao, F.; Chen, J.; Li, C.; and Zhu, J. 2022. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *NeurIPS*, 35: 5775–5787.

Meng, Z.; Yang, D.; Cao, X.; Shah, A.; and Lim, S.-N. 2022. Object-Centric Unsupervised Image Captioning. In *ECCV*, 219–235. Springer.

Mokady, R.; Hertz, A.; and Bermano, A. H. 2021. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*.

Nichol, A. Q.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; Mcgrew, B.; Sutskever, I.; and Chen, M. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *ICML*, volume 162, 16784–16804. PMLR.

Nukrai, D.; Mokady, R.; and Globerson, A. 2022. Text-Only Training for Image Captioning using Noise-Injected CLIP. In *EMNLP*, 4055–4063.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 311–318.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language Models are Unsupervised Multitask Learners. *OpenAI Blog*, 1(8): 9.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*, 10684–10695.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *NeurIPS*, 35: 36479–36494.

Song, Z.; Zhou, X.; Dong, L.; Tan, J.; and Guo, L. 2021. Direction relation transformer for image captioning. In *ACM MM*, 5056–5064.

Su, Y.; Lan, T.; Liu, Y.; Liu, F.; Yogatama, D.; Wang, Y.; Kong, L.; and Collier, N. 2022. Language models can see: plugging visual controls in text generation. *arXiv preprint arXiv:2205.02655*.

Tewel, Y.; Shalev, Y.; Schwartz, I.; and Wolf, L. 2022. Zerocap: Zero-shot image-to-text generation for visual-semantic arithmetic. In *CVPR*, 17918–17928.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *NeurIPS*, 30.

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *CVPR*, 4566–4575.

Villegas, R.; Babaeizadeh, M.; Kindermans, P.-J.; Moraldo, H.; Zhang, H.; Saffar, M. T.; Castro, S.; Kunze, J.; and Erhan, D. 2023. Phenaki: Variable Length Video Generation from Open Domain Textual Descriptions. In *ICLR*.

Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and tell: A neural image caption generator. In *CVPR*.

Wang, Y.; Xu, J.; and Sun, Y. 2022. End-to-end transformer based model for image captioning. In *AAAI*, volume 36, 2585–2594.

Wang, Z.; Yu, J.; Yu, A. W.; Dai, Z.; Tsvetkov, Y.; and Cao, Y. 2022. SimVLM: Simple Visual Language Model Pretraining with Weak Supervision. In *ICLR*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*, 35: 24824–24837.

Wu, M.; Zhang, X.; Sun, X.; Zhou, Y.; Chen, C.; Gu, J.; Sun, X.; and Ji, R. 2022. Difnet: Boosting visual information flow for image captioning. In *CVPR*, 18020–18029.

Young, P.; Lai, A.; Hodosh, M.; and Hockenmaier, J. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2: 67–78.

Yu, Y.; Chung, J.; Yun, H.; Hessel, J.; Park, J.; Lu, X.; Ammanabrolu, P.; Zellers, R.; Bras, R. L.; Kim, G.; et al. 2022. Multimodal knowledge alignment with reinforcement learning. *arXiv preprint arXiv:2205.12630*.

Zeng, Z.; Zhang, H.; Wang, Z.; Lu, R.; Wang, D.; and Chen, B. 2023. ConZIC: Controllable Zero-shot Image Captioning by Sampling-Based Polishing. In *CVPR*, 23465–23476.

Zhao, H.; Sheng, D.; Bao, J.; Chen, D.; Chen, D.; Wen, F.; Yuan, L.; Liu, C.; Zhou, W.; Chu, Q.; et al. 2023. X-Paste: Revisit Copy-Paste at Scale with CLIP and StableDiffusion. In *ICML*, volume 202, 42098–42109. PMLR.