PART OF SPEECH MASKING EFFECT ON VISION-LANGUAGE REPRESENTATION LEARNING

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 (Chairperson) Dr. Chantri Polprasert Dr. Attaphongse Taparugssanagorn

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 2025

AUTHOR'S DECLARATION

I, Pasit Tiwawongrut, declare that the research work carried out for this thesis was in accordance with the regulations of the Asian Institute of Technology. The work presented in it are my own and has been generated by me as the result of my own original research, and if external sources were used, such sources have been cited. It is original and has not been submitted to any other institution to obtain another degree or qualification. This is a true copy of the thesis, including final revisions.

Date: 22 August 2025

Name: PASIT TIWAWONGRUT

Signature:

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my parents, whose unconditional love, support, and encouragement have been the foundation of my journey.

I am sincerely thankful to the Asian Institute of Technology for providing me with the opportunity and environment to pursue my studies. My heartfelt appreciation also goes to my supervisor, whose guidance, patience, and invaluable insights have shaped the direction and quality of this research. I am equally grateful to the committee members for their constructive feedback and thoughtful suggestions, which have strengthened this work.

I would also like to thank ThaiSC for providing the computational resources essential to carrying out the experiments presented in this research.

Finally, I wish to acknowledge myself for the perseverance, dedication, and resilience that made this work possible.

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.

Our experiments show that the choice of masked tokens matters. For retrieval tasks, masking simpler tokens like determiners leads to higher accuracy than masking nouns, suggesting that freeing the model from predicting harder words can improve overall alignment. For VALSE and VQA, selective POS masking consistently performs better than random masking, and content-word masking helps most with fine-grained understanding. Even categories that perform less well in retrieval still add value in VQA, showing that different parts of speech support different aspects of cross-modal learning. We also confirm that models trained with MLM consistently outperform those trained without it, especially downstream task.

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 incorporated 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 modalities. 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, the effects of masking tokens based on sentence structure remain underexplored. Prior work has shown that effectiveness increases when the masked tokens are chosen to be semantically informative. For example, masking object words yields clear gains over random masking (Bitton, Stanovsky, Elhadad, & Schwartz, 2021); selectively masking infrequent words improves out-of-domain generalization during continued pre-training (Wilf et al., 2023); and curriculum-based masking reduces shallow reliance on local cues and promotes more consistent cross-modal interactions (Tou & Sun, 2024). 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 inspired by how human interpret the world through language, where each part of speech serves a distinct purpose. By selectively masking different parts of speech, we can better understand how each POS category affects the alignment between visual and linguistic information. This also allows us to probe what information the model can infer beyond the masked word itself. To further explore the effect of each POS, training without the MLM task and with different POS masking probabilities are compared.

The experiment is designed to answer the following questions:

1. How does masking each POS affect the performance and training loss of VL models during pre-training, and how does it influence downstream performance on visual question answering (VQA)?
2. What underlying representations do VL models acquire through MLM training, and does this process enable them to learn more than the masked word itself?
3. What is the difference between training without the MLM task compared to training with it, and when masking each POS with a 100 percent masking ratio?

The main contributions of this thesis are fourfold. First, we present a systematic study of POS masking strategies in VL pre-training, offering new insights into how each POS category contributes to cross-modal alignment. Second, we benchmark the effects of different POS masking strategies across retrieval and VQA tasks, identifying when and where specific linguistic categories are most influential. Third, we compare models trained with no MLM, standard random masking, and 100% POS-specific masking, providing a deeper understanding of the role of MLM in both retrieval and VQA. Finally, our findings offer insights for developing more efficient and linguistically informed masking strategies in future VL systems.

This study is aimed at serving as a foundation for future research on selective masking techniques, contributing to the development of VL models that are more robust, and data-efficient. By systematically exploring how different POS affect learning, we provide new insights into the role of language structure in multimodal pre-training.

1.2 Scope

The scope of this thesis is defined as follows:

1. The training and testing datasets are web-scale image–text pairs.
2. The scope is limited to cross-attention vision–language models trained with MLM, ITM, and ITC tasks, which are widely adopted objectives in modern VL pre-training.
3. This study concentrates on structured masking guided by POS, specifically focusing on nouns, verbs, adjectives, adverbs, proper nouns, determiners, auxiliaries, pronouns, and adpositions.

