big data and cognitive computing

MDPI

Article

Improving Entity Understanding for Vision-Language Pre-Training via Active Learning

Qunbo Wang [1], Sen Zhang [1], Boxuan Shao [2], Xize Guo [3], Jiayong An [1], Chao Fan [3], Yuanjun Jing [4], Junxian Li [5] and Wenjun Wu [6,*]

[1] School of Automation and Intelligence, Beijing Jiaotong University, Beijing 100044, China; 10311@bjtu.edu.cn (Q.W.)
[2] School of Statistics and Data Science, Capital University of Economics and Business, Beijing 100070, China
[3] School of Software, Northwestern Polytechnical University, Xi'an 710072, China
[4] Software Engineering Technology Office, Chongqing Chang'an Wangjiang Industrial Group Co., Ltd., Jiangbei District, Chongqing 401120, China
[5] Xi'an Aeronautics Computing Technique Research Institute, Aviation Industry Corporation of China, Xi'an 710065, China
[6] School of Artificial Intelligence, Beihang University, Beijing 100191, China
* Correspondence: wwj09315@buaa.edu.cn

Abstract

Although many researchers use pre-trained models to better solve downstream tasks, further exploration of more effective pre-training methods remains necessary, especially for multi-modal pre-training where high-quality training data is more difficult to obtain. This work aims to improve the knowledge-learning performance in multi-modal pre-training. Some researchers focus on injecting entity knowledge into language pre-trained models based on masked entity model (MEM) training, which masks entities randomly and lets the model recover. These methods cannot guarantee good performance due to the lack of consideration of which entities are more valuable for learning. Moreover, in multi-modal training data, some entities may be unrelated to visual content. In this work, for the vision-language pre-trained model, we propose a **M**asked **E**ntity **M**odel pre-training method based on **Active** learning (ActiveMEM). It is designed to actively mask important and informative entities—those that are both informative and uncertain—for the model to recover, thereby encouraging it to extract more valuable knowledge from the data. The proposed method is evaluated using three pre-training datasets and four downstream datasets, and the experimental results demonstrate the effectiveness of our method.

Keywords: vision-language pre-training; active learning; masked entity model; knowledge-enhanced pre-training; multi-modal representation learning

Check for updates

Academic Editor: Fabrizio Messina

Received: 29 April 2026
Revised: 7 June 2026
Accepted: 17 June 2026
Published: 22 June 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

With the introduction of pre-trained models [1,2], masked language model (MLM) pre-training has rapidly advanced the frontier of related works [3,4], which let models predict masked tokens given the remaining unmasked text. In many real-world scenarios, understanding and reasoning about entities (e.g., people, places, or objects) is crucial for tasks such as question answering and information retrieval. Consequently, pre-trained models need to have a good entity understanding ability. Research efforts have been devoted to injecting entity knowledge into models based on MEM pre-training [5], which lets models predict masked entities given the remaining unmasked text. However, these

methods are often used for language models rather than multi-modal models. Related research [6–8] often randomly masks entities for model recovery, without considering context and pre-training status. Limited research has been conducted on effectively integrating textual entity knowledge into multi-modal pre-trained models, despite its significance for real-world applications.

At present, high-quality multi-modal pre-training data is increasingly difficult to obtain, so it is necessary to study how to better mine pre-training data to improve knowledge-learning efficiency. MEM pre-training helps the model learn entity knowledge by memorizing the corresponding context. Existing methods usually adopt the random masking strategy in which different entities are masked with equal probability throughout the entire pre-training [2,3]. This strategy has obvious problems in the multi-modal setting where some entities are not the focus of attention of the visual content, or even irrelevant to the visual content, resulting in inefficient learning of entity knowledge. To acquire knowledge more effectively, the model is expected to recover important entities in the multi-modal setting, which can be determined based on the mutual information between the entities and the multi-modal context.

Moreover, the knowledge-learning efficiency of the model may be impacted by the pre-training status, which represents the model's cognitive state and changes accordingly as pre-training proceeds. Therefore, dynamically masking entities according to the model state can further tap the potential of knowledge-learning. Specifically, the knowledge-learning approach should be adjusted according to the pre-training status, in which the mask target should dynamically focus on the entities that the model is most uncertain about. In summary, compared with traditional strategies, it is necessary to dynamically mask entities for recovery according to the multi-modal context and the model state.

In the traditional scenario, active learning [9] aims to select informative samples for labeling to help the model achieve higher performance under a limited labeling budget. In MLM of self-supervised learning, masking can be seen as a kind of auto-labeling, and masking different objects can generate various training samples, where active learning can be used to obtain informative samples by selecting important objects for masking. Specifically, the active learning method can be implemented to dynamically mask the important and informative entities according to the multi-modal context and pre-training status. Rather than increasing the amount of pre-training data and model size to achieve new state-of-the-art performance, our work explores how to better mine the data to improve pre-training effectiveness. To the best of our knowledge, there has been little research on how to effectively incorporate active learning into the pre-training stage of models. This aligns with the way humans learn, as people typically acquire knowledge in an active, selective manner rather than passively absorbing all information.

