Leveraging per Image-Token Consistency for Vision-Language Pre-training

Yunhao Gou[1,2*] Tom Ko[3] Hansi Yang[2] James Kwok[2] Yu Zhang[1,4†]
Mingxuan Wang[3]

[1]Southern University of Science and Technology, [2]Hong Kong University of Science and Technology
[3]ByteDance AI Lab, [4]Peng Cheng Laboratory

{ygou, hyangbw}@connect.ust.hk jamesk@cse.ust.hk
{tom.ko, wangmingxuan.89}@bytedance.com yu.zhang.ust@gmail.com

Abstract

Most existing vision-language pre-training (VLP) approaches adopt cross-modal masked language modeling (CMLM) to learn vision-language associations. However, we find that CMLM is insufficient for this purpose according to our observations: (1) Modality bias: a considerable amount of masked tokens in CMLM can be recovered with only the language information, ignoring the visual inputs. (2) Under-utilization of the unmasked tokens: CMLM primarily focuses on the masked token but it cannot simultaneously leverage other tokens to learn vision-language associations. To handle those limitations, we propose ***EPIC*** *(l****E****veraging* ***P****er* ***I****mage-Token* ***C****onsistency for vision-language pre-training). In* ***EPIC****, for each image-sentence pair, we mask tokens that are salient to the image (i.e., Saliency-based Masking Strategy) and replace them with alternatives sampled from a language model (i.e., Inconsistent Token Generation Procedure), and then the model is required to determine* **for each token** *in the sentence whether it is consistent with the image (i.e., Image-Token Consistency Task). The proposed EPIC method is easily combined with pre-training methods. Extensive experiments show that the combination of the EPIC method and state-of-the-art pre-training approaches, including ViLT, ALBEF, METER, and X-VLM, leads to significant improvements on downstream tasks. Our coude is released at https://github.com/gyhdog99/epic*

1. Introduction

Vision-language pre-training (VLP) [5, 12, 21, 29, 30, 33, 37] aims to learn multi-modal representations from large-scale image-text pairs. A pre-trained vision-language model (VLM) fine-tuned with only a small amount of labeled data has shown state-of-the-art performance in many downstream tasks such as visual question answering and image-text retrieval.

A primary concern in developing pre-training objectives for VLP models is how to learn better vision-language associations. In addition to coarse-grained approaches such as image-text matching/contrasting [10, 14, 27] that align concepts from two modalities at the sample level, fine-grained approaches such as cross-modal masked language/image modeling (CMLM/CMIM) [16, 19, 32] learn vision-language associations at the token-object level. For example, Fig. 1 shows a picture paired with the sentence "Blue and yellow hydrant on the grass". When the word "hydrant" is masked, in order to correctly recover the token, the model has to find the actual object in the image and associate it with the word "hydrant".

While effective, CMLM is insufficient for learning vision-language associations because of (1) modality bias; and (2) under-utilization of unmasked tokens. In vision-language understanding, modality bias refers to leveraging only one modality for training/inference and so cross-modal knowledge is not well explored [23]. We argue that modality bias exists in CMLM, and prevents the model from learning sufficient vision-language associations. Specifically, in the CMLM task, we expect to mask *salient*[1] tokens (such as "blue", "yellow", "fire-hydrant", and "grass") as shown in the left of Fig. 1. These tokens are informative for learning vision-language association because masking them enforces the model to find the answer from the visual modality. However, in practice, whether a token is salient is unknown as we only have access to image-sentence level annotations. Given a fixed and relatively small masking ratio (typically 15% in CMLM), we might end up masking tokens that are less informative. For example, as shown

[*] Work was done when the author interned at ByteDance AI Lab.
[†] The corresponding author.

[1]The definition of "saliency" is given in Sec. 4.4.

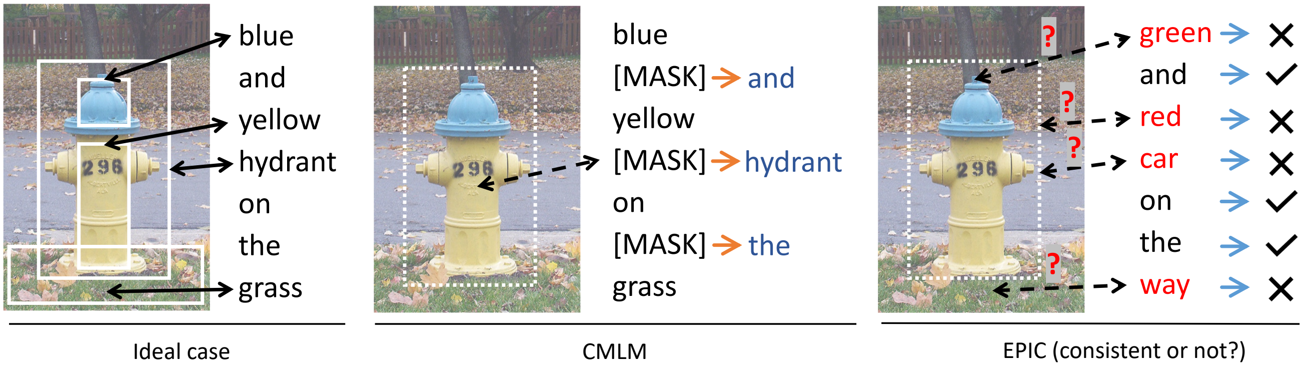


Figure 1. Illustrations of vision-language association learning. **Ideal case**: Fine-grained annotations (image regions and corresponding text tokens) are given, we can learn explicit associations (solid lines); **CMLM**: Without fine-grained annotations, we create supervision by masking, but this can be insufficient due to limited masking ratios and modality bias. **EPIC**: We find salient tokens and corrupt them to learn more associations. Both **CMLM** and **EPIC** learn implicit associations due to lack of region annotations.

in Fig. 1 (center), when "the" and "and" are masked, the model can predict these masked tokens with only language information. This thus is a form of modality bias as it circumvents using vision-language reasoning. Therefore, the modality bias can make CMLM insufficient to learn vision-language associations.

