Part of speech masking evaluation with vision-language model

by

Pasit Tiwawongrut

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of
Master of Science in Data Science and Artificial Intelligence

Examination Committee: Dr. Chaklam Silpasuwanchai
Dr. Mongkol Ekpanyapong
Dr.

Nationality: Thai
Previous Degree: Bachelor of Computer Engineering
Khon Kaen University
Thailand

Scholarship Donor: Asian Institute of Technology

Asian Institute of Technology
School of Engineering and Technology
Thailand
December 2023

ACKNOWLEDGMENTS

ABSTRACT

Vision language (VL) models have shown promising performance across multiple tasks in both zero-shot and fine-tuning setups. Most studies use masked language modeling as a pre-training task, applying random masking to image caption tokens. However, random token masking is not an optimal strategy for training VL models, and effective masking strategies in VL remain underexplored. In this work, we investigate the effects of part of speech (POS) masking, as each POS category contributes differently to sentence meaning. By pre-training models with different POS masking strategies, we evaluate each model on image-text retrieval and visual question answering tasks, categorizing each question type following the VALSE. Our findings contribute to a deeper understanding of how POS masking influences model performance, providing insights that can lead to more effective pre-training strategies for future VL models.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Background

Vision language (VL) models have gained significant attention due to their ability to perform both zero-shot and transfer learning, achieving high performance across numerous downstream tasks through pre-training with web-scale image-text pairs (Mo, Kim, Lee, & Shin, 2024; Z. Wang, Wu, Agarwal, & Sun, 2022; J. Zhang, Huang, Jin, & Lu, 2024). Many VL models incorporate masked language modeling (MLM) as a pre-training task, making it an important method to train VL models (J. Li et al., 2021; C. Li et al., 2022; Chen et al., 2020; W. Wang et al., 2023; Tan & Bansal, 2019). Typically, a subset of word tokens is randomly masked at a percentage during training, and the model is tasked with predicting these masked tokens using information from both visual and language modality. This masking approach has proven to enhance the alignment between visual and linguistic representations, boosting performance in VL tasks (Tan & Bansal, 2019).

Despite the widespread adoption of MLM in VL training, its effects on model performance, efficiency and training loss remain underexplored. Bitton, Stanovsky, Elhadad, and Schwartz (2021) demonstrated that many of the randomly masked tokens are often stop-words or punctuation, which the model can easily learn without any need for masking. Another study by Wilf et al. (2023) demonstrated that selectively masking infrequent words from the pre-training dataset can boost model performance on out-of-domain datasets during continued training. Additionally, Tou and Sun (2024) suggested that random masking causes the model to rely heavily on local text signals, and it result in inefficient and inconsistent interactions between modalities, leading to suboptimal performance. These findings emphasize the importance of strategic token selection in MLM to enhance VL model performance and efficiency.

In this work, we aim to address the gap in understanding how masking each part of speech (POS) impacts VL models. Each POS contributes distinctively to sentence meaning: nouns typically denote objects, while verbs describe actions and often demand contextual comprehension. By selectively masking different parts of speech, we can better understand how each category influences the alignment between visual and linguistic information. The experiment is designed to answer the following questions:

1. How does masking each POS impact the performance, efficiency and training loss of VL pre-training models?
2. How does each POS masking strategy affect visual question answering (VQA) performance when analyzed based on different question types?
3. Does POS masking improve fine-grained alignment between image and text based on relationships, attibutes and order of image caption.

1.2 Objective

The objectives for our experiment are as listed.

1. Develop a pre-trained VL model to evaluate the impact of masking each POS on performance and training dynamics.
2. Benchmark the performance of our masking approach using specialized datasets to gain a deeper understanding of masking effects.

1.3 Scope

1. The model architecture is a cross-attention model, chosen for its ability to jointly predict answers based on information from multiple modalities.

2. The fine-tuning dataset is in the same domain as the training dataset.

CHAPTER 2

LITERATURE REVIEW

This section of the literature review is organized around two key topics relevant to our study. The first topic addresses VL models, providing an overview of the model architectures recently used in VL models and discussing the choice of the base architecture for the VL model used in this research. The second topic MLM, an important pre-training approach that has improved VL model performance. Together, these sections provide a comprehensive overview of the methodological foundations of this study.