Recent multi-modal pre-training methods often focus on vision-language pre-training. To improve the learning effectiveness of entity knowledge for enhanced multi-modal understanding, this work proposes an entity-centric active learning method, ActiveMEM. This approach selectively masks important and relevant entities, allowing the model to recover them, thereby facilitating more effective utilization of pre-training data. Specifically, we formulate the MEM pre-training based on mutual information theory and determine the important entities that are more relevant to the context based on cross-modal attention. Then, to further select the interesting entities from the important entities, we use the prediction loss of each important entity to determine which are more difficult and have greater learning value, and mask these entities for recovery. In our experiments, compared to other baseline methods, ActiveMEM enables the pre-trained model to achieve higher performance on downstream tasks using the same training data within a classic multi-modal model architecture. This demonstrates that ActiveMEM can mask entities in the

multi-modal sample that are both substantial and surprising for learning knowledge. The main contributions of this paper are summarized as follows:

- To improve pre-trained model performance, recent efforts often focus on increasing the amount of pre-training data and model size. However, there is still a need to investigate more efficient pre-training methods. This work explores how to better mine the multi-modal data for pre-training.
- A novel method is proposed for improving the learning effectiveness of entity knowledge in vision-language pre-training. To the best of our knowledge, few studies have explored how to effectively apply active learning in the pre-training stage.
- To demonstrate that the proposed method is more effective, we pre-train the model on three different pre-training datasets respectively, and evaluate each pre-training model on four downstream tasks.

The remainder of this paper is structured as follows. Section 2 presents previous related work. Section 3 describes the proposed method, ActiveMEM. Section 4 presents sufficient experiments using three different pre-training datasets and four downstream datasets. Section 5 discusses the results, and Section 6 presents conclusions and future directions.

2. Related Work

2.1. Knowledge-Enhanced Pre-Training

More and more research efforts focus on knowledge-enhanced pre-training. Most studies adopt the MEM pre-training to inject knowledge into the pre-trained model. For example, Baidu-ERNIE [5] is a method encouraging pre-trained language models to learn entity knowledge by predicting masked entity mentions. BERT-WWM [10] adopts whole-word masking rather than single-character masking for pre-training BERT on Chinese data. EaE [11] is a method to learn entity knowledge directly from text and use the memorized knowledge for prediction. However, all these methods focus on the uni-modal setting and how to set the objective function to learn knowledge, rather than how to improve knowledge-learning effectiveness by better utilizing the data. ERNIE-ViL [7] masks objects based on scene graphs to help better learn knowledge in the vision-language scenario, but it does not consider using the model state to help determine masking objects. The knowledge-learning efficiency of the model may be impacted by pre-training status; it is necessary to dynamically mask entities for recovery according to the model state.

2.2. Pre-Training Efficiency

Recently, some works [12,13] have studied optimal settings for better MLM. For example, carefully considered masked units like span or entity are adopted throughout the pre-training stage [4,5,14]. Li and Zhao [14] show that such uniform masking minimizes its training objective by latching onto shallow local signals, resulting in pre-training inefficiency and sub-optimal performance on downstream tasks. Bitton et al. [15] propose alternative masking methods that better utilize the image during pre-training. However, the proposed methods do not consider model state to determine masking units. Zellers et al. [6] propose a method named Merlot to use attention weights from a language-only transformer to help mask important words that are mentioned multiple times, which is not designed to mask key entities for the corresponding multi-modal context and does not consider masking the entities with greater learning value.

In addition, most MLM methods adopt time-invariant masking, which cannot adapt to the changeable process of model pre-training. Therefore, Kang et al. [16] propose a method to automatically generate domain- and task-adaptive maskings of the given text for pre-training of the uni-modal model. Yang et al. [17] explore the influence of masked content and masking ratio in MLM pre-training and propose a time-variance MLM pre-

training method. However, these methods do not consider how to apply active learning to determine the mask units according to the model state in the multi-modal setting.

On the other hand, recent studies have demonstrated the importance of robust feature learning [18] and effective cross-modal fusion for complex visual understanding tasks [19]. In contrast, our work focuses on adaptive entity masking during pre-training rather than designing new fusion architectures.

2.3. Active Learning for Pre-Trained Models

Active learning methods can help to select training samples based on the model state. Recent research efforts on pre-trained models mostly focus on using active learning to reduce the amount of annotated data required to fine-tune on downstream tasks. For example, Dor et al. [20] prove that the active learning method could make the model achieve higher performance in the downstream task at the same annotation cost, and this performance advantage is more obvious when the data is unbalanced. Bengar et al. [21] train the task layer using active learning to reduce the annotated samples required to fine-tune the downstream task. In addition, Rotman and Reichart [22] consider the problem of multi-task active learning for pre-trained Transformer-based models. However, there is a lack of research on how to effectively apply active learning to the pre-training stage, and little work has investigated how to jointly leverage cross-modal relevance and model uncertainty to guide entity masking during vision-language pre-training.

3. Method

3.1. Problem Formulation

In the vision-language setting, each data sample includes an image and a text sequence, in which the text sequence contains multiple entities. For the MEM pre-training task, the target is to recover the masked entities given the surrounding context. We can formulate the problem as follows:

$$\min_{\theta} L(\theta; C) = \min_{\theta} \sum_{c \in C} -\log p_{\theta}(s^{*}|c) \tag{1}$$

where s^{*} denotes the masked entities, c denotes the surrounding context, and θ denotes the model parameters. Figure 1 gives the framework of our proposed method. In our active learning framework, different from randomly selecting the entities to mask, our method dynamically selects the important and interesting entities according to the model state.