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 is 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 Representation Learning

VL learning aims to align visual and linguistic information for multimodal tasks that require reasoning across both modalities, such as image captioning, visual question answering, and multimodal retrieval. The training objective can be divided roughly into three main categories: contrastive, generative, and alignment. Firstly, the contrastive learning objective trains VL representations by maximizing the alignment score between paired images and text while minimizing the score between unpaired images and text (Radford et al., 2021; Jia et al., 2021; J. Yang et al., 2022). Secondly, the generative learning objective focuses on reconstructing masked tokens in either the image or text modality, or both, to learn VL representations (Singh et al., 2022; J. Li et al., 2021; Alayrac et al., 2022). This objective requires the model to utilize both modalities to reconstruct missing tokens, which enhances alignment. Lastly, the alignment objective involves learning VL representations by predicting whether an image and text pair match (Bao et al., 2022). The combined use of these three training objectives has proven effective and is commonly applied across various pre-trained VL models.

Recent advancements in VL fusion methods can be roughly categorized into three main approaches. 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), 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), UNITER (Chen et al., 2020), FLAVA (Singh et al., 2022), and mPLUG (C. Li et al., 2022). 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. We also use all three training objectives with a modified MLM for this experiment.

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, D. 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-functional words, including 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 multimodalities 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

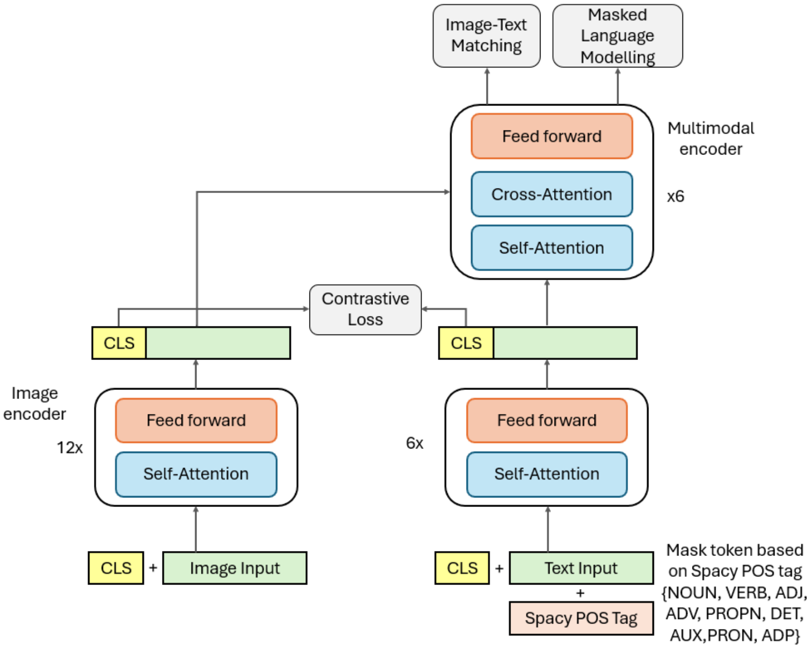
METHODOLOGY

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

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 were selected based on POS type instead of randomly masking. The model 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:

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

For the masking ratio, each POS token is masked with either a 70 percent or 100 percent probability. In this work, random token masking was also tested with a masking ratio of 15 percent.

3.2.2 Image-Text Contrastive Learning

To improve each unimodal encoder's representation, we used image-text constrative learning to improve alignment of each modality. ITC aims to improve alignment by maximizing the similarity score of image and text from the same pair with the score function $s(I,T) = v_{cls}^{\top} w_{cls}$, and minimizing the similarity score of image and text not from its pair. We then calculate the softmax-normalized similarity 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 $p^{t2i} \in \mathbb{R}^M$ scores 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 the same pair, and probability of 0 on the other hand. 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 was 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

For each image caption, each token was classified into POS categories for masking. We used POS-tagging tools SpaCy[1] to classify each word into POS classes based on the Universal POS tag set[2]. In this work, we modified the BERT tokenizer to integrate with SpaCy by using the Tokenizations[3] tool to align BERT token IDs with SpaCy tokens IDs.

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

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

[3] Tokenizations alignment library tool: https://github.com/explosion/tokenizations

In this experiment, we explored the effect of each POS on VL learning in terms of performance, and training loss. For the main experiment, each token was assigned to one of nine POS categories: NOUN (nouns), VERB (verbs), ADJ (adjectives), ADV (adverbs), PROPN (proper nouns), DET (determiners), AUX (auxiliaries), PRON (pronouns), and ADP (adpositions), and masked with a 70% probability. For evaluation, these POS were further classified as functional (determiners, auxiliaries, pronouns, adpositions) or non-functional (nouns, verbs, adjectives, adverbs, proper nouns). For the 100 percent masking setting, where all tokens corresponding to a specific POS are masked, we conducted experiments on non-functional parts of speech.