2.1 Vision-Language model

In the early stage of VL learning, the goal of training is to align fine-grained features of the image with text. Many works have adopted object detection to create fine-grained labels for the training images (Chen et al., 2020; Bao et al., 2022). However, the focus of VL training has shifted to using web-scale image-text pairs as a training set, demonstrating competitive performance, as shown by CLIP (Radford et al., 2021). Radford et al. (2021) proposed contrastive training for VL with a large-scale image-text pairs dataset by optimizing the alignment of image and text encodings from the same pair, which was proven to be scalable by Jia et al. (2021). This approach has become a foundational model for VL tasks (Bommasani et al., 2021).

Recent advancements in VL model training can be roughly categorized into three main methods. The first approach is a separate unimodal encoder for each modality, as seen in models like CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021). This method is trained with the objective of aligning the intermediate outputs of each modality's encoding. The second method uses a cross-attention layer to fuse multimodal inputs, e.g., Flamingo (Alayrac et al., 2022), mPLUG (C. Li et al., 2022), LXMERT (Tan & Bansal, 2019), and ALBEF (J. Li et al., 2021). The cross-attention layer enables the model to fuse each modality more deeply. Finally, the third approach uses a single large attention model with concatenated image and text tokens as input, as in BEIT-3 (W. Wang et al., 2023), OSCAR (X. Li et al., 2020). This approach allows for early-stage fusion of each modality, though it requires the highest amount of computational resources. In this work, we adopt the cross-attention method as the base model due to its effectiveness in fusing multimodal inputs. Additionally, this approach allows the model to be trained using the MLM task.

2.2 Masked Language Modelling

MLM is a widely used pre-training method in language model (LM) training (Devlin, Chang, Lee, & Toutanova, 2018; Lan, 2019; Yu et al., 2022; S. Zhang et al., 2022; Guu, Lee, Tung, Pasupat, & Chang, 2020) as a self-supervised task. BERT (Devlin et al., 2018) proposed MLM as a pre-training task, which has been proven effective for pre-training language models. The MLM task involves replacing some input tokens with a special [MASK] token, and the model must predict the masked tokens based on the given unmasked tokens. In the field of VL models, many VL models have also adopted MLM as a training task to train the model to predict masked text based on visual information (J. Li et al., 2021; C. Li et al., 2022; Chen et al., 2020; W. Wang et al., 2023).

In the field of selective masking strategies in natural language processing, several works have further refined MLM to enhance training efficiency. ERNIE (Sun et al., 2019), SpanBERT (Joshi et al., 2020), and *n*-gram Masking (Levine et al., 2021) propose span masking instead of single-token masking, which forces the model to rely more on long-range dependencies rather than adjacent tokens, resulting in better performance compared to BERT (Devlin et al., 2018). Considering linguistic features, Yang, Zhang, and Zhao (2023) conducted a training analysis based on POS masking focused on LM training. The results showed that focusing the masking of non-function words (ADJ, ADV, NOUN, PROPN, and VERB) in the later stages of training can encourage the LM model to develop a better contextual understanding.

For selective masking in VL training, Bitton et al. (2021) introduced an object token masking strategy, selectively masking object tokens in image captions and pre-training the model. This approach achieved superior performance compared to random masking. Another study by Wilf et al. (2023) showed that selectively masking infrequent words from the pre-training dataset during continued training enhances model performance on out-of-domain datasets. Additionally, (Tou & Sun, 2024) proposed a curriculum-based masking strategy in which a reinforcement learning agent dynamically selects masking spans based on cross-modal interactions. This method improved the model's mulitmodalities understanding while reducing the dataset size needed for effective training. In this work, we conduct experiments to analyze the impact of each POS on results within a VL setting.

