IConZIC : Image-Conditioned Zero-shot Image Captioning by Vision-Language Pre-Trained Model

Korea University COSE461 Final Project

******* Jeon
Department of Computer Science
Team **
****

Jiwon Jeong
Department of Data Science
Team **
****

******* Jeon
Department of Computer Science
Team **
****

******** Ko
Department of Computer Science
Team **
****

Abstract

The field of image captioning, which combines computer vision and natural language processing, has witnessed extensive research efforts. However, the exploration of zero-shot learning for image captioning remains relatively underexplored. Zero-shot image captioning research began with ZeroCap and was followed by ConZIC, which is currently considered state-of-the-art (SOTA). However, ConZIC still has certain limitations. To address these limitations and advance the field of zero-shot image captioning, we propose **IConZIC** (**I**mage-**Con**ditioned **Z**eroshot **I**mage **C**aptioning). IConZIC overcomes the initialization issue of ConZIC by leveraging the Vision-Language Pre-training(VLP) encoder, resulting in faster and more accurate caption generation. Our code is available at https://github.com/JeonSeongHu/IConZIC

1 Introduction

Image captioning (IC) is the task of automatically generating descriptive and coherent captions for images. Traditionally, IC methods have relied on paired and diverse image-text data for training, such as MSCOCO,[1] which has limitations when it comes to captioning outlier images that deviate from the training dataset distribution. However, recent developments have introduced zero-shot image captioning (ZIC) approaches, which overcome these limitations. ZIC methods allow for caption generation without the need for pre-training on specific image-caption pairs, making them more suitable for real-world applications.[2]

ConZIC is a ZIC model that uses pre-trained masked language models (MLMs) to generate captions. It refines the captions through an iterative process, combining predicted probabilities from Gibbs-BERT, CLIP scores, and control signals[2, 3, 4]. Each word in the caption is selected and masked, and the weighted sum of predicted probabilities guides the selection of the next word. This approach aims to produce high-quality captions in zero-shot image captioning tasks.

However, ConZIC has some limitations. Firstly, the use of random masks during candidate generation leads to an initialization problem, where the initial candidates may not be related to the image. As a result, the model needs to consider a larger number of candidates, leading to longer generation times.

Secondly, ConZIC is sensitive to hyperparameter setting. A lot of hyperparameters such as α, β, k, n and more should be controlled carefully. Improper settings can lead model to generate just sequence of random words, or sentences ignoring visual content. Finding the right hyperparameter configuration becomes a complex task in practice.

To address the challenges, we propose a novel approach called IConZIC (Image-Conditioned Zero-shot Image Captioning). Our approach incorporates image-conditioned masked language modeling (ICMLM) within VLP[5, 6, 7, 8]. It predicts masked language tokens from both text and visual tokens. While most ICMLMs cannot generate captions using only their encoder structure, our hypothesis is that by incorporating the "image-conditioned" aspect of ICMLM, we can generate appropriate captions for given images. In IConZIC, we use Gibbs-ViLT instead of Gibbs-BERT, which allows us to overcome the initialization problem in ConZIC by considering the visual tokens from the start. Furthermore, IConZIC provides improved flexibility in hyperparameter settings, with CLIP scores playing a crucial role in controlling diversity and enhancing control over the caption generation process.

2 Related Work

Zero-shot Image Captioning. Zero-shot capability has garnered significant attention in the field of deep learning as it enables machines to perform tasks without curated training data. One notable application of zero-shot learning is zero-shot image captioning, which allows machines to generate captions for images without relying on supervised training. One existing method in this domain is ZeroCap, which abandons supervised training and employs a sequential search strategy using large-scale pre-trained models to generate captions[9].

While ZeroCap has demonstrated effectiveness, its autoregressive generation approach and gradient-directed searching mechanism pose limitations in terms of caption diversity and inference speed, respectively[2, 9]. ConZIC overcomes these challenges using a novel sampling-based non-autoregressive language model named Gibbs-BERT, which enables the generation and continuous refinement of each word in the caption[2].

ConZIC offers more flexibility, faster generation speed, and higher diversity scores, it may still face challenges in accurately captioning images that significantly deviate from the training distribution[2]. Our insight is to change Gibbs-BERT[4] into ViLT[10] to reduce computational cost by minimizing candidates with less impact on performance.