3.4 Pre-Training Dataset

We pre-trained the model on the Conceptual Captions dataset (Sharma, Ding, Goodman, & Soricut, 2018) and the MSCOCO dataset, totaling 2.4 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.

3.5 Evaluation

In this work, we evaluated each model trained with different types of POS masking through image-text retrieval, image-text matching, 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, the model was tested by performing zero-shot evaluations on the Flickr30K (Plummer et al., 2015) dataset for both image retrieval (IR) and text retrieval (TR). The Flickr30K dataset is used to assess the model's overall performance in retrieval tasks. This setup allowed us to analyze how different POS masking strategies affect the model's retrieval performance and the alignment between visual and textual representations.

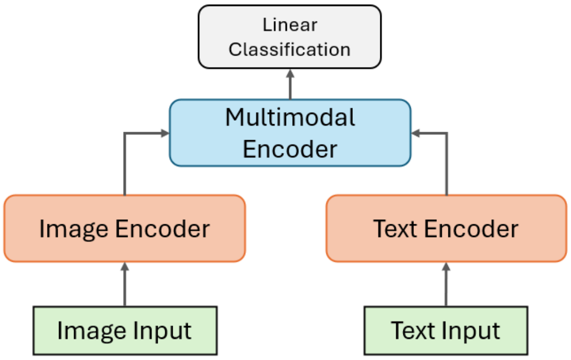
3.5.2 Image-Text Matching

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 benchmarked each pre-trained model with specific POS masking against VALSE benchmark (Parcalabescu et al., 2022). For the VALSE dataset, this benchmark categorizes each image-text sample into different linguistic phenomena as showed in Table 3.1, including 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 task is a classification task, where the model has to predict the correct caption for each image. We evaluated the model in a zero-shot manner by reusing the ITM head as a classifier. Evaluating models against this benchmark provides valuable insights into their semantic and contextual understanding of vision and language modality.

3.5.3 Visual Question Answering

In this work, the visual question answering (VQA) task was treated as a classification task. A classification head was appended to generate the answer, as shown in Figure 3.2. 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.

Figure 3.2
Visual question answering model architecture



3.6 Training

The model was pre-trained on a machine equipped with four NVIDIA A100 GPUs. The pre-training of the model was conducted using a batch size of 64 with 10 epochs. We used the AdamW optimizer with an initial learning rate of 1×10^{-4} and a weight decay of 0.02 to help regularize the training process. A cosine learning rate scheduler was applied, with the learning rate gradually increasing from a warm-up value of 1×10^{-5} during the first 5 epochs, before decaying towards a minimum learning rate of 1×10^{-5} by the end of training.

For the VQA task, the model was trained with a batch size of 32 on the same machine as pre-training. We used the AdamW optimizer with a learning rate of 2×10^{-5} and a weight decay of 0.02. A cosine learning rate scheduler was applied over 8 epochs, with a warm-up phase of 1 epoch starting at a learning rate of 1×10^{-5}, and decaying to a minimum of 1×10^{-6}.

VALSE Task	**Test**	**Example**
Existence Quantifier	Detect object presence or absence	“A cat on bed” vs. “A dog on bed”
Plurality Number	Singular vs. plural	“One flower” vs. “Some flowers”
Counting Balanced	Count with equal samples per class	3 apples vs. 5 apples
Counting Adversarial	Test for small-number bias	counts ≥ 4 vs counts 0–3
Counting Small Numbers	Count small numbers only	counts < 4
Spatial Relation	Understand positions	“Book on table” vs. “Book under table”
Action Replacement	Correct action	“Holding ball” vs. “Throwing ball”
Actant Swap	Correct roles	“Boy chases dog” vs. “Dog chases boy”
Coreference Standard	Pronoun–Entity Relation Understanding (from test set)	“Woman talks to girl. She smiles.”
Coreference Hard	Coreference Standard (from val set)	“Boy hugs dog. He is happy.”
Foil-COCO	Spot small caption error	Correct vs. nearly identical with one mistake

Table 3.1
VALSE dataset task explanations.

CHAPTER 4

Results

This chapter presents all the experiment results for each experiment and evaluation, aimed at addressing the research questions. The result are divided into four sections, including pre-training, image-text matching, visual question answering, and the evaluation result of varying the POS masking percentage.