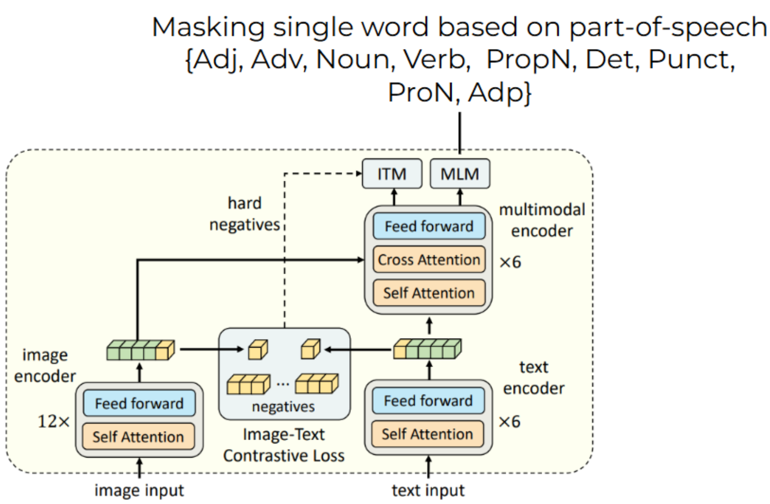
CHAPTER 3

METHODOLOGY

In this chapter, the methodology is detailed as follows. First, we describe the architecture of the model. Second, we explain all pre-training loss functions used in this experiment. Third, the details of POS tagging are provided. Fourth, we outline the datasets used in this experiment. Lastly, we provide details on the visual question answering setup.

Figure 3.1
Overall methodology

Pre-training the model with a MLM task by masking tokens based on the POS in the image captions.



3.1 Model architecture

As shown in Figure 3.1, our model includes three main components: an image encoder, a text encoder, and a multimodal encoder. The first component is the image encoder, for which we use ViT (Dosovitskiy et al., 2021), modified following (Radford et al., 2021), as the image encoder in this experiment. The second component is the text encoder, which employs a transformer architecture as BERT (Devlin et al., 2018) to encode image captions with BERT tokenizer for tokenization. The final component is the multimodal encoder, where VL interactions occur.

Given a training dataset D consisting of image-text pairs $(I_i, T_i) \in D$, where I_i is the image and T_i is the image caption of the i-th image, each image is first encoded as a sequence of tokens $\{v_{cls}, v_1, \ldots, v_n\}$ using ViT (Dosovitskiy et al., 2021). Here, v_{cls} represents the embedding of the [CLS] token prepended to the image patch sequence. In this experiment, the image encoder was initialized with ViT-B-32 pre-trained on ImageNet-21K (Deng et al., 2009). Next, we use a 6-layer transformer, randomly initialized, to encode the image caption T_i into text embeddings $\{w_{cls}, w_1, \ldots, w_n\}$, where w_{cls} is the embedding of the [CLS] token. Finally, both text and image encodings are passed through the multimodal encoder to fuse both inputs, producing multimodal encodings. For the multimodal encoder, a cross-attention layer is used, where both keys and values are the image encodings, and the text encoding serves as the query in the cross-attention layer.

3.2 Pre-training objectives

In this work, we pre-train our model with three objectives: masked language modeling (MLM), image-text constrative learning (ITC) and image-text matching (ITM).

3.2.1 Mask language modelling

Our model is trained with the masked language modeling task. Typically, a percentage of tokens $\{w_1, \ldots, w_T\}$ are replaced with a special [MASK] token to create a masked caption T^{mask}. However, in this work, the masked tokens are selected based on POS type instead of randomly masking. The model is then trained to predict the original tokens at the masked positions, conditioned on both the unmasked tokens in T^{mask} and the visual features of I as $p^{\text{mask}}(I, T^{\text{mask}})$. Let y^{mask} be a one-hot vector representing the ground-truth vocabulary for the masked token, where the masked token has a probability of 1. The model's objective is to minimize the cross-entropy $\mathbf{H}$, given by:

$$\mathcal{L}_{\text{MLM}} = \mathbf{H}(y^{\text{mask}}, p^{\text{mask}}(I, T^{\text{mask}})S))$$