VLP and Image-Conditioned Masked Language Modeling. VLP focuses on learning the semantic correspondence between different modalities, such as image-text and video-text pre-training, to associate words with their visual representations. VLP can be used for various downstream tasks and is trained with different objectives[11]. In order to use image captioning as a downstream task, a VLP model trained with a captioning objective is needed, or if the model does not have the captioning capability, an additional decoder architecture needs to be added and fine-tuned. [6, 7, 8]

ICMLM(Image-conditioned Masked Language Modeling)[5] is similar to MLM(Masked Language Modeling) but differs in that it uses images as an additional condition. ViLT[10] is a popular VLP model trained with the ICMLM objective. It simplified the image feature extraction process by using linear embeddings of image patches. This allows ViLT to achieve speeds more than 10 times faster than existing models.

Generally, VLP models trained with the ICMLMs objective cannot do captioning on their own and require additional fine-tuning. However, just as it is known that MLMs can generate sentences using Gibbs sampling[4], it is expected that ICMLMs can also generate image-conditioned text using the same method.

3 Approach

3.1 Framework of IConZIC

The objective of image captioning is to generate a suitable caption $x_{<1,n>}$, composed of n words, that maximizes the conditional probability $p(x_{<1,n>}|I)$ given an image I. Similar to ConZIC, we employ Gibbs sampling. However, a key distinction is that we utilize ICMLMs.

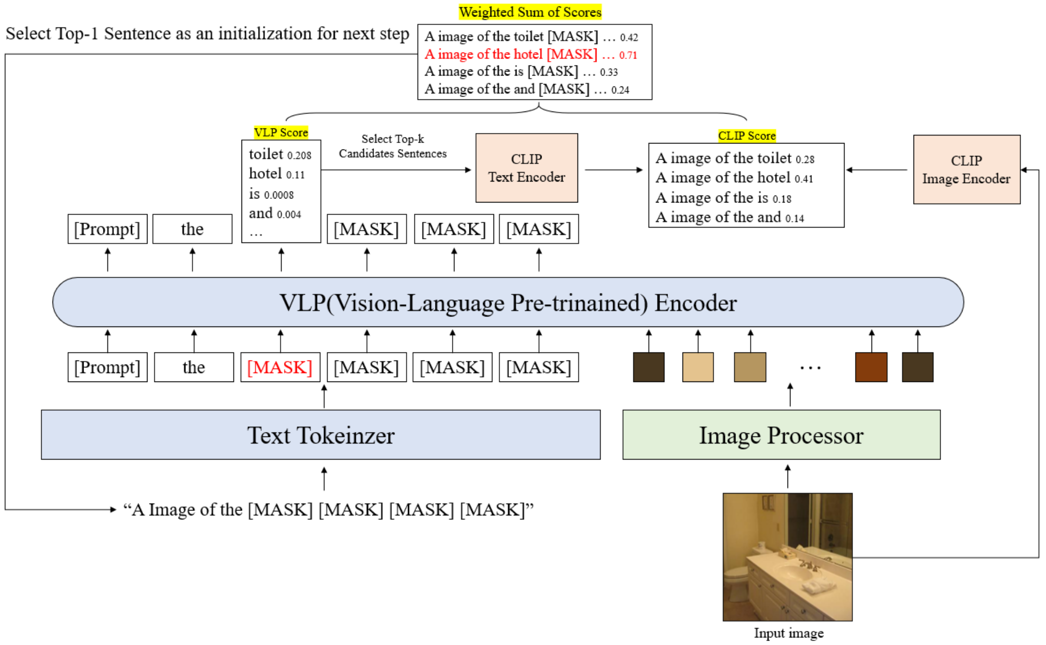

Weighted Sum of Scores
Select Top-1 Sentence as an initialization for next step
A image of the toilet [MASK] … 0.42
A image of the hotel [MASK] … 0.71
A image of the is [MASK] … 0.33
A image of the and [MASK] … 0.24
VLP Score
toilet 0.208
hotel 0.11
is 0.0008
and 0.004
…
Select Top-k
Candidates Sentences
CLIP
Text Encoder
CLIP Score
A image of the toilet 0.28
A image of the hotel 0.41
A image of the is 0.18
A image of the and 0.14
CLIP
Image Encoder
[Prompt]
the
[MASK]
[MASK]
[MASK]
VLP(Vision-Language Pre-trinained) Encoder
[Prompt]
the
[MASK]
[MASK]
[MASK]
[MASK]
…
Text Tokeinzer
Image Processor
"A Image of the [MASK] [MASK] [MASK] [MASK]"
Input image

Figure 1: An overview of our approach.