4.1 Pre-training

To address the question of how masking each POS affects the performance and training loss of VL pre-training models, we present all relevant results in this section. All losses, including MLM, ITC, and ITM, along with the Flickr30K evaluation results, are provided. The loss values are plotted on a logarithmic scale to visualize improvements over time across different POS masking strategies. The results from training the ALBEF model using the same dataset are also included for consistent comparison. We also provided the histrogram of each part-of-speech tag in the pre-training dataset as shown in Figure 4.4.

4.1.1 Flickr30K

The Flickr30K evaluation results are shown in Table 4.1, which presents the top-1, top-5, and top-10 retrieval scores for both TR and IR tasks across different training methodologies. By comparing r@1 performance for both TR, and IR, the model with determiner masking achieves the highest overall performance. Among the non-functional group, masking NOUN yields the best performance. By masking ADV and PROPN causes the most significant degradation compared to the random masking baseline.

From the training loss curves, it is evident that different POS categories affect

Masking Method		Flickr30K					
		TR			IR		
		r@1	r@5	r@10	r@1	r@5	r@10
ALBEF		70.40	89.50	94.00	54.66	82.02	88.70
Random Masking		67.00	88.00	93.75	52.61	80.14	87.76
Non-functional	NOUN	67.15	88.60	94.65	52.73	80.45	87.79
	VERB	54.85	82.85	90.05	43.82	73.84	82.82
	ADJ	62.30	87.30	92.40	47.39	75.47	84.06
	ADV	46.85	76.25	85.75	36.40	66.38	76.78
	PROPN	44.85	74.40	84.10	34.91	64.09	75.01
Functional	DET	71.05	92.00	95.30	56.01	81.93	88.59
	AUX	52.10	79.60	88.20	41.13	70.92	80.68
	PRON	51.45	78.80	87.10	39.97	69.58	79.32
	ADP	65.05	88.25	93.40	51.19	78.83	85.15

Table 4.1
Flickr30K benchmark image retrieval result.

the convergence behavior in difference ways. The loss for MLM, ITC, and ITM are displayed in the Figure 4.1, Figure 4.2, and Figure 4.3 respectively For both ITM and ITC, the loss curves are similar in behavior and follow a consistent order relative to each other. In the MLM loss graph, we can see that POS masking in the functional group result in lower loss, while those in the non-functional group show higher loss, and the random masking show the highest loss by the end of training.

Taken together, the results show that masking each POS impacts both the training loss trajectory and final model performance in distinct ways. By observing the MLM loss graph, we find that non-functional POS are more difficult for the model to learn through the MLM task, whereas functional POS are learned more quickly. The ranking of the performance for each POS masking method aligns with the ITM and ITC loss curves, where a lower loss corresponds to higher retrieval accuracy.

Figure 4.1

MLM loss curves for different POS masking strategies (log scale).

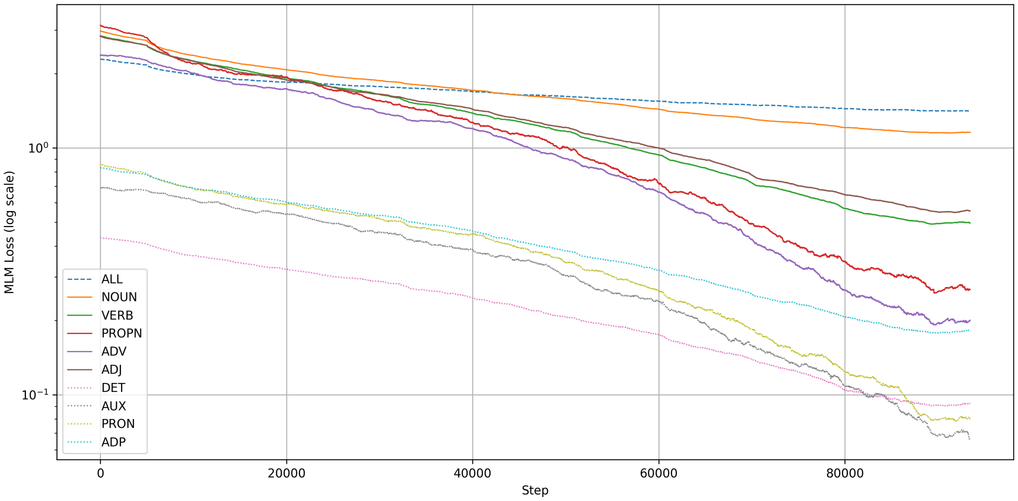


Figure 4.2

ITC loss curves for different POS masking strategies (log scale).