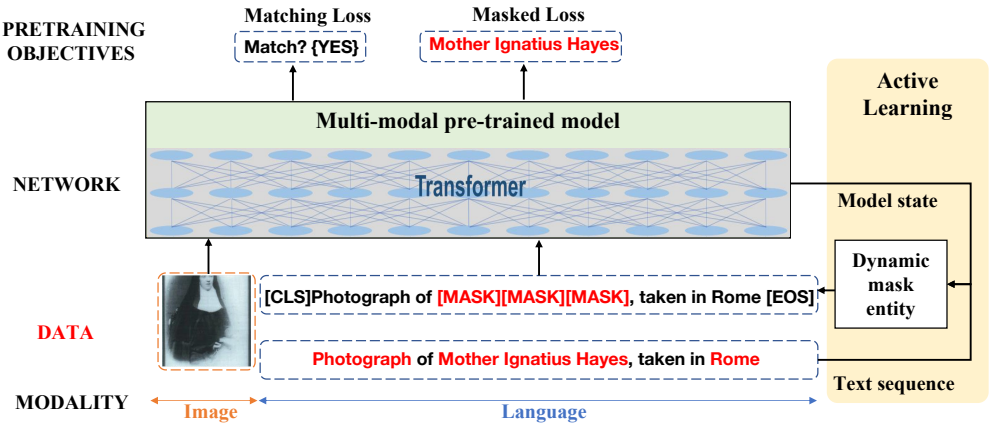


Figure 1. The framework of ActiveMEM. To study how to apply active learning to the masking-based multi-modal pre-training, we adopt METER [23] as the model framework, which is a benchmark framework of the vision-language model.

3.2. Information-Theoretic Framework for MEM

The objective of MEM pre-training is to predict masked entities given the surrounding context. We first formulate MEM pre-training from an information-theoretic perspective. Specifically, the mutual information between the masked entities s^* and the surrounding context c can be defined as follows:

$$I(c; s^*) = H(s^*) - H(s^*|c) \tag{2}$$

where $H(s^*)$ is the marginal entropy of s^*, and $H(s^*|c)$ is the conditional entropy of s^* conditioned on c. Mutual information measures the dependency between s^* and c. In other words, given the masked entities s^* and the corresponding context c, it can be understood as how much knowing c reduces the uncertainty in s^*. In the pre-training, we need to maximize their mutual information $I(c; s^*)$, so that masked entities can be predicted given the context. Referring to the mutual information maximization principle [24], existing pre-training tasks can be viewed as maximizing the lower bound of mutual information. Specifically, according to InfoNCE [25] (Information Noise-Contrastive Estimation), the following lower bound can be obtained

$$I(c; s^*) \geq E_{q(S)}\left[\log \frac{f_\theta(c; s^*)}{\sum_{s \in S} f_\theta(c, s)}\right] + \log |S| \tag{3}$$

where f_θ is a function that scores whether the context c and the masked content s^* are a positive pair. $q(S)$ is the distribution of S, and S is a set that contains s^* and $|S| - 1$ negative samples. For the masked loss in the pre-training, the optimization objective is to minimize the cross-entropy loss of the masked content given context, which can be approximated to maximize the lower bound of $I(c; s^*)$ [26].

However, in visual-language pre-training, the context includes both an image and a text sequence after masking the entities. Some entities in the text are not the focus of attention of the visual content, or even irrelevant to the visual content, where masking these entities and training the model to maximize their mutual information with the context actually has a negative impact on learning effectiveness. Therefore, it is necessary to find anchor points in different modalities that are associated. In other words, finding and masking the key entities s^*_{key} that are related to the visual content helps to better maximize the mutual information.

On the other hand, only the key entities s^*_{key} can have a bridge effect between different modalities, where MEM pre-training can obtain a v-structure dependency "$c_{text} \rightarrow s^*_{key} \leftarrow c_{image}$", leading to interaction information. Although the data are different in different modalities, based on the entities s^*_{key} related to the image content, the information bottleneck principle can transform the cross-modal structural similarity into isomorphic representations because these entities encourage the correlation between cross-modal contents, which helps to better learn entity knowledge.

In our work, we obtain the entities s^*_{key} using cross-modal attention by averaging the cross-attention scores across all layers. The entities with higher cross-modal attention scores are more attended to by the visual modality, meaning they are more relevant to the image content. These entities are more likely to be semantically relevant to the visual content and therefore provide a useful proxy signal for estimating entity importance. Masking these entities and training the model to maximize their mutual information with the context can impart the greatest change to the current model, which can help the model train better in the multi-modal setting. Instead of directly estimating mutual information, cross-modal attention serves as an efficient proxy for image-entity relevance. Highly attended entities drive stronger cross-modal interactions, providing richer learning signals.

3.3. Active Learning for MEM Pre-Training

Based on the analysis in the previous section, this section introduces an active learning method, ActiveMEM, that selects entities according to the multi-modal context and model state, which helps to mask important and interesting entities to obtain training samples for MEM pre-training. Specifically, ActiveMEM first selects $\lambda \times k$ most important entities as a candidate set E_C ($|E_C| = \lambda \times k$, $\lambda > 1$), and then selects only k interesting entities from E_C to mask for learning. λ is a hyperparameter that determines the number of candidate entities, and we give an analytical study of λ in Section 4.6.

To select the important entities, ActiveMEM selects entities based on cross-modal attention according to Section 3.2. The $\lambda \times k$ entities with the highest cross-modal attention are selected to form the candidate set. Specifically, an entity can be denoted as $e = (token_1, \ldots, token_J)$, where J is the number of tokens in e. The attention score of the entity e is calculated as the average attention of all its tokens.

On the other hand, at the beginning of pre-training, all entities are unfamiliar to the model. As training proceeds, the relative difficulty of entities changes as the model parameters are updated. Therefore, it may be unnecessary to keep masking "easy" entities that the model can already predict with high confidence, since they contribute little additional information for learning. To better utilize training capacity, the goal of ActiveMEM is to prioritize "difficult" entities that are harder for the model to predict, which are often more interesting for model learning. Specifically, given the masked entity candidates and their surrounding context, we can obtain the model outputs and calculate the loss of each entity. The prediction loss is related to the model uncertainty regarding a candidate entity. From an information-theoretic perspective, larger prediction loss indicates higher residual uncertainty under the current model state, suggesting greater potential information gain after learning. Then, only k entities with the highest loss are selected and masked for MEM pre-training.