In ConZIC, the log data likelihood is derived from a Bayes Rule as $\log p(x_{<1,n>}|I) = \alpha \log p(I|x_{<1,n>}) + \beta \log p(x_{<1,n>})$[2]. The former term is computed using CLIP score, and the latter term is computed using BERT score. The data distribution is then modeled based on the weighted sum of these scores.

In contrast, ICMLMs predict masked words using both visual tokens and other words. By applying Gibbs sampling to this approach, we can effectively model the conditional probability $p(x_{<1,n>}|I)$. Theoretically, it is possible to generate captions based on the top-1 candidate predicted by the model alone. This method offers improved efficiency compared to ConZIC, which required selecting a minimum of 200 candidates for better performance.

However, subsequent experiments have shown that considering CLIP score for generating captions with k candidates yields significantly better performance. Mathematically, this can be seen as modeling $\log p(x_{<1,n>}|I)$ as $\alpha \log p(I|x_{<1,n>}) + \beta \log p(x_{<1,n>}|I)$.

As $\alpha \log p(I|x_{<1,n>}) + \beta \log p(x_{<1,n>}|I)$ can be approximated as $(\alpha + \beta) \log p(I|x_{<1,n>}) + \beta \log p(x_{<1,n>})$ with Bayes Rule, we can view it as modeling $\log p(x_{<1,n>}|I)$ as well.

3.2 Iterative sampling-based Modeling for $p(x_{<1,n>}|I)$

Gibbs sampling is a Markov chain Monte Carlo (MCMC) algorithm used in Gibbs-BERT[4]. It samples from the distribution $p(x_{<1,n>})$ by iteratively sampling individual words x_i from $p(x_i|x_{-i})$. However, this approach has a limitation: the distribution $p(x_i|x_{-i})$ does not have image information. So, in ConZIC, Gibbs-BERT does not incorporate images in its process. It serves as an auxiliary model to enhance the fluency of generated captions produced by CLIP, rather than generating a random sequence of image-related words.

At the initialization stage, especially when all words are masked, we cannot expect good candidates. As the iterations progress, the generated sentence incorporates image information, enabling the sampling of better candidates. Nevertheless, the generation of accurate captions with a small number of candidates is not feasible because of the initialization phase. This leads to the need for increasing the number of candidates proposed by Gibbs-BERT, which in turn increases computational complexity and generation time due to the calculation of CLIP scores for multiple sentences.

We introduce Gibbs-ViLT, pre-trained with the ICMLM objective. It directly samples each word x_i from the distribution $p(x_i|x_{-i}, I)$ [5] using Gibbs-sampling. It retains all the advantages of Gibbs-

BERT's approach. The flexible ordering of sampling allows for the generation of diverse captions, addressing issues like collapsed mode [12] and lack of diversity often encountered in traditional autoregressive sampling methods.

Moreover, the distribution $p(x_i|x_{-i}, I)$ is conditioned on the image information. Therefore, with a smaller value of k (the number of candidates considered), there is a higher probability of obtaining suitable candidates that align well with the image. Additionally, even during the initialization stage, the distribution provides image-informed candidates, resulting in improved generation performance.

Each word is initially initialized as [MASK]. Each iteration consists of a polishing step corresponding to the number of words in the sentence. In the initialization phase, the words are assigned to indices for each step, and according to the predefined algorithm, the most suitable word x_i is sampled in the order of the assigned indices. The initial sentence for each iteration and step will utilize the output sentence from the previous iteration and step.

3.3 Enhancing Caption with CLIP-Guided Scoring

The CLIP-guided candidate selection method used in ConZIC is incorporated directly into Gibbs-ViLT. For a [MASK] token in an input sentence, the top-k word candidates are selected using ViLT's output. Each candidate word replaces the [MASK] token, resulting in the generation of k new sentences. CLIP scores between the generated sentences and the input image are then calculated. These scores are combined with the ViLT probabilities using a weighted sum, and the top-1 word is selected as a substitute for the [MASK] token. This allowed us to enhance the generation process by prioritizing more relevant and contextually appropriate phrases that align with the visual content of the input image.