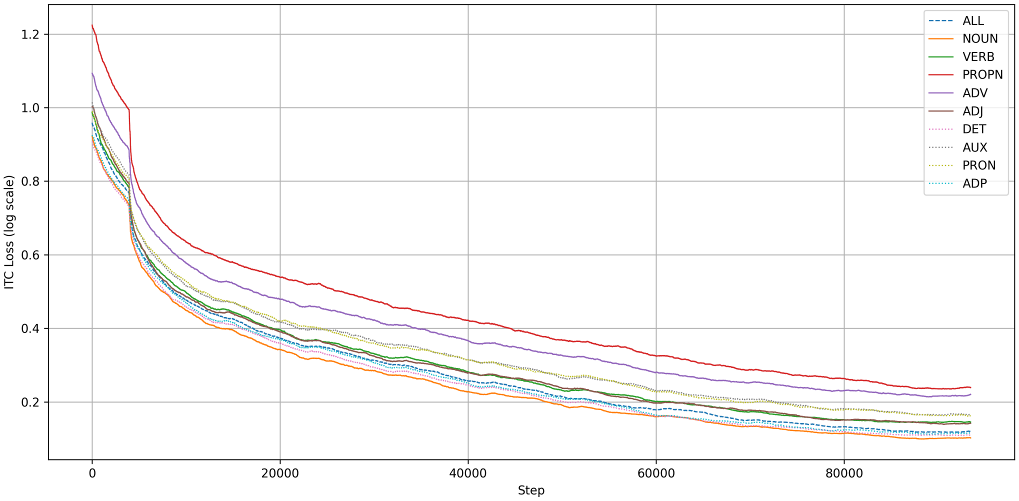
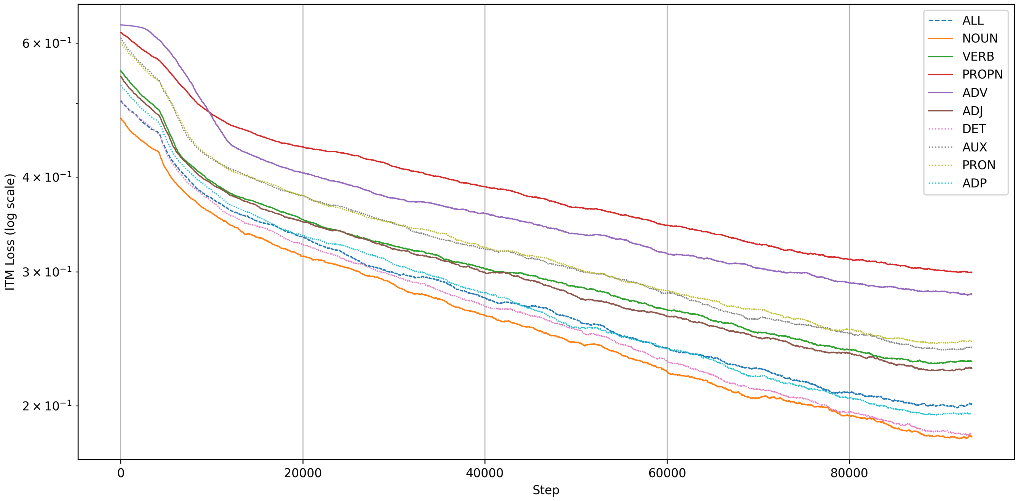


Figure 4.3

ITM loss curves for different POS masking strategies (log scale).

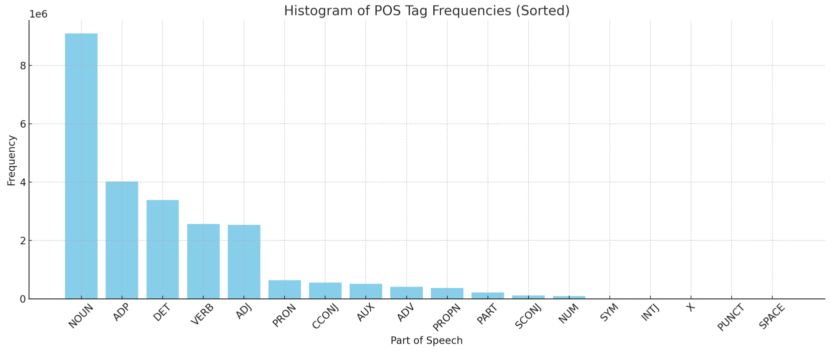


4.1.2 Histogram of POS tag

This section provides a visualization of tokens categorized by their POS from the training dataset, as shown in Figure 4.4. The histogram illustrates the frequency distribution of POS tags, sorted from the most to least common. NOUN tokens dominate the dataset, followed by ADP, DET, VERB, and ADJ, while categories such as SYM, INTJ, X, PUNCT, and SPACE appear rarely in the training dataset.