Another source of insufficiency in CMLM comes from the under-utilization of unmasked tokens. Similar to Masked Language Modeling (MLM) [6] in language pre-training, the CMLM loss is computed over masked tokens rather than all tokens in the sentence. As a result, learning of cross-modal association is possible only for the masked tokens but not for the remaining unmasked ones. For example, in Fig. 1, ideally, there are four associations (shown in black arrows) between text tokens and the corresponding regions, while there is only one association for CMLM. Therefore, CMLM cannot leverage all tokens (including the unmasked ones) for learning vision-language associations.

To expedite the learning of cross-modal associations in VLP, we propose **EPIC** (l**E**veraging **P**er **I**mage-Token **C**onsistency for vision-language pre-training). For each image-sentence pair, we mask tokens that are salient to the image (Saliency-based Masking Strategy) and make them "inconsistent"[2] with the image by replacing them with alternatives from a BERT-like language model (Inconsistent Token Generation Procedure). The model is then required to determine whether ***each token*** in the sentence is consistent with the image (Image-Token Consistency (ITC) Task). As this masks salient tokens and applies a language model to generate inconsistent tokens from them, the model has to refer to the visual modality to determine whether a token is inconsistent. Therefore, the modality bias problem can be alleviated. Moreover, we can make better use of the unmasked tokens for learning vision-language association as the ITC task requires the model to determine whether ***each token*** is consistent with the image.

The proposed EPIC method is easy to implement and widely applicable to a lot of vision-language model architectures. We demonstrate the effectiveness of our approach on various pre-training approaches, including ViLT [21], ALBEF [14], METER [8], and X-VLM [39], and observe significant improvements on downstream tasks. For example, on MSCOCO image-text retrieval, the proposed EPIC method achieves an absolute gain of 2.5% and 4.7% over METER and ViLT, respectively, in terms of the Recall@1 score. On visual reasoning tasks (e.g., NLVR2), the proposed method improves over ALBEF by 1.8% and X-VLM (the state-of-the-art within its model scale) by 1.3%. The proposed method also allows better generalization of pre-training models. For example, in zero-shot image-text retrieval, we improve X-VLM by 3.9% (COCO) and ViLT by 9.9% (Flickr30k).

2. Related Work

Learning Vision-Language Associations. In vision-language pre-training, a contrastive objective is often used to learn the coarse-grained associations between sentences and images [10, 14, 17, 27, 38]. However, models pre-trained with this objective give unsatisfactory results on vision-language reasoning tasks (such as VQA [1] and NLVR2 [31]) which require understanding fine-grained vision-language associations. Cross-Modal Masked Language/Image Modeling (CMLM/CMIM) [2, 16, 19, 20] and its variants are applied to learn fine-grained vision-language associations in a self-supervised manner. Besides CMLM/CMIM, one can further leverage annotated region-phrase (e.g., bounding boxes) image-text data to enable the model with more sophisticated vision-language reasoning

[2] A formal definition of (in)consistency tokens will be provided in Sec. 4.2.

abilities [7, 11, 15, 39].

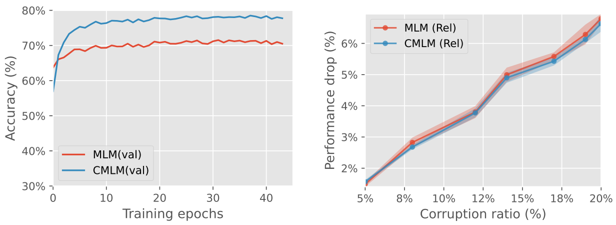
Cross-Modal Masked Language/Image Modeling. CMLM and CMIM are widely adopted as pre-training objectives in vision-language pre-training. They corrupt a token/patch of a sentence/image and train the model to reconstruct the original data with information from both modalities. LXMERT [32] found that loading BERT into the text encoder harms pre-training because the pre-trained BERT can have high CMLM accuracy and thereby circumventing the cross-modal reasoning process. Recently, similar to our observations, Bitton et al. [2] showed that roughly 50% of the masked tokens in CMLM are punctuation or stop-words, leading to sample inefficiency of CMLM. They propose a rule-based masking strategy to mask object words, content words, or words with high concreteness [3] in CMLM according to different downstream tasks, expecting such words to be more related to the visual input. Such a strategy lacks generalization ability and cannot scale to more downstream tasks.

Modality Bias in Vision-Language Understanding. In CMLM, predicting the masked token without referring to the visual modality can be seen as a consequence of modality bias towards language in vision-language understanding. Due to the existence of modality bias, the learning of vision-language associations is weakened. Similarly, modality bias also happens in visual question answering (VQA) [1] where the model is required to answer a question given a visual input. For example, simply answering "tennis" to the sport-related questions can achieve approximately 40% accuracy [23] on the VQA v1.0 dataset. To reduce such a bias, CF-VQA [23] proposes a counterfactual framework which directly subtracts the language-based predictions from the answers. In the proposed EPIC method, we mitigate the negative influence of language bias by allowing the model to learn more vision-language associations over a wider range of tokens.

3. Empirical Analysis

In this section, we empirically analyse the existence of modality bias and under-utilization of unmasked tokens in CMLM. Experimental details are in Appendix A.

Identifying Modality Bias. If there is a strong modality bias towards language in Cross-Modal Language Modeling (CMLM), a pure language model (LM) that is blind to the visual inputs can achieve comparable accuracy on the Masked Language Modeling (MLM) task. Otherwise, the LM will be outperformed by the vision-language model (VLM) by a large margin. Based on this, we run CMLM on a VLM and MLM on a LM, respectively. They share the same input sentences and masked tokens, while the former additionally receives visual inputs. As shown in Fig. 2a, though using visual modality is helpful to recover the original token (80% Acc.), a pure LM can already achieve an accuracy of 70%. Therefore, a modality bias towards language does exist in the training of CMLM.



(a) Accuracy gaps.

(b) Accuracy drops.

Figure 2. Left: CMLM/MLM accuracy of VLM/LM (y-axis) as the training process proceeds (x-axis). Right: relative accuracy drop of CMLM/MLM (y-xais) as we corrupt more unmasked tokens (x-axis). The colored bands indicate the variance of the results as each configuration is repeated for 5 times.