3.4 Overall Algorithm

Algorithm 1 Our overall algorithms

```
Data: initial caption: $\mathbf{x}^0_{<1,n>} = (x^0_1, ..., x^0_n)$; iterations=T, candidates=K;
position sequence P = Shuffle([1, ..., n]);
Result: The final Caption: $\mathbf{x}^T_{<1,n>} = (x^T_1, ..., x^T_n)$;
for iteration $t \in [1, ..., T]$ do
    state: $\mathbf{x}^{t-1}_{<1,n>} = (x^{t-1}_1, ..., x^{t-1}_n)$
    for position $i \in P$ do
        1. Replace $x_i^{t=1}$ with [MASK];
        2. Predict the word distribution over vocabulary by Gibbs-ViLT: $p(x_i|\mathbf{x}^{t-1}_{-i}, I)$;
        3. Select top-$K$ candidate words $\{x^t_{ik}\}_{k=1}$ by $p(x_i|\mathbf{x}^{t-1}_{-i}, I)$, whose probability $p_k^{ViLT}$;
        4. Get $K$ candidate sentences $\{s_k\}^K_{k=1} : (x^{t-1}_1, ..., x^{t-1}_{i-1}, x^t_{ik}, x^{t-1}_{i+1}, ..., x^{t-1}_n)^K_{k=1}$;
        5. Compute the CLIP score for $\{s_k\}^K_{k=1}$
        6. Select $x^t_i$ with largest probability by $\alpha p_k^{ViLT} + \beta p^{CLIP}$;
        7. Replace $x^{t-1}_i$ with $x^t_i$;
    end
    state: $\mathbf{x}^{t-1}_{<1,n>} = (x^{t-1}_1, ..., x^{t-1}_n)$
end
```

4 Experiments

4.1 Dataset

MSCOCO caption[1] : MSCOCO Caption dataset is a huge dataset that consists of 82,783 training images with 5 captions each, 40,504 validation images with 5 captions each, and 40,775 testing images with 379,249 captions. Due to limitations in time and resources, we were unable to generate captions for the entire image dataset. Instead, we conducted experiments by randomly selecting either about 1000 or 256 images from the validation set.

4.2 Evaluation method

Accuracy. Image Captioning models are evaluated using two main types of metrics: supervised metrics and unsupervised metrics. In our case, we will compare our model with ConZIC, which achieved SOTA performance in zero-shot image captioning. Both methods utilize zero-shot methods that do not involve any training, making it meaningless to evaluate the models using supervised metrics. Therefore, we will use an unsupervised metric called CLIP-S (CLIPScore) [3] for evaluation. CLIP-S is a reference-free metric that measures the cosine similarity between image embeddings and text embeddings, and it is widely used for evaluating zero-shot image captioning. For our evaluation, we will calculate accuracy based on the average score of the caption with the highest CLIP-S value among all iterations.

Diversity. Another important criterion in Zero-shot Image Captioning is diversity. Unlike traditional supervised captioning, which is limited by the word distribution of the training dataset, zero-shot captioning aims to generate sentences with various words. In our evaluation, we will measure diversity by considering the number of unique words used in the best captions.

Time. Generation time is an important factor that determines the feasibility of ZIC in real-world applications. The average time taken to generate a caption per image is measured in seconds, as an evaluation criterion.

4.3 Experimental details

We conducted two main experiments:

- Performance comparison between ConZIC and our model for α
- Performance comparison between ConZIC and our model for k

In the first experiment covering a total of 256 images using a maximum caption length of 7 words; we adjusted the value α. Following ConZIC's recommendation where their optimal value was established as being at 0.02, we tested other ranges across 0.2, 0.002, and 2. During this same experiment, we set the value of k to 50 for our model, which differs from ConZIC which uses 200.

In the second experiment, captions were generated for 1000 images with a maximum length of 12. We varied the value of k as 10, 20, 50, 100, and 200 while keeping α fixed at 0.02.

For both experiments, the number of iterations was fixed at 15. We followed the "Shuffle" method to take advantage of flexible-order sampling, which is one of the strengths of Gibbs sampling. In the Shuffle method, step indices are randomly assigned at the start of caption generation, and Gibbs sampling is performed accordingly. All other hyperparameters were kept the same as those used in the implementation of ConZIC.

All experiments were conducted using pre-trained models without any fine-tuning. The CLIP-ViT-B/32 model was used for calculating the CLIP Score, while the ViLT-B32-MLM-ITM model was used for calculating the ViLT Score. The experiments were performed on a single RTX 3050 Laptop GPU and Google Colab.