Figure 4.4

Histogram of POS tag frequencies in the training dataset (sorted by frequency).



4.2 Image-Text Matching

In this section, we evaluate the impact of POS masking on image-text matching by reporting zero-shot classification accuracy on the VALSE benchmark with a random masking method as a baseline. This experiment is designed to investigate whether the model acquires knowledge not only of the masked word itself but also of surrounding context.

4.2.1 VALSE

Table 4.2 summarizes zero-shot classification accuracy on the VALSE benchmark for each POS masking strategy. For completeness, we include the results of training without any masking, which serves as a reference point. Random masking is treated as the baseline, and for clarity, the best-performing method in each task (excluding the baseline, and no masking) is highlighted in gray, while the second-best method is highlighted in lightgray. If random masking achieves the highest score, the corresponding result is underlined. This visualization allows us to focus on the relative contributions of different POS categories without being overshadowed by the random masking baseline.

The results reveal that different parts of speech contribute most strongly to tasks that align with their linguistic functionals, while also offering useful cues beyond their primary roles. NOUN masking not only achieves the strongest performance on object-centric tasks such as existence task, but also shows competitive results on counting and action tasks. VERB masking, as expected, performs well on action task, but it also improves performance on counting task. Similarly, adjective masking achieves strong results on counting task. By contrast, PROPN masking underperforms in many tasks.

For the functional POS group, we observe that DET masking perform the best overall, achieving the highest scores in plurality, counting, and Foil-it! tasks, while also showing competitive performance in spatial relation task. AUX masking yields the strongest results on coreference tasks and additionally perform on counting. PRON also contribute most effectively to coreference. ADP, on the other hand, achieve their best results on action actant swap task.

When comparing the best-performing POS masking results against the no-masking baseline, we find mixed outcomes. Existence Quantifier shows no significant improvement through masking, while Plurality experiences a modest drop of 1.61In contrast, Counting tasks benefit substantially, with gains of 1.12%, 4.74%, and 7.74% for the Balanced, Small Number, and Adversarial sub-tasks, respectively. Spatial Relation shows only a minor decrease of 0.58%, while Action Replacement improves by 2.44% but Actant Swap drops slightly by 0.45%. Coreference tasks improve by approximately 3%, whereas Foil-it! shows no significant difference.

Masking Method		VALSE											
		Existence	Prularity	Counting			Sp.Re [1]	Action		Coreference		Foil-it!	Avg
		quantifiers	number	balanced	small number	adversarial	relations	replacement	actant swap	standard	clean		
Random Masking		65.06	61.43	54.64	57.81	62.83	61.61	68.04	51.88	49.70	43.37	85.79	60.20
Non-functional	NOUN	67.63	62.60	52.59	54.64	64.39	59.84	68.15	48.87	51.31	49.21	85.69	60.45
	VERB	60.37	60.50	54.83	56.30	61.52	57.68	68.24	48.62	51.31	42.40	83.45	58.66
	ADJ	60.85	60.55	54.00	56.84	67.01	57.68	65.68	50.92	50.34	44.74	83.01	59.24
	ADV	62.56	58.74	53.08	57.32	59.92	58.10	65.74	49.11	49.04	41.30	84.28	58.11
	PROPN	61.51	59.23	52.49	56.25	61.26	55.86	64.31	50.85	50.36	43.03	82.62	57.98
Functional	DET	60.14	63.33	53.47	57.86	65.40	59.06	66.67	50.43	50.09	38.99	87.94	59.40
	AUX	56.73	60.60	51.76	57.32	60.59	56.48	65.04	50.65	49.33	51.39	84.62	58.59
	PRON	56.05	61.33	50.39	54.88	58.87	58.93	64.36	48.05	53.23	50.48	83.40	58.18
	ADP	66.27	61.23	53.52	57.03	66.04	58.28	67.73	52.14	50.05	46.13	86.38	60.44
No Masking		67.61	64.94	53.71	53.12	59.27	60.42	65.8	52.59	50.47	48.31	87.60	60.35

Table 4.2

VALSE benchmark for image-text matching result.