In our MEM pre-training, the active learning process runs multiple steps of data preparation and model training. At the beginning of every pre-training step, for the text sequence in each input, our method selects $\lambda \times k$ important entities with the largest cross-attention as a candidate set E_C. The slots within these entities are then replaced with the `[MASK]` token, and a forward pass is performed, allowing us to compute the losses associated with these entities. From this, we select the top k entities with the highest losses from E_C. It is worth noting that this process incurs minimal computational cost, as the forward pass is performed via backpropagation during the model's inference phase. Thirdly, the model is trained to predict the original tokens of these entities, and the model parameters θ are updated. Specifically, the workflow of processing each input for model pre-training is shown in Figure 2.

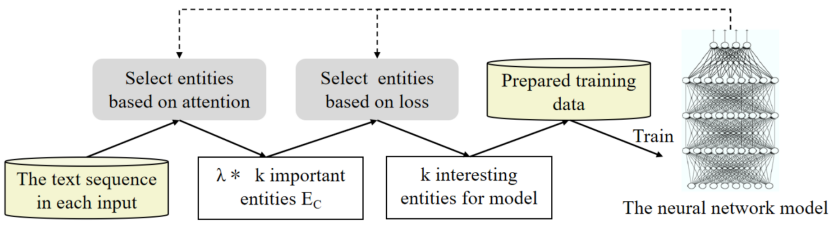


Figure 2. Workflow of processing each input for model pre-training.

4. Experiments

In this section, we pre-train the model on three different pre-training datasets and evaluate it on four downstream tasks, including knowledge-based tasks. In Section 4.4, we compare ActiveMEM with several baseline methods. Then, we evaluate the effect of the

specific design choices made for ActiveMEM through ablation experiments in Section 4.5, and we give an analysis of hyperparameter λ in Section 4.6 and hyperparameter k in Section 4.7.

4.1. Pre-Training Setting

To avoid the influence of other factors in the experiment and only study the effectiveness of applying active learning to the masking-based pre-training objective, it is necessary to adopt a benchmark multi-modal model framework. In our experiments, we directly adopt the benchmark multi-modal model architecture METER [23] and its recommended setup, where CLIP-VIT [27] is adopted as the Vision Encoder and RoBERTa [3] is adopted as the text encoder. The learning rates for the bottom and top layers are set to 1×10^{-5} and 5×10^{-5}, respectively. The warm-up ratio is set to 10%, and the learning rate is linearly decayed after 10% of the total pre-training steps. The model is trained using 8 NVIDIA V100 GPUs, each with 32 GB of memory, with a batch size of 8 per GPU. Referring to Dou et al. [23], the model is pre-trained with masked language modeling (MLM) and image-text matching (ITM) objective functions. In our ActiveMEM method, we replace the masked language modeling (MLM) with the proposed masking strategy.

Since our experiments involve multiple pre-training datasets containing texts of varying lengths, with some texts containing fewer entities, we set $k = 1$ to handle cases where only a single entity is present. In the majority of instances, the texts contain multiple entities. In addition, we set $\lambda = 2$ in our ActiveMEM method, which achieves better performance according to Section 4.6. When a text contains only a single entity, we restrict the selection to exactly one candidate entity.

To verify the generalization of ActiveMEM, we use three datasets with different characteristics to pre-train the model: Wikipedia-based Image Text Dataset (WIT) [28], Microsoft COCO (COCO) [29], and Conceptual Captions (CC3M) [30]. For these pre-training datasets, we annotate the entities of each sample based on entities from Wikipedia. These entities contain organizations, persons, locations, etc., which can be denoted with a proper name and have a physical existence. Specifically, we utilized Wikipedia as a unified, large-scale repository to annotate WIT, COCO, and CC3M. Candidate entities were automatically extracted via named entity recognition and matched to Wikipedia entries, discarding ambiguous or low-confidence matches. Despite distinct caption styles and entity distributions across datasets, applying this uniform procedure ensures cross-dataset consistency. Furthermore, since our masking strategy relies on intra-sample relative importance rankings, the selection process is expected to be less sensitive to moderate annotation noise. In the MEM pre-training, the annotated entities are used for masking based on the corresponding strategy.

4.2. Downstream Tasks

After pre-training with three datasets, we have three different pre-trained models. We evaluate each model on four downstream representative tasks, including VQA-V2 [31], OK-VQA [32], A-OKVQA [33], and SNLI-VE [31]. The setup details of model fine-tuning in VQA and VE tasks can be referred to Dou et al. [23]. Each experiment is repeated three times and the results are averaged for the report.

4.3. Baselines

We apply active learning to the pre-training stage in our method. To isolate the effect of the proposed active masking strategy, all baselines are implemented under the same METER architecture and training configuration. Therefore, we primarily compare against representative masking policies rather than methods involving additional architectural modifications, curriculum-learning frameworks [34,35], or hard-example mining mechanisms [36], etc. Specifically, we compare our method with the following related methods:

Token-Random: The random masking strategy introduced in BERT.

ContentWords-Random: Masks exactly one word in each sentence, increasing the probability of masking content words to 80%.

Whole-Random: A random masking strategy that masks the entire word instead of the token, introduced by the authors of BERT.

Span-Random: A random masking strategy that selects multiple consecutive tokens.

4.4. Experimental Results

As shown in Table 1, the three groups of experiments adopt different pre-training datasets respectively. Under each pre-training dataset, we give the results of different methods, including Token-Random, ContentWords-Random, Whole-Random, Span-Random, and ActiveMEM under different downstream tasks. Note that not all results are better than the results of state-of-the-art models because we only use a benchmark multi-modal model framework and use one dataset for pre-training in each experiment. Recent efforts often focus on increasing the amount of pre-training data and model size to achieve new state-of-the-art performance; our work explores how to better mine the pre-training data.