4.4 Quantitative Results

ConZIC

Parameter	Accuracy	Diversity
	CLIP-S	Vocab
α = 2.0	0.61	457
α = 0.2	0.81	969
α = 0.02	0.94	**1712**
α = 0.002	0.95	**2249**

Ours

Parameter	Accuracy	Diversity
	CLIP-S	Vocab
α = 2.0	**0.75**	**495**
α = 0.2	**0.89**	**977**
α = 0.02	**0.96**	1461
α = 0.002	**0.96**	1653

Table 1: Performance comparison of our method and ConZIC with different α values.

ConZIC

Parameter	Accuracy	Diversity	Time
	CLIP-S	Vocab	s/image
k = 10	0.85	**2894**	**6.8**
k = 20	0.92	**4756**	**7.6**
k = 50	0.98	**6545**	**11.6**
k = 100	1.02	**8006**	**20.4**
k = 200	1.03	**9434**	**36.0**

Ours

Parameter	Accuracy	Diversity	Time
	CLIP-S	Vocab	s/image
k = 10	**0.98**	2363	8.1
k = 20	**1.01**	3391	9.2
k = 50	**1.04**	4259	13.6
k = 100	**1.05**	5056	21.4
k = 200	**1.05**	5921	39.4

Table 2: Performance comparison of our method and ConZIC with different k values.

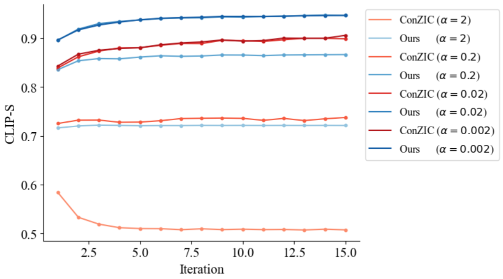


Figure 2: Comparison of CLIP-s for α

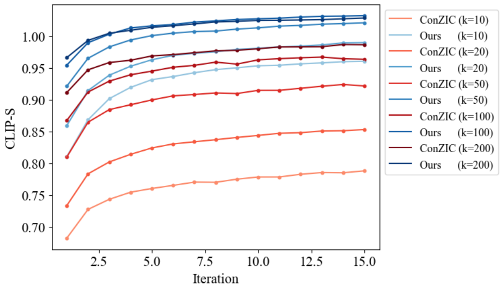


Figure 3: Comparison of CLIP-S for k

The quantitative results are presented in Tables 1, 2 and Figure 2, 3. Accuracy is reported as the average of the best captions obtained during 15 iterations, and the figures illustrate the average accuracy across iterations.

Performance comparison based on α. Our model outperformed ConZIC in terms of accuracy for the best captions. Setting a very high value for α (e.g., 2) resulted in a significantly larger BERT Score or ViLT Score compared to the CLIP Score, leading to the selection of the top-1 candidate without considering the CLIP Score. When α was set to 2, ConZIC initially achieved the highest accuracy in the first iteration but then converged to a lower value of around 0.5. In contrast, our model converged at around 0.7. However, this is still an unsatisfactory result, indicating that Gibbs-ViLT alone cannot produce satisfactory captions. More detailed results can be found in the Ablation Study in the Appendix. On the other hand, diversity showed a decreasing trend as the CLIP score decreased compared to ConZIC. This can be attributed not only to the lower value of k in our model but also to the fundamental differences between Gibbs-BERT and Gibbs-ViLT, as demonstrated in the second experiment results (related to the k value).

Performance comparison based on k. Our model exhibited superior accuracy across all k values.In terms of generation time, due to the nature of ViLT, which requires additional image tokens for embedding, our model appeared relatively slower for the same k value. However, our model achieved accuracy comparable to ConZIC's best result even with smaller k values. When our model used $k = 20$ and ConZIC used $k = 200$, similar accuracy was achieved, but our model was approximately four times faster in terms of generation time. Our model maintained high accuracy regardless of the value of k, and the accuracy almost converged when k exceeded 50; in fact, the results for $k = 200$ were not better than those for $k = 100$. This demonstrates that ViLT performs better than BERT in image-conditioned sampling. Our model showed lower diversity performance once again. This can be attributed to the difference in the pre-training process between BERT and ViLT. While BERT is pre-trained on a vast amount of text data available on the internet, ViLT is trained only on caption data created by human annotators for images. Therefore, ViLT inherently has a limited diversity of recommended candidates.