Under-utilization of Unmasked Tokens. If the CMLM is weak at reasoning associations between images and *unmasked* tokens, replacing the unmasked ones with alternatives based on the language context should not prevent the model from recovering the masked tokens; otherwise, the performance of CMLM will drop significantly because the language context is inconsistent with the image. Therefore, we use a pre-trained VLM and LM to perform inference on CMLM and MLM, respectively, under a corrupted context in which we randomly replace unmasked tokens by sampling from the MLM head of a BERT. As illustrated in Fig. 2b, both the VLM and LM suffer performance deterioration under different corruption ratios. However, even though the text contexts are inconsistent with the image after corruption, the performance drop is less severe for the VLM as compared to the LM. In fact, the performance curves for CMLM and MLM overlap in terms of relative performance drop. Hence, we conclude that the CMLM is weak at learning associations between unmasked tokens and images.

4. Methodology

Sec. 4.1 first introduces preliminaries of the proposed method. We then introduce the three components of **EPIC**. We formulate the ITC task and introduce the concept of token consistency w.r.t. an image in Sec. 4.2. We design the inconsistent token generation procedure in Sec. 4.3, and the saliency-based masking strategy in Sec. 4.4. Fig. 3 gives an overview of the proposed method. The complete EPIC algorithm is shown in Appendix B.

4.1. Preliminaries

Pre-training Framework. In vision-language pre-training, we have access to parallel image-text data $\mathcal{D} = \{(\boldsymbol{w}_i, \boldsymbol{v}_i)\}_{i=1}^{N} \sim P_{\mathcal{W},\mathcal{V}}$. Specifically, for any sentence $\boldsymbol{w} = [w_1, \ldots, w_n]$, there is a corresponding image $\boldsymbol{v} =$

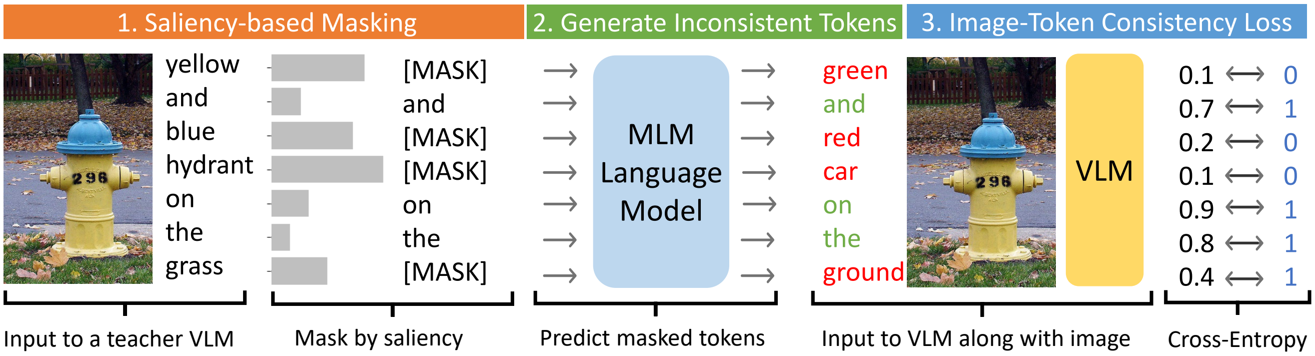


Figure 3. An overview of EPIC. The input image and sentence are input to a teacher VLM to obtain the saliency for each text token w.r.t the image. Then we mask accordingly and generate tokens inconsistent with the image though a language model (fine-tuning during pre-training). Finally we train the VLM to determine for each token whether it is inconsistent with the image (ITC).

$[v_1, \dots, v_m]$ in the form of grid-based or region-based features [8]. Without loss of generality, we assume a vision-language model $f_{\mathrm{VL}}(\cdot)$ in METER [8] of the following form: Given a pair of sentence $\boldsymbol{w}$ and image $\boldsymbol{v}$, it first extracts text features and visual features via a text encoder and vision encoder, respectively. The text and visual features are then fed into a multi-modal fusion module, consisting of several layers of self-attention followed by cross-attention, to produce cross-modal representations for text and image. Note that the proposed method is also applicable to other architectures such as ViLT [12], ALBEF [14], and X-VLM [39].

4.2. Image-Token Consistency

First, we introduce the concept of consistency of a text token with respect to an image.

Consistency of Text Tokens w.r.t Images. In vision-language pre-training, we assume that the dataset is clean (without noise) and that we have access to the marginal distribution of the text corpus $P_{\mathcal{W}}$. For a pair of text $\boldsymbol{w}$ and image $\boldsymbol{v}$, a text token w_i in $\boldsymbol{w}$ is "consistent" with the image given the remaining tokens, i.e., $P_{\mathcal{W},\mathcal{V}}(\boldsymbol{W} = \boldsymbol{w}, \boldsymbol{V} = \boldsymbol{v})$ is large. If we replace a token w_j with w_j' such that the new sentence $\bar{\boldsymbol{w}}$ (1) does not match the semantic content of the image, i.e., $P_{\mathcal{W},\mathcal{V}}(\boldsymbol{W} = \bar{\boldsymbol{w}}, \boldsymbol{V} = \boldsymbol{v})$ is small and (2) is linguistically proper, i.e., $P_{\mathcal{W}}(\boldsymbol{W} = \bar{\boldsymbol{w}})$ is large, then w_j' is inconsistent with the image given the remaining tokens.

An example is shown in Fig. 3. In the sentence "Yellow and blue hydrant on the grass", the word "hydrant" is consistent with the image given the remaining words. If we replace "yellow" with "green", it produces a sentence that does not match the image, and so "green“ is inconsistent.

Task Formulation. Given an image $\boldsymbol{v}$ and a sentence $\bar{\boldsymbol{w}}$ with inconsistent tokens at positions $\mathcal{T}$, the model has to determine whether each token is consistent with the image. We will introduce the generation process of $\bar{\boldsymbol{w}}$ and $\mathcal{T}$ in Sec. 4.3.