4.3 Visual Question Answering

Table 4.3 presents the VQA2.0 test-dev performance after fine-tuning on the VQA task for each POS masking strategy, with results reported for Yes/No, Number, and Other question types. NOUN masking achieved the highest overall accuracy (70.29%), closely followed by random masking (70.28%). Within the non-functional group, NOUN masking performed best, while VERB (69.13%) and ADJ (69.09%) achieved similar scores. ADV masking yielded the lowest performance (64.12%), largely due to reduced accuracy in the Number and Other categories. For functional categories, DET and ADP masking achieved similar overall results (68.98% and 68.96%), with AUX (67.09%) and PRON (66.55%) performing lower. Comparing the best-performing non-functional and functional POS masking strategies shows a performance difference of 1.31%.

[1] Spacial Relation

Masking Method		VQA2.0 test dev			
		Yes/No	Number	Other	Overall
Random Masking		87.88	49.64	59.63	70.28
Non-functional	NOUN	87.84	49.49	60.03	70.29
	VERB	87.17	48.39	58.43	69.13
	ADJ	86.69	48.86	58.64	69.09
	ADV	83.10	43.83	52.49	64.12
	PROPN	85.07	46.38	56.60	67.71
Functional	DET	87.35	49.49	57.68	68.98
	AUX	85.25	46.59	56.24	67.09
	PRON	84.13	46.29	56.15	66.55
	ADP	87.07	48.82	58.07	68.96

Table 4.3

VQA2.0 test-dev benchmark result.

4.4 Masking Ratio

In this experiment, we report the effect of masking probability on both the Flickr30K and VQA benchmarks by comparing 0%, 70%, and 100% masking levels. For Flickr30K, we focus on exploring POS masking for non-functional POS categories, as shown in Table 4.4. For the VQA task, we compare NOUN masking—identified as the best-performing strategy on VQA, as shown in Table 4.5.

For the Flickr30K benchmark, we observe that the no-masking method outperforms models trained with the MLM objective. Specifically, it achieves a 7.45%, and 5.31% improvement in TR and IR r@1, respectively, compared to NOUN masking with 70% masking probability, which is the best among the masking-based methods. We also observe performance improvements when increasing the masking probability for VERB, ADV, and PROPN categories. On the other hand, performance deteriorates as masking probability increases for NOUN and ADJ masking.

For the VQA results, models trained with the MLM objective consistently outperform those trained without masking, showing an average improvement of approximately 2% in overall accuracy. The improvement occurs most significantly on the Other question type, improving by 2.81% and 2.60% for NOUN masking at 70% and 100% masking probabilities, respectively. Under NOUN masking, increasing the masking probability to 100% results in slight improvements for the Yes/No and Number question types, while performance slightly declines for the Other category.

Masking Method	**Masking probability**	**Flickr30K**					
		TR			IR		
		r@1	r@5	r@10	r@1	r@5	r@10
No Masking	0	74.60	92.50	95.90	58.04	83.82	90.04
NOUN	70	67.15	88.60	94.65	52.73	80.45	87.79
	100	65.80	90.40	94.90	53.34	78.94	86.72
VERB	70	54.85	82.85	90.05	43.82	73.84	82.82
	100	56.70	83.40	90.70	44.52	74.24	83.52
ADJ	70	62.30	87.30	92.40	47.39	75.47	84.06
	100	62.20	87.30	92.50	47.08	75.78	84.22
ADV	70	46.85	76.25	85.75	36.40	66.38	76.78
	100	50.10	78.80	87.90	37.74	67.78	78.00
PROPN	70	44.85	74.40	84.10	34.91	64.09	75.01
	100	49.10	78.30	85.90	36.06	66.88	77.22

Table 4.4
Flickr30K benchmark image retrieval result.

Masking Method	Masking Probability	VQA2.0 test dev			
		Yes/No	Number	Other	Overall
NOUN	70	87.84	49.49	60.03	70.29
NOUN	100	87.88	49.81	59.82	70.24
No Masking	0	87.42	49.43	57.22	68.78

Table 4.5
VQA2.0 test-dev benchmark result.

CHAPTER 5

DISCUSSION

This chapter presents a discussion of our experimental results and reflects on their implications in relation to the main research questions: How does masking each POS affect the performance and training loss of VL models during pre-training, and how does it influence downstream performance on VQA? We begin by examining how masking different POS categories impacts model performance during both pre-training and downstream evaluation. We then discuss the results on VQA, followed by an analysis of the no-masking baseline. Finally, we outline the limitations of our approach.