3.2.2 Image-text contrative learning

To improve each unimodal encoders representation, we use image-text constrative learning to improve alignment of each modality. ITC aims to improve alignment by maximize similarity score of image and text from the same pair with score function $s(I,T) = v_{cls}^{\top} w_{cls}$, and minimize similarity score of image and text not from its pair. We then calculate softmax-normalized similiarity score for each image to any text and each text to any image, identified as image-to-text $p^{i2t} \in \mathbb{R}^M$ and text-to-image $p^{t2i} \in \mathbb{R}^M$ score as:

$$p_i^{i2t}(I) = \frac{\exp\left(s(I,T_i)\right)/\tau}{\sum_{m=1}^{M} \exp\left(s(I,T_m)/\tau\right.}, \quad p_i^{t2i}(T) = \frac{\exp\left(s(T,I_i)\right)/\tau}{\sum_{m=1}^{M} \exp\left(s(T,I_m)/\tau\right.}$$

where τ is a learnable temperature parameter. Let $y^{i2t}(I) \in \{0,1\}^M$ and $y^{t2i}(T) \in \{0,1\}^M$ be a ground truth with probability of 1 at a position of same pair, and probability of 0 on the otherhand. The ITC loss is calculated as cross-entropy $\mathbf{H}$ between p and y:

$$\mathcal{L}_{\text{ITC}} = \frac{1}{2}(\mathbf{H}(y^{i2t}, p^{i2t}) + \mathbf{H}(y^{t2i}, p^{t2i}))$$

3.2.3 Image-text matching

To further improve multimodal alignment in the VL model, image-text matching is employed to enhance alignment. The model is trained to predict whether an image and caption are from the same pair. A fully connected layer, followed by a softmax function, is added over the model. This layer takes the [CLS] embedding from the multimodal encoding as input to predict whether the pair is positive (matched) or negative (unmatched).

The loss function for ITM, using cross-entropy loss, is defined as:

$$\mathcal{L}_{\text{ITM}} = \mathbf{H}(y^{\text{itm}}, p^{\text{itm}}(I,T)),$$

where y^{itm} is a one-hot ground-truth label, and $p^{\text{itm}}(I,T)$ is the predicted class probability.

The full pre-training objective of our work can be written as:

$$\mathcal{L} = \mathcal{L}_{\text{MLM}} + \mathcal{L}_{\text{ITC}} + \mathcal{L}_{\text{ITM}}$$

3.3 Part of speech masking

In this experiment, we explored the effect of each POS on VL learning in term of performance, efficiency and training loss. For each image caption, each token have to be classified into POS categories for masking. We use POS-tagging tools SpaCy[1] to classify each word into POS classes based-on the Universal POS tag set[2].

3.4 Pre-training dataset

We pre-trained the model on the Conceptual Captions dataset (Sharma, Ding, Goodman, & Soricut, 2018), which consists of 3.3 million image-text pairs. In Conceptual Captions dataset, an automated process was used to select, filter, and refine these image-caption pairs to ensure they are clear, informative, and suitable for effective model training.

[1] POS-tagging tool SpaCy: https://spacy.io/

[2] Universal POS tag set: https://universaldependencies.org/u/pos/

3.5 Evaluation

In this work, we evaluate each model trained with different types of POS masking through image-text retrieval and visual question answering tasks. Details of the evaluation methods and datasets used in these tasks are provided in this section.

3.5.1 Image-text retrieval

For the image-text retrieval task, we evaluate the effect of masking on each POS category by performing zero-shot evaluations on the Flickr30K (Plummer et al., 2015) and VALSE (Parcalabescu et al., 2022) datasets for both image-to-text and text-to-image retrieval. The Flickr30K dataset is used to assess the model's overall performance in retrieval tasks. To gain deeper insights, we evaluate our model using the VALSE dataset, which categorizes linguistic phenomena into six distinct types: existence, plurality, counting, relation, action, and coreference. Each image caption in the VALSE dataset also includes a "foil" version, where words related to each caption category are modified. This setup allows us to analyze how different POS masking strategies affect the model's retrieval performance and the alignment between visual and textual representations.