The $\bar{\boldsymbol{w}}$ and $\boldsymbol{v}$ are fed into VLM to obtain the last-layer hidden representations $\left\{\bar{\boldsymbol{h}}_i^{\mathrm{VL}}\right\}_{i=1}^{n}$ of $\bar{\boldsymbol{w}}$. Let $\boldsymbol{\beta}$ be the weight vector for decision. The probability that each token (from position $i = 1$ to n) is consistent with the image is:

$$D_{\mathrm{ITC}}^{i} = \operatorname{sigmoid}\left(\boldsymbol{\beta}^{\intercal}\bar{\boldsymbol{h}}_i^{\mathrm{VL}}\right). \tag{1}$$

Therefore, the EPIC model minimizes the following binary classification loss:

$$\mathcal{L}_{\mathrm{ITC}} = -\sum_{i \notin \mathcal{T}} \log D_{\mathrm{ITC}}^{i} - \sum_{i \in \mathcal{T}} \log\left(1 - D_{\mathrm{ITC}}^{i}\right). \tag{2}$$

Leveraging Unmasked Tokens. It is easy to see that in Eq. (2), the EPIC model is tasked to determine for all the tokens whether they are consistent with the image or not. Therefore, the ITC task leverages more tokens for learning vision-language associations.

Mitigating Modality Bias. In the second condition for inconsistent tokens, we require them to be linguistically proper for the sentence. We argue that this is essential for the ITC task to alleviate the modality bias problem because if the sentence after replacement $\bar{\boldsymbol{w}}$ is not linguistically proper, then the VLM can identify the replaced tokens simply by using the language context only. As a result, the VLM does not learn vision-language associations by decreasing $\mathcal{L}_{\mathrm{ITC}}$. Instead, it simply conducts language modeling.

4.3. Generating Inconsistent Tokens

In this section, we discuss how to generate sentences $\bar{\boldsymbol{w}}$ with tokens that fulfill the two conditions of inconsistent tokens in Sec. 4.2. Strictly speaking, we have to model the distributions $P_{\mathcal{W},\mathcal{V}}$ and $P_{\mathcal{W}}$ for generation, but this is hard. Instead, we propose to approximately generate inconsistent tokens with a BERT-like language model.

Assume that we have a set $\mathcal{M}$ of masking positions (we will discuss how to obtain $\mathcal{M}$ in Sec. 4.4). Given the original text sequence $\boldsymbol{w}$, we first mask the corresponding tokens by replacing them with the `[MASK]` symbol:

$$\boldsymbol{w}^{\text{mask}} = \text{MASK}\left(\boldsymbol{w}, \mathcal{M}\right). \tag{3}$$

The resulting sentence $\boldsymbol{w}^{\text{mask}}$ is fed to an auxiliary BERT-like language model $f_L(\cdot)$ to obtain a sequence of contextual representations $\boldsymbol{H}^{\text{L}} = \{\boldsymbol{h}_i^{\text{L}}\}_{i=1}^{n}$. We then obtain the probability of predicting a particular token w_k as:

$$p_{\text{MLM}}\left(w_k \mid \boldsymbol{h}_i^{\text{L}}\right) = \frac{\exp\left(\boldsymbol{e}\left(w_k\right)^\intercal \boldsymbol{h}_i^{\text{L}}\right)}{\sum_{w'} \exp\left(\boldsymbol{e}\left(w'\right)^\intercal \boldsymbol{h}_i^{\text{L}}\right)}, \tag{4}$$

where $\boldsymbol{e}(\cdot)$ denotes the token embedding. The new sentence, with the inconsistent tokens, is $\bar{\boldsymbol{w}} = [\bar{w}_1, \ldots, \bar{w}_n]$:

$$\bar{w}_i \sim p_{\text{MLM}}\left(w \mid \boldsymbol{h}_i^{\text{L}}\right). \tag{5}$$

The positions of the inconsistent tokens $\mathcal{T}$ are obtained by:

$$\mathcal{T} = \{t \mid \bar{w}_t \neq w_t, \forall t \in \mathcal{M}\}. \tag{6}$$

We also train the language model to fit to the sentence during pre-training by a MLM objective:

$$\mathcal{L}_{\text{GEN}} = -\sum_{i \in \mathcal{M}} \log p_{\text{MLM}}\left(w_i \mid \boldsymbol{h}_i^{\text{L}}\right). \tag{7}$$

We provide justification for using a language model to approximately fulfill the two conditions of inconsistent tokens. First, the language model has no access to the visual inputs. $P_{\mathcal{W},\mathcal{V}}(\boldsymbol{W} = \bar{\boldsymbol{w}}, \boldsymbol{V} = \boldsymbol{v})$ is likely to be small, and so it generates samples approximating the first condition. Second, since the language model is fine-tuned during pre-training, $P_{\mathcal{W}}(\boldsymbol{W} = \bar{\boldsymbol{w}})$ is likely to be large. Thus, it generates samples that can approximate the second condition.

4.4. Saliency-based Masking

In this section, we discuss how to obtain masking positions $\mathcal{M}$. First, we introduce the concept of saliency.

Saliency of Text Tokens w.r.t. Images. For a pair of sentence $\boldsymbol{w}$ and image $\boldsymbol{v}$, a text token w_i is salient w.r.t. to the image if the meaning of the token is strongly related to the content in the image, otherwise, it is not salient. In Fig. 3, words "yellow", "blue", "hydrant" and "grass" are salient w.r.t. the image, while "and", "the" are not.

Masking Salient Tokens. As discussed in Sec. 4.3, we first mask the original sentence based on $\mathcal{M}$, and obtain the positions $\mathcal{T}$ for inconsistent tokens by comparing the generated tokens with the original tokens (Eq. (6)). Since the language model is fine-tuned on the corpus, if we mask non-salient tokens w.r.t. an image, the language model is likely to recover the original tokens by attending to the remaining language context. In this case, no inconsistent tokens will be generated, and the model cannot learn vision-language associations. However, if we mask salient tokens w.r.t. an image, the language model is incapable of recovering such tokens because it has no access to the visual input. Therefore, it is expected to mask salient tokens for generating inconsistent tokens.