Specifically, as demonstrated in Table 1, baseline methods cannot achieve better performance because they cannot help the model capture more entity knowledge, which subsequently hinders their efficacy in downstream tasks. In the case of the WIT pre-training dataset, all evaluated methods exhibit superior performance. This finding suggests that pre-training on entity-rich corpora is inherently more effective for downstream task adaptation. Furthermore, the adaptive nature of our proposed method, ActiveMEM, consistently outperforms all baselines across all downstream tasks. This superiority underscores ActiveMEM as the optimal masking policy for effectively injecting entity knowledge. In other words, ActiveMEM achieves consistently competitive performance across different downstream tasks and pre-training datasets; while the degree of improvement is task-dependent, the significance of our method is validated, confirming the practical viability of our method.

Table 1. Performance of various masking strategies on four downstream tasks. The experiment for each downstream task is repeated three times and the results are averaged to report mean values alongside corresponding standard deviations. Best results are shown in bold.

Dataset	Task	Token Random	ContentWord Random	Whole Random	Span Random	ActiveMEM
WIT	VQA-V2	73.3 ± 0.29	72.5 ± 0.26	72.0 ± 0.36	73.5 ± 0.32	$\mathbf{74.9 \pm 0.28}$
	OK-VQA	38.5 ± 0.26	39.0 ± 0.23	39.1 ± 0.24	39.1 ± 0.27	$\mathbf{39.8 \pm 0.18}$
	A-OKVQA	37.4 ± 0.25	37.2 ± 0.19	37.2 ± 0.22	37.3 ± 0.28	$\mathbf{38.2 \pm 0.23}$
	SNLI-VE	79.7 ± 0.24	79.3 ± 0.23	79.4 ± 0.26	79.2 ± 0.30	$\mathbf{80.5 \pm 0.24}$
COCO	VQA-V2	72.8 ± 0.30	71.8 ± 0.29	71.5 ± 0.37	73.1 ± 0.30	$\mathbf{74.0 \pm 0.27}$
	OK-VQA	37.2 ± 0.28	37.6 ± 0.24	37.7 ± 0.25	37.9 ± 0.26	$\mathbf{38.9 \pm 0.19}$
	A-OKVQA	37.2 ± 0.19	36.4 ± 0.25	36.5 ± 0.30	37.1 ± 0.24	$\mathbf{38.0 \pm 0.25}$
	SNLI-VE	79.1 ± 0.25	78.7 ± 0.26	79.1 ± 0.27	78.6 ± 0.32	$\mathbf{79.9 \pm 0.24}$
CC3M	VQA-V2	72.5 ± 0.31	72.0 ± 0.32	71.6 ± 0.41	72.7 ± 0.34	$\mathbf{73.8 \pm 0.30}$
	OK-VQA	37.3 ± 0.23	37.4 ± 0.25	37.5 ± 0.30	37.6 ± 0.26	$\mathbf{38.6 \pm 0.23}$
	A-OKVQA	36.1 ± 0.24	35.1 ± 0.27	35.2 ± 0.31	36.1 ± 0.29	$\mathbf{37.2 \pm 0.28}$
	SNLI-VE	79.0 ± 0.28	78.8 ± 0.23	78.9 ± 0.26	78.8 ± 0.31	$\mathbf{79.8 \pm 0.26}$

4.5. Ablation Experiments

In this section, we investigate the effectiveness of each design choice through an ablation study, including the "w/o attention" setting and the "w/o loss" setting. As shown in in Table 2, our method improves the model performance over all downstream tasks by comparing "w/o attention" and "w/o loss" settings, where the improvement amplitude varies for different tasks.

First, we evaluate the effectiveness of selecting the important entities based on attention. The results of ActiveMEM "w/o attention" show that the attention-based strategy is

a critical component of ActiveMEM, as "w/o attention" leads to poor performance. We can observe that "w/o attention" causes performance to drop more drastically than "w/o loss". This is because the attention-based strategy can effectively help to find the important entities by considering the multi-modal context, which learns vision-entity alignment and is more important for mining the data.

Second, we evaluate the effectiveness of selecting the interesting entities based on loss. The results in Table 2 show that our method obtains better performance than its counterpart without loss. This result verifies that selecting interesting entities based on loss can also help the model to learn entity knowledge. The results show that training without either component hurts performance, demonstrating the necessity of selecting the important and interesting entities simultaneously. In addition, without selecting the interesting entities based on the prediction loss, the method can still outperform the baseline methods without introducing any additional time cost.

Table 2. Ablation results. Best results are shown in bold.

Dataset	Task	ActiveMEM w/o Attention	ActiveMEM w/o Loss	ActiveMEM
WIT	VQA-V2	74.3	74.5	**74.9**
	OK-VQA	39.3	39.4	**39.8**
	A-OKVQA	37.5	37.6	**38.2**
	SNLI-VE	79.5	79.8	**80.5**
COCO	VQA-V2	73.3	73.5	**74.0**
	OK-VQA	38.2	38.4	**38.9**
	A-OKVQA	37.3	37.4	**38.0**
	SNLI-VE	79.0	79.2	**79.9**
CC3M	VQA-V2	73.1	73.3	**73.8**
	OK-VQA	37.9	38.1	**38.6**
	A-OKVQA	36.4	36.7	**37.2**
	SNLI-VE	79.0	79.1	**79.8**

4.6. Analysis of Hyperparameter λ

ActiveMEM first selects $\lambda \times k$ most important entities as a candidate set E_C, and then selects only k interesting entities from E_C to mask for learning. Note that $k = 1$ in our experiments. As shown in Table 3, 'ActiveMEM $\lambda = 2$' is better than 'ActiveMEM $\lambda = 3$', illustrating that introducing too many candidate entities can have a negative impact on performance. Therefore, $\lambda = 2$ is adopted in other experiments.