5.1 The Effect of Each POS in Visual Pre-training

Our experimental findings highlight that masking different POS during pre-training leads to distinct outcomes. On the Flickr30K benchmark, DET masking achieved the highest retrieval performance across both TR and IR tasks, followed by NOUN masking among the non-functional POS group. Intuitively, one might expect nouns typically carrying more semantic should outperform determiners. However, the results reveal the opposite, suggesting that determiners are easier for the model to learn. When combined with the loss curves, we observe that retrieval performance correlates more closely with the ITM and ITC objectives than with the MLM task. The ranking of each POS masking method aligns with the ITM and ITC loss curves, where lower loss corresponds to higher retrieval accuracy. This suggests that for retrieval tasks, masking simpler tokens such as determiners may reduce the burden on the MLM task and lead to better performance.

5.2 The Effect of Each POS in Based on Linguistic Phenoma VALSE Benchmark

The VALSE benchmark evaluates a model's fine-grained linguistic understanding. We find that selectively masking specific POS categories consistently outperforms random masking, and that each POS-based masking strategy also performs well on tasks related to its corresponding linguistic functional. This suggests that the MLM objective plays a particularly important role in tasks requiring fine-grained understanding, and that performance is highly sensitive to which tokens are masked, highlighting the importance of strategic token selection.

Furthermore, We also observe that POS masking captures more than just the task directly tied to its linguistic function. For example, training the model with VERB and ADJ masks also yields strong performance on the counting task. From the results, there remains a clear opportunity to improve performance by leveraging the strengths of each POS category into a single model.

5.3 Visual Question Answering

In the VQA task, we observe that even some POS categories that performed poorly in retrieval still enable the model to retain fine-grained image understanding. Specifically, most non-functional POS masking methods lead to better performance compared to functional POS masking. This result emphasizes that masking more content words leads to better fine-grained alignment. If we combine the results of VALSE with VQA, we can see that performance on the VALSE dataset is directly related to performance on the VQA dataset, as both tasks require fine-grained understanding.

5.4 Masking Probability

When compared to training without masking, models trained without MLM may perform reasonably well in a zero-shot setting. However, models trained with the MLM objective show clear improvements when fine-tuned on specific tasks. This highlights the importance of MLM for VL model improvement.

5.5 Limitations

The limitation in term of dataset this study is the imbalance in POS distribution, as observed in the POS histogram. Certain POS categories, such as nouns, appear far more frequently than others, which may introduce bias in the model's learning process and affect the generalizability of the results. While this imbalance reflects the natural distribution of language in real-world datasets, it may confound our interpretation of how each POS contributes to VL learning. Another factor is that the performance of our work may not be optimal, since we deliberately adopted standard methods to ensure fair comparison. Additionally, our experiments are limited to a specific set of benchmarks. Evaluation the effects of POS masking across a broader range of datasets would be necessary to confirm the consistency and robustness of the observed phenomena.

CHAPTER 6

CONCLUSION

6.1 Conclusion

In this work, we systematically investigate POS masking strategies in VL pre-training and their effects on cross-modal alignment, retrieval, and VQA tasks. Our findings show that different POS categories yield distinct outcomes: for retrieval, simpler tokens such as determiners reduce the burden on the MLM objective and lead to higher accuracy, while for fine-grained benchmarks like VALSE and VQA, masking strategies aligned with linguistic functionals consistently outperform random masking and reveal strong sensitivity to token selection. Notably, even POS categories that perform poorly in retrieval still enhance fine-grained understanding, with content-word masking contributing most to alignment. Moreover, models trained with MLM consistently surpass those without it when fine-tuned, underscoring the crucial role of MLM in VL model improvement. Together, these findings not only deepen our understanding of linguistic contributions in VL pre-training but also show opportunities for designing more effective and linguistically grounded models.

6.2 Future work

Building on our findings, we identify a clear gap in fully leveraging POS masking strategies in combination to enhance model performance. Future work could explore adaptive masking approaches that integrate multiple POS categories, selecting them dynamically to suit different tasks. In addition, adjusting loss scheduling presents another promising direction, as our results suggest that certain tasks depend more heavily on specific objectives.

Due to our scope, we focus only on a single POS to ensure fairness and comparability. However, further research following a curriculum-based approach across multiple POS categories could be beneficial. By progressively combining words from different POS categories to build more simpler sentence or a increase complexity of the sentences.

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*.

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.*

Singh, A., Hu, R., Goswami, V., Couairon, G., Galuba, W., Rohrbach, M., & Kiela, D. (2022). Flava: A foundational language and vision alignment model. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 15638–15650).

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.*

Yang, J., Li, C., Zhang, P., Xiao, B., Liu, C., Yuan, L., & Gao, J. (2022). Unified contrastive learning in image-text-label space. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 19163–19173).

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

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