In practice, we find salient tokens by selecting the tokens/positions with higher attention scores to the image. The cross-modal representation of the `[CLS]` token of the image $\boldsymbol{v}$ is query $\boldsymbol{q}_k^V \in \mathbb{R}^{d_h}$, and that of all the elements in $\boldsymbol{w}$ are keys $\boldsymbol{K}_k^W \in \mathbb{R}^{n \times d_h}$ at head k, where n is the sequence length of the sentence, $d_h = d/h$ is the dimension of a single-head output. For simplicity, we only consider the representations from the penultimate layer of the fusion module. The image-token saliency $\boldsymbol{\alpha} \in \mathbb{R}^d$ can then be written as

$$\boldsymbol{\alpha} = \text{softmax}\left(\frac{1}{h}\sum_{k=1}^{h} \boldsymbol{q}_k^V (\boldsymbol{K}_k^W)^\intercal\right). \tag{8}$$

Finally, the masking positions $\mathcal{M}$ are sampled without replacement from the categorical distribution $p(t = i; \boldsymbol{\alpha}) = \alpha_i$ such that $|\mathcal{M}| = m$, where m is the expected number of inconsistent tokens. In this way, tokens with high saliencies are more likely to be masked. Notice that we use a pre-trained vision-language model to obtain $\boldsymbol{q}_k^V$ and $\boldsymbol{K}_k^W$. Details are in Appendix C.

5. Experiments

5.1. Pre-training Settings

Baselines. The proposed method has no constraints on the network architectures, training objectives, and visual representations. It supports architectures conducting either cross-attention or self-attention for multi-modal fusion. In this study, we plug-in our method into four recent approaches with diversified cross-modal learning techniques: (i) METER, [8] which adopts modality-specific encoders with cross-attention, (ii) ALBEF [14], which utilizes cross-modal contrastive learning with cross-attention, (iii) X-VLM [39], which performs cross-modal alignments from multiple granularities with cross-attention, and (iv) ViLT [12], which fuses images and text in a single encoder via self-attention. We reproduce the pre-training for all the baselines with the settings where they achieve their best results.

Datasets. There are four widely adopted datasets for vision-language pre-training: (i) COCO [18], (ii) Visual Genome (VG) [13], (iii) SBU Captions [24] and (iv) Conceptual Captions 3M (CC) [4]. We refer the combination of these datasets as 4M because there are 4 million unique images in them. We also experiment with a large-scale web dataset

Method	Data	EPIC	MSCOCO (5K test set)		Flickr30K (1K test set)		Flickr30K ZS(1K test set)	
			TR	IR	TR	IR	TR	IR
			R@1/R@5/R@10	R@1/R@5/R@10	R@1/R@5/R@10	R@1/R@5/R@10	R@1/R@5/R@10	R@1/R@5/R@10
METER	4M	✗	77.2 / 93.7 / 97.1	59.2 / 84.0 / 90.8	94.2 / **99.6 / 99.9**	83.8 / 97.3 / 98.6	92.2 / **99.3** / 99.7	78.8 / **94.5** / 96.9
		✓	**79.0 / 94.5 / 97.5**	**61.2 / 85.2 / 91.6**	**95.8** / 99.3 / 99.6	**85.1 / 97.4 / 98.7**	**93.1** / 99.1 / **99.8**	**79.0 / 94.5 / 97.1**
	16M	✗	78.2 / 93.8 / 96.9	59.8 / 84.2 / 90.8	94.6 / 99.7 / 99.8	85.2 / 97.4 / 98.8	93.1 / 99.4 / **99.8**	81.1 / **96.0** / 97.8
		✓	**79.7 / 94.8 / 97.5**	**62.5 / 85.4 / 91.9**	**96.5 / 99.9 / 99.9**	**86.7 / 97.8 / 99.0**	**94.7 / 99.5 / 99.8**	**82.6** / 95.9 / **99.8**
ALBEF	4M	✗	73.3 / 92.4 / **96.4**	56.4 / 81.8 / 88.9	94.4 / 99.4 / 99.8	82.1 / 95.7 / 97.9	91.1 / 98.7 / 99.4	76.1 / 93.0 / 95.9
		✓	**75.1 / 92.9 / 96.4**	**58.6 / 82.7 / 89.3**	**95.6 / 99.7 / 99.9**	**83.7 / 96.7 / 98.4**	**91.7 / 99.2 / 99.8**	**78.1 / 93.8 / 96.4**
	16M	✗	78.3 / 93.9 / 96.8	61.3 / 84.3 / 90.6	95.9 / 99.8 / **100**	85.7 / 97.2 / 98.8	93.6 / 99.5 / 99.8	83.2 / 95.9 / 97.7
		✓	**79.2 / 94.7 / 97.5**	**62.9 / 85.4 / 91.3**	**96.4 / 100 / 100**	**87.1 / 97.3 / 98.9**	**94.8 / 99.7 / 99.9**	**84.1 / 96.5 / 97.9**
X-VLM	$4M^+$	✗	79.8 / 95.1 / 97.7	62.7 / 85.6 / 91.4	96.7 / **99.9 / 100**	85.3 / 97.4 / **98.7**	83.1 / 97.8 / 99.4	70.5 / 92.8 / 96.4
		✓	**81.0 / 95.3 / 97.9**	**64.1 / 86.1 / 91.6**	**97.2 / 99.9 / 100**	**87.0 / 97.6 / 98.7**	**86.2 / 98.6 / 99.8**	**74.4 / 94.3 / 97.0**
	$16M^+$	✗	79.5 / 95.4 / 97.8	63.3 / 85.6 / 91.4	96.8 / **100 / 100**	86.7 / 97.5 / 98.7	86.4 / **99.2** / 99.6	**76.1** / 94.1 / 96.7
		✓	**80.7 / 95.6 / 98.0**	**64.1 / 85.9 / 91.8**	**97.4 / 100 / 100**	**87.3 / 97.6 / 98.8**	**89.0** / 99.0 / **99.7**	75.4 / **94.2 / 96.8**
ViLT	4M	✗	60.4 / 85.8 / 92.2	41.3 / 71.4 / 82.3	80.8 / 95.9 / 98.7	61.2 / 88.0 / 93.5	72.6 / 93.0 / 96.8	53.4 / 80.8 / 88.7
		✓	**65.0 / 87.5 / 93.7**	**46.0 / 74.8 / 84.6**	**85.2 / 97.3 / 99.3**	**66.9 / 90.0 / 94.4**	**79.8 / 95.8 / 97.9**	**63.3 / 86.3 / 92.0**

Table 1. Image-text retrieval results on the MSCOCO (fine-tuned) and Flickr30K (fine-tuned and zero-shot) datasets. IR: Image Retrieval and TR: Text Retrieval. Recall@K with K = 1, 5, and 10 is used as the evaluation metric. Better results under the same baseline are marked in **bold**.