Table 3. Experimental results with different values of hyperparameter λ. Best results are shown in bold.

Dataset	Task	ActiveMEM $\lambda = 3$	ActiveMEM $\lambda = 2$
WIT	VQA-V2	74.4	**74.9**
	OK-VQA	39.4	**39.8**
	A-OKVQA	37.5	**38.2**
	SNLI-VE	79.7	**80.5**
COCO	VQA-V2	73.2	**74.0**
	OK-VQA	38.4	**38.9**
	A-OKVQA	37.5	**38.0**
	SNLI-VE	79.2	**79.9**
CC3M	VQA-V2	73.2	**73.8**
	OK-VQA	38.1	**38.6**
	A-OKVQA	36.6	**37.2**
	SNLI-VE	79.1	**79.8**

4.7. Analysis of Hyperparameter k

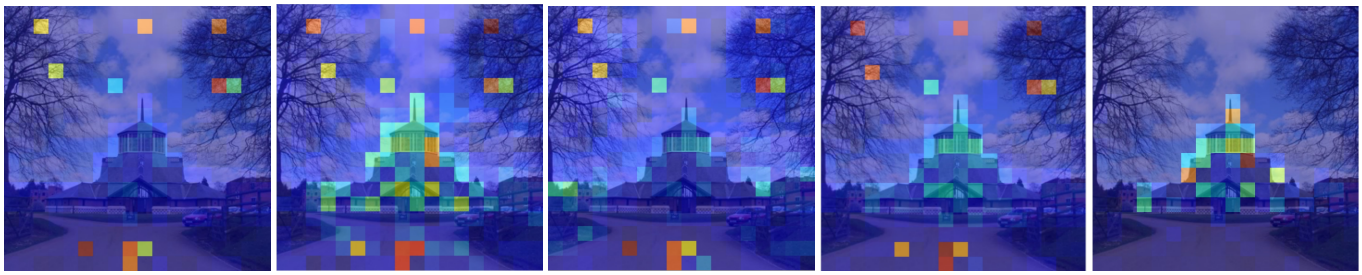
In the entity-rich pre-training dataset WIT, there are multiple entities in the text of each sample. In this section, we give an analysis of the effect of hyperparameter k on performance. Following the analysis of Section 4.6, we set $\lambda = 2$ in the experiments. As shown in Table 4, 'ActiveMEM $k = 3$' is better than 'ActiveMEM $k = 2$' and 'ActiveMEM $k = 1$', illustrating that masking more entities can help achieve better performance, although performance improvement is limited.

Table 4. Experimental results with different values of hyperparameter k.

Dataset	Downstream Task	ActiveMEM $k = 1$	ActiveMEM $k = 2$	ActiveMEM $k = 3$
WIT	VQA-V2	74.9	75.0	75.0
	OK-VQA	39.8	39.8	39.9
	A-OKVQA	38.2	38.3	38.4
	SNLI-VE	80.5	80.6	80.6

4.8. Qualitative Analysis

For qualitative analysis, as shown in Figure 3, we provide a data example from the WIT dataset. Specifically, for an entity 'Abbey' in the caption of the data example, we provide the visualization of cross-modal attention in the corresponding image, generated by averaging the cross-attention scores across all layers. Based on the cross-modal attention of the entity 'Abbey', we can successfully obtain the correct region in the image, even though understanding the entity in WIT requires more background knowledge. In contrast, other methods, especially Token-Random, yield less satisfactory results, with attention that is scattered and unfocused. The results show that ActiveMEM can help the model acquire a better ability of entity understanding in the cross-modal setting. We also observe a similar trend across WIT, COCO, and CC3M, where entities associated with salient objects appear more likely to focus attention on relevant areas, while weakly grounded entities tend to exhibit more diffuse patterns. Note that the use of cross-modal attention is an indicator of image-entity relevance, while recognizing that attention alone may not fully capture semantic alignment.



(**a**) Token-Random (**b**) ContentWord-Random (**c**) Whole-Random (**d**) Span-Random (**e**) ActiveMEM

Figure 3. For the entity 'Abbey' in the caption of a data example, we provide the visualization of cross-modal attention in the corresponding image. It is obvious that our method, ActiveMEM, can effectively align the entity to the corresponding image region. The results show that ActiveMEM can help the model acquire a strong ability of entity understanding in the cross-modal setting. In contrast, other methods, especially Token-Random, yield less satisfactory results, with attention that is scattered and unfocused.

5. Discussion

While recent approaches in multi-modal pre-training primarily focus on scaling up model size or increasing the volume of pre-training data to achieve state-of-the-art performance, our work takes a different direction. Rather than merely enlarging data or models,

we aim to better mine and utilize the existing multi-modal data for pre-training. To this end, we propose a novel masking-based pre-training method grounded in active learning principles, which can be used in the pre-training of other multi-modal models.

To the best of our knowledge, existing active learning studies mainly focus on selecting labeled samples for downstream task adaptation. In contrast, ActiveMEM formulates entity masking as an active sample selection problem during pre-training and introduces a dynamic masking mechanism that jointly considers cross-modal relevance and model uncertainty. Therefore, our contribution is not merely the combination of two criteria, but the introduction of an active-learning paradigm into multi-modal pre-training.

Beyond active learning, several recent vision-language pre-training studies have explored complementary strategies, such as curriculum learning and hard example mining, to improve training efficiency and representation quality. These approaches mainly focus on scheduling strategies or training curricula, whereas ActiveMEM focuses specifically on active entity selection during MEM pre-training. Since these techniques address different aspects of the learning process, they are not mutually exclusive and could potentially be combined in future research.