4.5 Qualitative Results

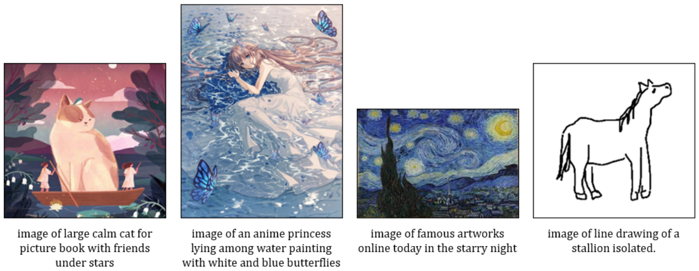


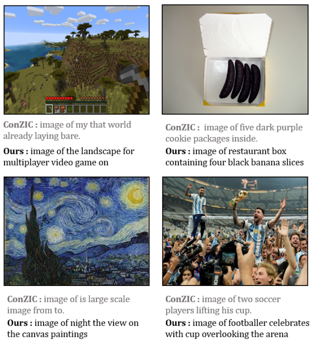
image of large calm cat for picture book with friends under stars

image of an anime princess lying among water painting with white and blue butterflies

image of famous artworks online today in the starry night

image of line drawing of a stallion isolated.

Figure 4: Qualitative Results



ConZIC : image of my that world already laying bare.

Ours : image of the landscape for multiplayer video game on

ConZIC : image of five dark purple cookie packages inside.

Ours : image of restaurant box containing four black banana slices

ConZIC : image of is large scale image from to.

Ours : image of night the view on the canvas paintings

ConZIC : image of two soccer players lifting his cup.

Ours : image of footballer celebrates with cup overlooking the arena

Figure 5: Comparison with ConZIC

As shown in Figure 4, we generated the captions using the example images in the ConZIC paper. The results demonstrate that our model can effectively caption images even when they deviate from the real-world image data distribution that ViLT is pre-trained on, such as MSCOCO [1] and Visual Genome [13].

Figure 5 compares the captions generated by ConZIC and our model when the number of candidates, k, is set to 20. ConZIC, due to the limited number of candidates, either fail to capture the elements of the image correctly (top right) or generates completely random captions (top left). In contrast, IConZIC consistently generates high-quality captions without encountering such issues.

5 Analysis

5.1 Ablation Study

Comparison of captions with and without CLIP Scores. We compared the captions with and without CLIP score in the same image. The results are shown in Figure 6. By combining the VLP encoder with CLIP, our model demonstrated an improved ability to generate that is not only more fluent but also retains a higher level of image semantics. The integration of CLIP scores facilitated a better understanding of the visual context, enabling the generation of more vivid and informative captions.

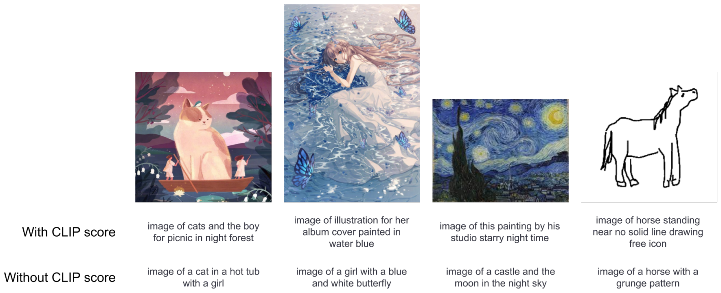
5.2 Contribution

Resolve the issues associated with the random initialization of ConZIC. Despite utilizing a smaller number of candidates, our approach achieved high accuracy, enabling us to reduce the value of k and significantly decrease generation time. Furthermore, our method exhibited robustness to hyperparameter settings, consistently surpassing the performance of ConZIC under various configurations. Notably, even with extreme values of hyperparameters, our approach demonstrated substantial performance improvements.

Propose that ICMLMs can not only "see", but also "speak". Similar to how Gibbs sampling allows MLMs to leverage their latent text generation capabilities, ICMLMs also utilize Gibbs sampling to harness their ability to generate image-conditioned text, specifically captions.