Method	Data	EPIC	MSCOCO (5K test set)					
			TR			IR		
			R@1	R@5	R@10	R@1	R@5	R@10
X-VLM	$4M^+$	✗	69.2	91.9	96.5	55.3	82.5	89.7
		✓	**72.0**	**93.4**	**97.3**	**57.3**	**83.4**	**90.2**
	$16M^+$	✗	73.1	92.8	**97.0**	56.9	83.0	89.9
		✓	**73.2**	**93.6**	96.9	**57.5**	**83.3**	**90.0**
ViLT	4M	✗	54.6	81.4	89.1	38.6	69.0	80.2
		✓	**64.1**	**86.5**	**92.5**	**47.1**	**74.8**	**84.6**

Table 2. Zero-shot image-text retrieval results on MSCOCO.

CC12M [4]. We consider the combination of 4M and 12M datasets as 16M. Additionally, the X-VLM requires fine-grained annotations (e.g., object and region descriptions) for both the 4M and 16M settings. Therefore, the resulting dataset for X-VLM is called $4M^+$ and $16M^+$, respectively. A more detailed description of the data statistics is in Appendix E.

Implementation Details. For the pre-training baselines, we follow the official implementations provided by the authors of METER, X-VLM, ALBEF and ViLT. More details on their implementations are in Appendix F. For the proposed method, the auxiliary language model is chosen to be identical to the text encoder of the vision-language model. For ViLT without modality-specific encoders, we use BERT-base [6]. All the auxiliary language models are loaded directly from HuggingFace repository [34] with pre-trained checkpoints. For every sentence, the number of masked tokens m is calculated as m = ceil(mask_ratio × len(sentence)). The mask ratio is set to 0.35 for all baselines (based on the hyper-parameter search results in Sec. 5.3). The loss weight λ of the Image-Token Consistency task is set to 8.

5.2. Results on Downstream Tasks

Evaluation is performed on the following downstream tasks: (i) Image-Text Retrieval, (ii) Visual Question Answering (VQA), (iii) Natural Language for Visual Reasoning (NLVR2), and (iv) Visual Entailment (VE). Details on these tasks are in Appendix D.

5.2.1 Results for Image-Text Retrieval

Tables 1 and 2 show that EPIC is universally effective over the different baselines/datasets, since we achieve non-trivial improvement nearly for all the settings. In addition, compared with all the baselines, EPIC brings more significant improvement to ViLT on different datasets. For example, on the fine-tuned retrieval tasks, we achieve an absolute improvement of around 5% in terms of the TR/IR @1 metric on MSCOCO and Flickr30K. Furthermore, for zero-shot tasks, the improvement on the Flickr30K dataset is 7.2% in terms of TR@1, and 9.9% in terms of IR@1. The gap between ViLT and the other baselines can be explained by the fact that ViLT adopts a simplified architecture for vision-language pre-training and it does not incorporate more sophisticated tasks to learn vision-language associations. Nevertheless, the results on ViLT can be seen as an indicator for the effectiveness of the proposed method on a "clean" baseline.

For the other baselines with advanced strategies to learn vision-language associations (such as cross-modal contrasting in ALBEF, fine-grained reasoning in X-VLM, and powerful pre-trained encoders in METER), EPIC still demonstrates significant improvements. For example, compared with METER on the MSCOCO dataset (fine-tuned), EPIC

Method	Data	EPIC	VQA		NLVR2		SNLI-VE	
			dev	std	dev	std	dev	std
METER	4M$^+$	✗	77.7	77.9	81.8	82.5	81.4	81.0
		✓	**77.9**	**78.0**	**83.5**	**83.5**	**81.6**	**81.8**
	16M$^+$	✗	78.3	78.4	82.7	84.3	81.7	81.8
		✓	**78.6**	**78.7**	**85.0**	**85.2**	**82.1**	**82.3**
ALBEF	4M$^+$	✗	74.6	74.6	79.5	80.0	80.1	80.1
		✓	**75.1**	**75.2**	**81.3**	**82.2**	**80.6**	**80.7**
	16M$^+$	✗	75.8	76.0	82.6	82.5	80.8	80.9
		✓	**76.7**	**76.7**	**84.1**	**84.0**	**81.3**	**81.7**
X-VLM	4M$^+$	✗	78.1	78.2	83.3	84.1	/	/
		✓	**78.5**	**78.5**	**84.6**	**84.5**	/	/
	16M$^+$	✗	78.0	78.2	84.3	84.5	/	/
		✓	**78.3**	**78.3**	**85.2**	**85.5**	/	/
ViLT	4M	✗	71.3	71.4	75.0	75.2	/	/
		✓	**71.8**	**71.8**	**77.2**	**77.1**	/	/

Table 3. Evaluation results on downstream vision-language tasks: VQA, NLVR2, and SNLI-VE. "/" indicates that the original baseline does not conduct experiment on this task.

achieves an absolute IR@1 improvement of 2% and 2.7% with 4M and 16M data, respectively. When using the X-VLM as baseline, we observe an absolute improvement of 3.9% and 3.1% in terms of IR@1 and TR@1, respectively, on Flickr30K (zero-shot).

5.2.2 Results for VQA, NLVR2, and VE

Table 3 shows that EPIC is effective among all the vision-language tasks. We improve over METER in terms of the VQA dev accuracy by 0.3% under the 16M setting and X-VLM by 0.4% under the 4M$^+$ setting. Such improvement is non-trivial given the fact that these two baselines are quite competitive on this task. We also notice that EPIC improves ALBEF on VQA by 0.9% in dev accuracy. Further, we observe significant improvements over all baselines on NLVR2 (e.g., +2.3% dev accuracy on METER 16M; +2.2% std accuracy on ALBEF 4M). The proposed method is also effective on the SNLI-VE dataset. It brings an absolute improvement of 0.8% over ALBEF under the 16M setting.

Method	NLVR2	Flickr30K-ft		Flickr30K-zs		MSCOCO-ft	
	dev	TR1	IR1	TR1	IR1	TR1	IR1
vanilla METER	79.6	89.2	76.6	83.2	67.7	71.0	52.5
ITC (rand.)	79.9	91.5	77.8	83.2	69.0	70.8	53.5
ITC+LM	80.9	92.1	78.9	83.8	71.5	73.4	**55.6**
EPIC	**81.0**	**92.9**	**79.0**	**84.3**	**72.5**	**74.1**	**55.6**

Table 4. Ablation studies on the effect of the ITC task, negative samples generation and the saliency-based masking strategy.