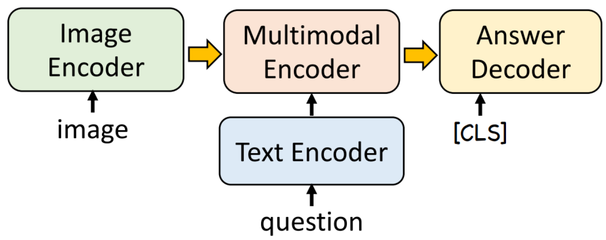
3.5.2 Attribution, Relation and Order benchmark

As demonstrated by Tou and Sun (2024), the results suggest that masking strategies impact a model's ability to understand attributes, relationships, and word order. In this work, we also benchmark each pre-trained model with specific POS masking against four tasks proposed by Yuksekgonul, Bianchi, Kalluri, Jurafsky, and Zou (2023): Visual Genome Relation (VGR), Visual Genome Attribution (VGA), COCO order (Co), and Filckr order (Fo). This benchmark is designed to assess the VL model's understanding of compositional relationships by swapping and replacing words in image captions, such as altering "The horse is eating the grass" to "The grass is eating the horse." Evaluating models against this benchmark provides valuable insights into their semantic and context understanding of vision and language modality.

Figure 3.2

Visual question answering model architecture

Modified model architecture for VQA task.



3.5.3 Visual question answering

The visual question answering (VQA) task requires the model to generate answers, which require an additional decoder over the multimodal encoder, as shown in Figure 3.2. In this work, we append a 6-layer transformer as a decoder, initialized with the parameters of the multimodal encoder. Answers are generated in an auto-regressive manner, with multimodal embeddings as input and a special start-of-sequence token ([CLS]) as the initial input to the decoder. The benchmark dataset for the VQA task is the VQA2.0 dataset (Goyal, Khot, Summers-Stay, Batra, & Parikh, 2017), which is constructed using images from COCO (Lin et al., 2014). This dataset includes 83,000 images for training, 41,000 for validation, and 81,000 for testing. We further train our model using the VQA2.0 training set. We categorized each question into six categories following VALSE linguistic phenomena categories to provide deeper insights.

CHAPTER 4

Results

In this chapter, we provide all the experiment results for each experiment and evaluation.

4.0.1 Pre-training

CHAPTER 5

DISCUSSION

CHAPTER 6

CONCLUSION

REFERENCES

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., ... others (2022). Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, *35*, 23716–23736.

Bao, H., Wang, W., Dong, L., Liu, Q., Mohammed, O. K., Aggarwal, K., ... Wei, F. (2022). Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *Advances in Neural Information Processing Systems*, *35*, 32897–32912.

Bitton, Y., Stanovsky, G., Elhadad, M., & Schwartz, R. (2021). Data efficient masked language modeling for vision and language. *arXiv preprint arXiv:2109.02040*.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... others (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Chen, Y.-C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., ... Liu, J. (2020). Uniter: Universal image-text representation learning. In *Computer vision – eccv 2020: 16th european conference, glasgow, uk, august 23–28, 2020, proceedings, part xxx* (p. 104–120). Berlin, Heidelberg: Springer-Verlag. Retrieved from `https://doi.org/10.1007/978-3-030-58577-8_7` doi: 10.1007/978-3-030-58577-8_7

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 ieee conference on computer vision and pattern recognition* (pp. 248–255).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International conference on learning representations.* Retrieved from `https://openreview.net/forum?id=YicbFdNTTy`

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., & Parikh, D. (2017). Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 6904–6913).

Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M.-W. (2020). *Realm: Retrieval-augmented language model pre-training.* Retrieved from https://arxiv.org/abs/2002.08909

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., ... Duerig, T. (2021). Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning* (pp. 4904–4916).