In addition, this work explores how to inject entity knowledge into multi-modal pre-trained models more effectively. It is an exploratory investigation into the application of active learning to the pre-training stage. Our method can achieve the best performance by using cross-attention and prediction loss to select the important and interesting entities for mask-and-recover. Note that although direct mutual information estimation is theoretically appealing, it can be challenging to apply in vision-language pre-training due to instability in high-dimensional multi-modal spaces and sensitivity to training settings.

On the other hand, the additional forward pass introduced by ActiveMEM for entity selection does incur some computational overhead. However, since the selection process is performed via backpropagation during the model's inference phase, it does not significantly increase training costs. More importantly, ActiveMEM is designed to maximize performance gains using a fixed training dataset, enabling more informative masking and improving learning efficiency compared to conventional random masking methods. Furthermore, ActiveMEM does not introduce additional trainable parameters or auxiliary networks, as the attention scores and prediction losses used for entity selection are naturally available within the standard training pipeline. Therefore, the computational overhead of ActiveMEM remains relatively small in practical vision-language pre-training scenarios.

6. Conclusions

For multi-modal pre-training, this work proposes a masked entity model (MEM) pre-training method based on active learning. It is designed to actively mask important and interesting entities based on the multi-modal context and model state for model recovery. In our experiments, we demonstrate the effectiveness of our method across four downstream tasks. This work represents an initial exploration of applying active learning during the pre-training stage. In future work, we plan to extend this approach to models incorporating additional modalities to further enhance multi-modal representation learning. Additionally, we aim to investigate the scalability and generalization of ActiveMEM across a broader range of tasks and domains.

Author Contributions: Conceptualization: Q.W.; Methodology: Q.W. and S.Z.; Software: Q.W.; Validation: Q.W., S.Z. and X.G.; Formal analysis: J.A.; Investigation: Y.J. and J.L.; Resources: Y.J.; Data curation: B.S.; Writing—original draft preparation: Q.W.; Writing—review and editing: S.Z., B.S., X.G., J.A., C.F., Y.J., J.L. and W.W.; Visualization: S.Z. and X.G.; Supervision: C.F. and W.W.; Project administration: C.F. and W.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Shanghai Aerospace Science and Technology Innovation Foundation, grant numbers SAST2021-054 and SAST2023-005; the Aviation Science Fund, grant numbers 20240001053008 and 201907053004; and the Beijing Natural Science Foundation, grant number L252146.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The pre-training datasets used in this study are publicly available: WIT (https://github.com/google-research-datasets/wit (accessed on 16 June 2026)), COCO (https://cocodataset.org (accessed on 16 June 2026)), and CC3M (https://ai.google.com/research/ConceptualCaptions (accessed on 16 June 2026)). The downstream task datasets are available at: VQA-V2 (https://visualqa.org/ (accessed on 16 June 2026)), OK-VQA (https://okvqa.allenai.org/ (accessed on 16 June 2026)), A-OKVQA (https://github.com/allenai/aokvqa (accessed on 16 June 2026)), and SNLI-VE (https://github.com/necla-ml/SNLI-VE (accessed on 16 June 2026)).

Acknowledgments: The authors thank the reviewers for their constructive comments and suggestions.

Conflicts of Interest: Author Yuanjun Jing was employed by Chongqing Chang'an Wangjiang Industrial Group Co., Ltd. and author Junxian Li was employed by Aviation Industry Corporation of China. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

MLM	Masked Language Model
MEM	Masked Entity Model
VLP	Vision-Language Pre-training
ITM	Image-Text Matching
VQA	Visual Question Answering
VE	Visual Entailment
WIT	Wikipedia-based Image Text
COCO	Common Objects in Context
CC3M	Conceptual Captions 3M

References

1. Li, Y.; Ma, Y.; Yan, S.; Zhang, C.; Liu, J.; Lu, J.; Xu, Z.; Chen, M.; Wang, M.; Zhan, S.; et al. Model merging in pre-training of large language models. *Adv. Neural Inf. Process. Syst.* **2026**, *38*, 133668–133691.
2. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019 ; Volume 1 (Long and Short Papers), pp. 4171–4186.
3. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv* **2019**, arXiv:1907.11692.
4. Joshi, M.; Chen, D.; Liu, Y.; Weld, D.S.; Zettlemoyer, L.; Levy, O. Spanbert: Improving pre-training by representing and predicting spans. *Trans. Assoc. Comput. Linguist.* **2020**, *8*, 64–77.
5. Sun, Y.; Wang, S.; Li, Y.; Feng, S.; Chen, X.; Zhang, H.; Tian, X.; Zhu, D.; Tian, H.; Wu, H. Ernie: Enhanced representation through knowledge integration. *arXiv* **2019**, arXiv:1904.09223.
6. Zellers, R.; Lu, X.; Hessel, J.; Yu, Y.; Park, J.S.; Cao, J.; Farhadi, A.; Choi, Y. Merlot: Multimodal neural script knowledge models. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 23634–23651.
7. Yu, F.; Tang, J.; Yin, W.; Sun, Y.; Tian, H.; Wu, H.; Wang, H. Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. In Proceedings of the AAAI Conference on Artificial Intelligence, Online, 2–9 February 2021; Volume 35, pp. 3208–3216.
8. Kwon, G.; Cai, Z.; Ravichandran, A.; Bas, E.; Bhotika, R.; Soatto, S. Masked Vision and Language Modeling for Multi-modal Representation Learning. In Proceedings of the Eleventh International Conference on Learning Representations, Virtual, 25–29 April 2022.