Generator	NLVR2	Flickr30K-ft		Flickr30K-zs		MSCOCO-ft	
	dev	TR1	IR1	TR1	IR1	TR1	IR1
LM (cond.)	79.9	92.4	78.7	84.0	71.8	72.5	54.5
VLM (SAS)	80.5	91.0	79.2	83.2	70.4	72.4	54.9
LM (trained)	80.7	91.3	**79.5**	**84.8**	72.1	73.6	55.1
LM (fixed)	80.7	91.9	79.0	83.8	71.1	72.4	54.8
LM (fine-tune)	**81.0**	**92.9**	79.0	84.3	**72.5**	**74.1**	**55.6**

Table 5. Ablation study on different negative sample generators.

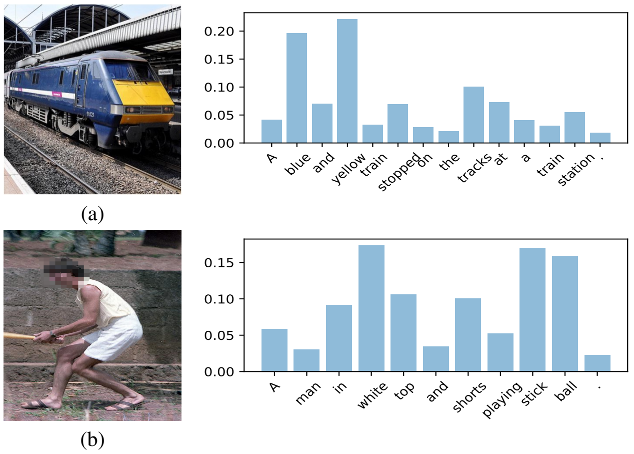
5.3. Ablation Study

We conduct ablation studies based on the METER model due to its superior performance with basic training objectives (i.e., image-text matching and cross-modal masked language modeling). However, training a full-version of METER is expensive (roughly 4 days with 32 A100 80G). Hence, we conduct ablations on a smaller scale that is still representative. Specifically, the model is still pre-trained on the 4M dataset, but the input image is resized to 224 × 224 (instead of 288 × 288 in the normal METER setting). To speed up the training process, we replace the image encoder CLIP-16 [26] with CLIP-32 for less memory usage, and also shorten the training schedule to 50k steps.

We evaluate the pre-trained model on the NLVR2 task and retrieval task (Flickr30K fine-tuned/zero-shot, MSCOCO fine-tuned). For the NLVR2 task, we report the development accuracy. For the retrieval tasks, we report the IR1/TR1 accuracies on the corresponding validation sets.

Component Analysis. As shown in Table 4, on top of vanilla METER, we first add an ITC task (*ITC (rand.)*) with inconsistent tokens sampled uniformly from the vocabulary (selecting 35% of positions for inconsistent tokens). In this case, though some of the sampled tokens do not satisfy the conditions for inconsistent tokens in Sec. 4.2, the pre-trained model can still benefit from the ITC task. When we replace the random strategy with a language model, *ITC+LM*, we generate inconsistent tokens approximating the conditions in Sec. 4.2. Finally, we choose to mask the tokens that are salient w.r.t. the image to generate inconsistent tokens, *ITC+saliency+LM*, and this further improves the performance on downstream tasks. Therefore, each component of EPIC is effective to improve the performance of the pre-trained model.

Conditions of Inconsistent Tokens. In this ablation experiment, we study the importance of the two conditions of inconsistent tokens in Sec. 4.2. *LM (fine-tune)* is proposed in EPIC and it produces inconsistent tokens approximating the two conditions. First, we drop the first condition, that is, $P_{\mathcal{W},\mathcal{V}}(\boldsymbol{W} = \bar{\boldsymbol{w}}, \boldsymbol{V} = \boldsymbol{v})$ is no longer small. We achieve this in two ways. (i) *LM (cond.)*: When generating inconsistent samples and fine-tuning the LM, we replace the class token (first token) in the text with the one from the image encoder of the VLM (with gradient propagation canceled).



(a)

(b)

Figure 4. Visualization on the token saliency distribution.

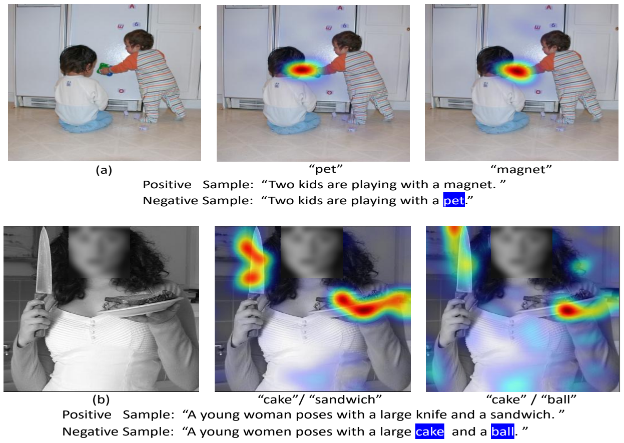

(a)
“pet”
“magnet”
Positive Sample: “Two kids are playing with a magnet. ”
Negative Sample: “Two kids are playing with a pet.”
(b)
“cake”/ “sandwich”
“cake” / “ball”
Positive Sample: “A young woman poses with a large knife and a sandwich. ”
Negative Sample: “A young women poses with a large cake and a ball. ”

Figure 5. Grad-CAM visualization on the cross-attention maps of the image with respect to the ITC task.

(ii) *VLM (SAS)*: We adapt SAS [36] from the text modality to the multi-modality setting. An auxiliary VLM (previous checkpoints of the VLM, detailed in Appendix G) is used to produce inconsistent tokens. As shown in Table 5, we observe that both *LM (cond.)* and *VLM (SAS)* suffer performance deterioration compared to *LM (fine-tune)*. This indicates the importance of the first condition.