5.3 Limitation

Zero-shot issues : ViLT, trained on curated text-image pair datasets, may not qualify as a complete zero-shot captioning approach. Unlike models which are ZeroCap and ConZIC that utilize pre-trained models like CLIP, ViLT's training on curated datasets raises doubts about its zero-shot captioning capabilities. The reason why ZeroCap and ConZIC could claim their methods as zero-shot captioning was that using only CLIP[2, 9], which was trained on a large amount of web data, did not demonstrate



With CLIP score	image of cats and the boy for picnic in night forest	image of illustration for her album cover painted in water blue	image of this painting by his studio starry night time	image of horse standing near no solid line drawing free icon
Without CLIP score	image of a cat in a hot tub with a girl	image of a girl with a blue and white butterfly	image of a castle and the moon in the night sky	image of a horse with a grunge pattern

Figure 6: Caption Generation: CLIP Scores vs. No CLIP Scores

good performance in captioning. Similarly, using ViLT alone for caption generation with Gibbs sampling may not yield satisfactory results, but the difference lies in ViLT being trained on curated datasets rather than CLIP[10].

Generation issues : Setting excessively long sentence lengths can lead to repetitive words and reduced sentence fluency. This is a limitation of Gibbs sampling-based generation, which exhibits lower fluency compared to generative language models.[4] Subjective pronouns like "I," "we," or "you" are occasionally used, stemming from CLIP not being specifically designed for image captioning. Another issue in the generation is that there are insufficient data regarding the effectiveness of utilizing proper nouns or text recognition within images for generating captions. Additionally, it is necessary to verify whether the limitations observed in ZeroCap, such as national bias, have been addressed.

Evaluation issues : The CLIP-S metric itself presents challenges. While it is effective in describing image-text alignment[3], solely relying on CLIP-S for evaluating caption fluency is problematic. A more refined metric is required, as higher CLIP-S do not necessarily indicate better caption quality. Training on curated datasets, mentioned in previous ZIC approaches, may result in reduced diversity of words in generated captions, as observed in our experimental results. However, additional research is needed to investigate whether the reduction in the vocabulary size extends to caption quality and diversity.

6 Conclusion

We introduced IConZIC, an image captioning model based on the ViLT which is an image-conditioned language model. We aimed to address the limitations of existing zero-shot image captioning models by incorporating a slight influence of image features during caption initialization, while still preserving the essence of zero-shot captioning. By utilizing VLP trained with ICMLM instead of BERT, our model achieves state-of-the-art(SOTA) with higher accuracy and faster generation compared to ConZIC, which was a cutting-edge method in zero-shot image captioning. The use of ViLT enables the generation of words that are more relevant to images, resulting in higher accuracy even with a reduced value of k. At the same value of k and α, ViLT has a smaller vocabulary size of generated captions compared to BERT. However, the evaluation of vocabulary size of generated captions alone does not provide a comprehensive assessment of whether ViLT generates a smaller number of image-related words compared to BERT or reduces the generation of unrelated words. It would be future work to evaluate under various metrics and compare each model to find the best parameters.

References

[1] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *CoRR*, abs/1504.00325, 2015.

[2] Zequn Zeng, Hao Zhang, Zhengjue Wang, Ruiying Lu, Dongsheng Wang, and Bo Chen. Conzic: Controllable zero-shot image captioning by sampling-based polishing, 2023.

[3] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning, 2022.

[4] Alex Wang and Kyunghyun Cho. BERT has a mouth, and it must speak: BERT as a markov random field language model. *CoRR*, abs/1902.04094, 2019.

[5] Mert Bulent Sariyildiz, Julien Perez, and Diane Larlus. Learning visual representations with caption annotations, 2020.

[6] Haiyang Xu, Ming Yan, Chenliang Li, Bin Bi, Songfang Huang, Wenming Xiao, and Fei Huang. E2E-VLP: end-to-end vision-language pre-training enhanced by visual learning. *CoRR*, abs/2106.01804, 2021.

[7] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *CoRR*, abs/2108.10904, 2021.

[8] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and VQA. *CoRR*, abs/1909.11059, 2019.

[9] Yoad Tewel, Yoav Shalev, Idan Schwartz, and Lior Wolf. Zero-shot image-to-text generation for visual-semantic arithmetic. *CoRR*, abs/2111.14447, 2021.

[10] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision, 2021.

[11] Feilong Chen, Duzhen Zhang, Minglun Han, Xiuyi Chen, Jing Shi, Shuang Xu, and Bo Xu. Vlp: A survey on vision-language pre-training, 2022.

[12] Qi Chen, Chaorui Deng, and Qi Wu. Learning distinct and representative modes for image captioning, 2022.

[13] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. Visual genome: Connecting language and vision using crowdsourced dense image annotations, 2016.