9. Settles, B. *Active Learning Literature Survey*; Technical Report; University of Wisconsin-Madison Department of Computer Sciences: Madison, WI, USA , 2009.
10. Cui, Y.; Che, W.; Liu, T.; Qin, B.; Yang, Z. Pre-training with whole word masking for chinese bert. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 3504–3514. [CrossRef]
11. Févry, T.; Soares, L.B.; FitzGerald, N.; Choi, E.; Kwiatkowski, T. Entities as experts: Sparse memory access with entity supervision. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020; pp. 4937–4951.
12. Liu, H.; Li, C.; Li, Y.; Lee, Y.J. Improved baselines with visual instruction tuning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 26296–26306.
13. Edman, L.; Fraser, A. Mask and You Shall Receive: Optimizing Masked Language Modeling For Pretraining BabyLMs. In Proceedings of the First BabyLM Workshop, Suzhou, China, 5–9 November 2025; pp. 445–453.
14. Li, Y.; Zhao, H. Pre-training universal language representation. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Online, 1–6 August 2021; Volume 1, pp. 5122–5133.
15. Bitton, Y.; Elhadad, M.; Stanovsky, G.; Schwartz, R. Data Efficient Masked Language Modeling for Vision and Language. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2021, Punta Cana, Dominican Republic, 7–11 November 2021; pp. 3013–3028.
16. Kang, M.; Han, M.; Hwang, S.J. Neural Mask Generator: Learning to Generate Adaptive Word Maskings for Language Model Adaptation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020; pp. 6102–6120.
17. Yang, D.; Zhang, Z.; Zhao, H. Learning better masking for better language model pre-training. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Toronto, ON, Canada, 9–14 July 2023; Volume 1, pp. 7255–7267.
18. Liu, Y.; Wang, C.; Lu, M.; Yang, J.; Gui, J.; Zhang, S. From simple to complex scenes: Learning robust feature representations for accurate human parsing. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 5449–5462. [CrossRef] [PubMed]
19. Wang, S.; Wang, C.; Shi, C.; Liu, Y.; Lu, M. Mask-guided mamba fusion for drone-based visible-infrared vehicle detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5005712.
20. Dor, L.E.; Halfon, A.; Gera, A.; Shnarch, E.; Dankin, L.; Choshen, L.; Danilevsky, M.; Aharonov, R.; Katz, Y.; Slonim, N. Active learning for BERT: An empirical study. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020; pp. 7949–7962.
21. Bengar, J.Z.; van de Weijer, J.; Twardowski, B.; Raducanu, B. Reducing label effort: Self-supervised meets active learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 1631–1639.
22. Rotman, G.; Reichart, R. Multi-task active learning for pre-trained transformer-based models. *Trans. Assoc. Comput. Linguist.* **2022**, *10*, 1209–1228. [CrossRef]
23. Dou, Z.Y.; Xu, Y.; Gan, Z.; Wang, J.; Wang, S.; Wang, L.; Zhu, C.; Zhang, P.; Yuan, L.; Peng, N.; et al. An empirical study of training end-to-end vision-and-language transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 18166–18176.
24. Kong, L.; d'Autume, C.d.M.; Ling, W.; Yu, L.; Dai, Z.; Yogatama, D. A mutual information maximization perspective of language representation learning. *arXiv* **2019**, arXiv:1910.08350.
25. Oord, A.v.d.; Li, Y.; Vinyals, O. Representation learning with contrastive predictive coding. *arXiv* **2018**, arXiv:1807.03748.
26. Chi, Z.; Dong, L.; Wei, F.; Yang, N.; Singhal, S.; Wang, W.; Song, X.; Mao, X.L.; Huang, H.Y.; Zhou, M. InfoXLM: An Information-Theoretic Framework for Cross-Lingual Language Model Pre-Training. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online, 6–11 June 2021; pp. 3576–3588.
27. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 18–24 July 2021; pp. 8748–8763.
28. Srinivasan, K.; Raman, K.; Chen, J.; Bendersky, M.; Najork, M. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual, 11–15 July 2021; pp. 2443–2449.
29. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 740–755.
30. Sharma, P.; Ding, N.; Goodman, S.; Soricut, R. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018; Volume 1, pp. 2556–2565.

31. Song, H.; Dong, L.; Zhang, W.; Liu, T.; Wei, F. Clip models are few-shot learners: Empirical studies on vqa and visual entailment. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, Dublin, Ireland, 22–27 May 2022; Volume 1, pp. 6088–6100.
32. Marino, K.; Rastegari, M.; Farhadi, A.; Mottaghi, R. Ok-vqa: A visual question answering benchmark requiring external knowledge. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3195–3204.
33. Schwenk, D.; Khandelwal, A.; Clark, C.; Marino, K.; Mottaghi, R. A-okvqa: A benchmark for visual question answering using world knowledge. In Proceedings of the European Conference on Computer Vision; Springer: Berlin/Heidelberg, Germany, 2022; pp. 146–162.
34. Wang, Y.; Yue, Y.; Lu, R.; Han, Y.; Song, S.; Huang, G. Efficienttrain++: Generalized curriculum learning for efficient visual backbone training. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 8036–8055. [CrossRef] [PubMed]
35. Fu, Y.G.; Chen, X.; Xu, S.; Li, J.; Yao, X.; Huang, Z.; Wang, Y.M. GSSCL: A framework for Graph Self-Supervised Curriculum Learning based on clustering label smoothing. *Neural Netw.* **2025**, *181*, 106787. [PubMed]
36. Lin, C.; Ji, X.; Yang, Y.; Li, Q.; Zhao, Z.; Peng, Z.; Wang, R.; Fang, L.; Shen, C. Hard adversarial example mining for improving robust fairness. *IEEE Trans. Inf. Forensics Secur.* **2024**, *20*, 350–363. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.