We then demonstrate the importance of the second condition, that is $P_{\mathcal{W}}(\boldsymbol{W} = \bar{\boldsymbol{w}})$ is large. Intuitively, we fail to satisfy this condition when we stop fine-tuning the LM on the text corpus, i.e., *LM (fixed)*. Further, for *LM (fine-tune)*, we experiment its possible alternatives, *LM (trained)*, where we fit the LM on the text corpus before instead of during pre-training. We observe that *LM (fixed)* clearly decreases the performance (especially in MSCOCO-ft). This validates the effectiveness of the second condition. Note that *LM (fine-tune)* outperforms *LM (trained)* by a small margin. This gap can be attributed to the training dynamics brought by fine-tuning [22]. As the LM fits the text corpus gradually better during finetuning, it becomes increasingly hard for a model to identify whether a token is replaced. This dynamics encourages the model to perform curriculum learning, which improves performance.

5.4. Visualization

Token Saliency. Fig. 4 shows the saliency distribution of the tokens w.r.t. the image for an image-sentence pair. We can see that salient tokens have higher densities in the distribution. For example, in Fig. 4a, “blue” and “yellow” are highlighted. In Fig. 4b, the teacher model gives more mass on “white”, “stick” and “ball”. These confirm that salient tokens can be detected by EPIC.

Grad-CAM of ITC. Fig. 5 visualizes the cross-attention maps of the image w.r.t. the inconsistent/consistent tokens using Grad-CAM [28]. In each row, we show the original image (left), the attention map w.r.t. the inconsistent tokens (middle), and that w.r.t. the consistent ones (right). For example, in the first row, when we input the model with the original sentence, the model predicts that the token “magnet” is consistent and attends to the actual magnet in the image. However, when we replace “magnet” with “pet”, the model predicts the token “pet” as “inconsistent” and still attends to the magnet in the image. This means the model is aware that the object in the image is actually a magnet instead of a pet. Similar observation exists for the second row. These demonstrate the model’s ability to reason on cross-modal relationship.

6. Conclusion

In this paper we propose **EPIC**, a pre-training approach that leverage more text tokens for learning vision-language associations. It is less affected by the modality bias problem compared with CMLM. Specifically, we propose an ITC task to identify inconsistent tokens generated by a language model coupled with a saliency-based masking strategy. The task formulation of the ITC task and the design of inconsistent samples address the problems of under-utilization of unmasked tokens and modality bias. We perform extensive experiments and show that EPIC brings consistent performance gains over several baselines on a wide range of downstream tasks. Possible directions for future research can be: (i) approaching the conditions of inconsistent tokens in Sec. 4.2 more precisely; (ii) finding salient tokens without using a pre-trained teacher VLM.

Acknowledgements

This work is supported by NSFC key grant 62136005, NSFC general grant 62076118, and Shenzhen fundamental research program JCYJ20210324105000003.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 2, 3, 11

[2] Yonatan Bitton, Michael Elhadad, Gabriel Stanovsky, and Roy Schwartz. Data efficient masked language modeling for vision and language. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3013–3028, Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. 2, 3

[3] Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior Research Methods*, 46:904–911, 2014. 3

[4] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021. 5, 6

[5] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020. 1

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2, 6

[7] Zi-Yi Dou, Aishwarya Kamath, Zhe Gan, Pengchuan Zhang, Jianfeng Wang, Linjie Li, Zicheng Liu, Ce Liu, Yann LeCun, Nanyun Peng, et al. Coarse-to-fine vision-language pre-training with fusion in the backbone. *arXiv preprint arXiv:2206.07643*, 2022. 3

[8] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, Zicheng Liu, and Michael Zeng. An empirical study of training end-to-end vision-and-language transformers. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 4, 5, 11, 12

[9] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6325–6334, 2017. 11

[10] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 1, 2

[11] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790, 2021. 3

[12] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021. 1, 4, 5, 12

[13] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017. 5

[14] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq R. Joty, Caiming Xiong, and Steven C. H. Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*, 2021. 1, 2, 4, 5, 12

[15] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022. 3

[16] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020. 1, 2

[17] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*, 2021. 2

[18] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 5, 11

[19] Yongfei Liu, Chenfei Wu, Shao-yen Tseng, Vasudev Lal, Xuming He, and Nan Duan. Kd-vlp: Improving end-to-end vision-and-language pretraining with object knowledge distillation. *arXiv preprint arXiv:2109.10504*, 2021. 1, 2

[20] Haoyu Lu, Nanyi Fei, Yuqi Huo, Yizhao Gao, Zhiwu Lu, and Ji-Rong Wen. Cots: Collaborative two-stream vision-language pre-training model for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15692–15701, 2022. 2

[21] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 1, 2

[22] Yu Meng, Chenyan Xiong, Payal Bajaj, Paul Bennett, Jiawei Han, Xia Song, et al. Coco-lm: Correcting and contrasting text sequences for language model pretraining. *Advances in Neural Information Processing Systems*, 34:23102–23114, 2021. 8

[23] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition, pages 12700–12710, 2021. 1, 3

[24] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24, 2011. 5

[25] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 11

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 7, 11

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021. 1, 2

[28] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 8

[29] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650, 2022. 1

[30] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019. 1

[31] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491*, 2018. 2, 11

[32] Hao Tan and Mohit Bansal. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. 1, 3

[33] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021. 1

[34] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, Oct. 2020. Association for Computational Linguistics. 6, 11

[35] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*, 2019. 12

[36] Yifei Xu, Jingqiao Zhang, Ru He, Liangzhu Ge, Chao Yang, Cheng Yang, and Ying Nian Wu. Sas: Self-augmented strategy for language model pre-training. *arXiv preprint arXiv:2106.07176*, 2021. 8

[37] Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15671–15680, 2022. 1

[38] Haoxuan You, Luowei Zhou, Bin Xiao, Noel Codella, Yu Cheng, Ruochen Xu, Shih-Fu Chang, and Lu Yuan. Learning visual representation from modality-shared contrastive language-image pre-training. In *European Conference on Computer Vision*, pages 69–87. Springer, 2022. 2

[39] Yan Zeng, Xinsong Zhang, and Hang Li. Multi-grained vision language pre-training: Aligning texts with visual concepts. *arXiv preprint arXiv:2111.08276*, 2021. 2, 3, 4, 5, 12