Joshi, M., Chen, D., Liu, Y., Weld, D. S., Zettlemoyer, L., & Levy, O. (2020). SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, *8*, 64–77. Retrieved from https://aclanthology.org/2020.tacl-1.5 doi: 10.1162/tacl_a_00300

Lan, Z. (2019). Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Levine, Y., Lenz, B., Lieber, O., Abend, O., Leyton-Brown, K., Tennenholtz, M., & Shoham, Y. (2021). {PMI}-masking: Principled masking of correlated spans. In *International conference on learning representations.* Retrieved from https://openreview.net/forum?id=3Aoft6NWFej

Li, C., Xu, H., Tian, J., Wang, W., Yan, M., Bi, B., ... others (2022). mplug: Effective and efficient vision-language learning by cross-modal skip-connections. *arXiv preprint arXiv:2205.12005*.

Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., & Hoi, S. C. H. (2021). Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, *34*, 9694–9705.

Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., ... others (2020). Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer vision–eccv 2020: 16th european conference, glasgow, uk, august 23–28, 2020, proceedings, part xxx 16* (pp. 121–137).

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *Computer vision–eccv 2014: 13th european conference, zurich, switzerland, september 6-12, 2014, proceedings, part v 13* (pp. 740–755).

Mo, S., Kim, M., Lee, K., & Shin, J. (2024). S-clip: Semi-supervised vision-language learning using few specialist captions. *Advances in Neural Information Processing Systems*, *36*.

Parcalabescu, L., Cafagna, M., Muradjan, L., Frank, A., Calixto, I., & Gatt, A. (2022, May). VALSE: A task-independent benchmark for vision and language models centered on linguistic phenomena. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 8253–8280). Dublin, Ireland: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2022.acl-long.567

Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., & Lazebnik, S. (2015). Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the ieee international conference on computer vision* (pp. 2641–2649).

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... others (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763).

Sharma, P., Ding, N., Goodman, S., & Soricut, R. (2018). Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of acl.*

Sun, Y., Wang, S., Li, Y., Feng, S., Chen, X., Zhang, H., ... Wu, H. (2019). ERNIE: enhanced representation through knowledge integration. *CoRR, abs/1904.09223*.

Tan, H., & Bansal, M. (2019). Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 conference on empirical methods in natural language processing.*

Tou, K., & Sun, Z. (2024, June). Curriculum masking in vision-language pretraining to maximize cross modal interaction. In K. Duh, H. Gomez, & S. Bethard (Eds.), *Proceedings of the 2024 conference of the north american chapter of the association for computational linguistics: Human language technologies (volume 1: Long papers)* (pp. 3672–3688). Mexico City, Mexico: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2024.naacl-long.203 doi: 10.18653/v1/2024.naacl-long.203

Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., ... others (2023). Image as a foreign language: Beit pretraining for vision and vision-language tasks. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 19175–19186).

Wang, Z., Wu, Z., Agarwal, D., & Sun, J. (2022). *Medclip: Contrastive learning from unpaired medical images and text.* Retrieved from https://arxiv.org/abs/2210.10163

Wilf, A., Akter, S. N., Mathur, L., Liang, P. P., Mathew, S., Shou, M., ...

Morency, L.-P. (2023). Difference-masking: Choosing what to mask in continued pretraining. *arXiv preprint arXiv:2305.14577*.

Yang, D., Zhang, Z., & Zhao, H. (2023). *Learning better masking for better language model pre-training.*

Yu, W., Zhu, C., Fang, Y., Yu, D., Wang, S., Xu, Y., … Jiang, M. (2022). *Dict-bert: Enhancing language model pre-training with dictionary.* Retrieved from `https://arxiv.org/abs/2110.06490`

Yuksekgonul, M., Bianchi, F., Kalluri, P., Jurafsky, D., & Zou, J. (2023). When and why vision-language models behave like bags-of-words, and what to do about it? In *International conference on learning representations.* Retrieved from `https://openreview.net/forum?id=KRLUvxh8uaX`

Zhang, J., Huang, J., Jin, S., & Lu, S. (2024). Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *46*(8), 5625-5644. doi: 10.1109/TPAMI.2024.3369699

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., … Zettlemoyer, L. (2022). *Opt: Open pre-trained transformer language models.* Retrieved from `https://arxiv.org/abs/2205.01068`