2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) | 979-8-3503-0129-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/CVPR52729.2023.01060

ViLEM: Visual-Language Error Modeling for Image-Text Retrieval

Yuxin Chen[1,2,4*], Zongyang Ma[1,2,4*], Ziqi Zhang[1,4*], Zhongang Qi[2], Chunfeng Yuan[1†], Ying Shan[2], Bing Li[1], Weiming Hu[1,4,5], Xiaohu Qie[3], JianPing Wu[6]

[1]State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences; [2]ARC Lab, [3]Tencent PCG; [4]School of Artificial Intelligence, University of Chinese Academy of Sciences; [5]CAS Center for Excellence in Brain Science and Intelligence Technology; [6]Tsinghua University

{chenyuxin2019,mazongyang2020}@ia.ac.cn, {ziqi.zhang,cfyuan,bli,wmhu}@nlpr.ia.ac.cn
{zhongangqi,yingsshan,tigerqie}@tencent.com, jianping@cernet.edu.cn

Abstract

*Dominant pre-training works for image-text retrieval adopt "dual-encoder" architecture to enable high efficiency, where two encoders are used to extract image and text representations and contrastive learning is employed for global alignment. However, coarse-grained global alignment ignores detailed semantic associations between image and text. In this work, we propose a novel proxy task, named **V**isual-**L**anguage **E**rror **M**odeling (**ViLEM**), to inject detailed image-text association into "dual-encoder" model by "proofreading" each word in the text against the corresponding image. Specifically, we first edit the image-paired text to automatically generate diverse plausible negative texts with pre-trained language models. ViLEM then enforces the model to discriminate the correctness of each word in the plausible negative texts and further correct the wrong words via resorting to image information. Furthermore, we propose a multi-granularity interaction framework to perform ViLEM via interacting text features with both global and local image features, which associates local text semantics with both high-level visual context and multi-level local visual information. Our method surpasses state-of-the-art "dual-encoder" methods by a large margin on the image-text retrieval task and significantly improves discriminativeness to local textual semantics. Our model can also generalize well to video-text retrieval.*

1. Introduction

Pre-training vision-language models on massive image-text pairs to learn transferable representations for image-text retrieval has attracted a lot of attention in recent years. Previous dominant methods [11, 29, 38] adopt "dual-encoder" architecture to enable efficient retrieval, where two separate encoders are used to extract image and text representations. They learn a joint image-text embedding space via constraining the coarse-grained alignment between global image and text features. However, the coarse-grained alignment constraint ignores the capture of detailed image and text semantics, and associations between them, impeding the performance improvement of image-text retrieval.

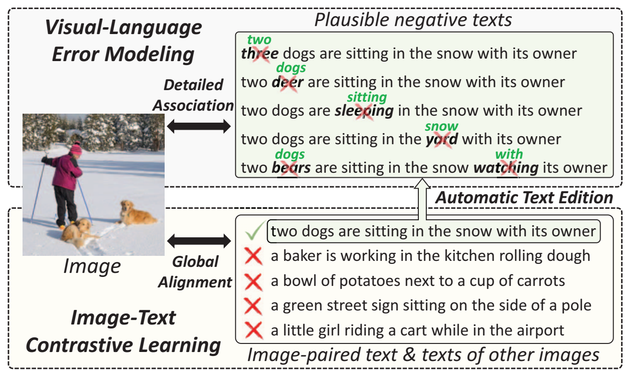


Figure 1. Illustration of image-text contrastive learning (ITC) and visual-language error modeling (ViLEM). ITC learns image-text global alignment by distinguishing paired data from unpaired data. ViLEM establishes detailed image-text association via discriminating and correcting wrong words in plausible negative texts.

Humans achieve accurate image-text matching by carefully discriminating whether there exists semantic divergence between image and text, *i.e*., determining whether each word can be precisely grounded to the image, which requires a comprehensive perception of each modality and well association between them. Humans can also eliminate semantic divergence effortlessly by correcting text errors through their powerful semantic association capability. Inspired by these, we propose a novel proxy task, named

* Equal contribution. † Corresponding author.

979-8-3503-0129-8/23/$31.00 ©2023 IEEE
DOI 10.1109/CVPR52729.2023.01060

Visual-**L**anguage **E**rror **M**odeling (**ViLEM**), for image-text retrieval. As shown in Figure 1, compared with image-text contrastive learning for global alignment, ViLEM enforces the model to discriminate and eliminate the local semantic divergence by "proofreading" plausible negative texts against image information, which enhances fine-grained semantic perception and establishes detailed image-text association. Collaborating with image-text contrastive learning, ViLEM significantly improves the retrieval performance of "dual-encoder" architecture.

ViLEM is divided into two sub-tasks: text error detection and text error correction. Given an image and a plausible negative text, the goal of error detection is training the model to exhaustively discriminate the correctness of each word in the form of binary classification. Meanwhile, error correction enforces the model to predict the correct words for the wrong ones from a fixed vocabulary under the condition of image information. However, finding plausible negative text for images and obtaining corresponding labels of error detection and correction requires high human annotation costs. Thus, we propose to automatically construct plausible negative texts and corresponding labels with a pre-trained language model BERT [12], where we exploit its rich linguistic knowledge to edit the image-paired texts and generate local text errors. The generated errors can be related to objects, actions, scenes, relationships, *etc.* (as shown in Figure 1), with which the model can learn various fine-grained semantics. The detection and correction labels can also be obtained by comparing generated negative texts with image-paired texts.

To further leverage ViLEM's ability to establish semantics associations, we propose a multi-granularity interaction framework to enable effective interaction between visual and textual encoders while maintaining high retrieval efficiency. Specifically, global visual features and local visual features are both fully exploited for text error detection and correction. For global visual features, we inject them into the local text representations to provide visual conditions for discriminating and correcting text errors, which associates local text information with high-level visual context and enhances the discriminativeness to fine-grained text semantics. For local visual features, we employ additional cross-attention modules to adaptively aggregate them into word-related visual concepts for error detection and correction, which establishes the association between detailed text semantics with multi-level local visual information and facilitates fine-grained image-text alignment. The cross-attention modules will be removed in the inference, introducing no additional computation cost and parameters compared with vanilla "dual-encoder".

The contributions of this work are listed as follows:

(1) We introduce a novel proxy task, Visual-Language Error Modeling (ViLEM), to inject detailed semantic association between images and texts into "dual-encoder" architecture.

(2) We propose a multi-granularity interaction framework to further leverage the ability of ViLEM while maintaining the high retrieval efficiency, which enhances the capture of fine-grained semantics and associates local text semantics with both high-level visual context and multi-level local visual information.

(3) The extensive experimental results show that our method surpasses previous state-of-the-art "dual-encoder" methods by a large margin on the image-text retrieval task and significantly improves the discriminativeness to local text semantics. Moreover, our model can also generalize well to video-text retrieval.

2. Related Work

Pre-training for Image-text Retrieval. Previous pre-training works for image-text retrieval can be divided into two categories, *i.e.*, "joint-encoder" methods and "dual-encoder" methods. "Joint-encoder" methods [4, 16–18, 22, 42] contain a multi-modal encoder to enable fine-grained feature interaction between image and text. The binary classification objective is utilized to predict whether the input image and text are matched. Despite their promising performance, every image-text pair needs to be fed into the joint encoder, leading to extreme inefficiency. "Dual-encoder" methods [11, 21, 29, 34, 38] adopt two individual encoders to extract the image and text features separately, and project global representations into a shared embedding space. These methods allow the pre-computing of global image and text features and achieve efficient retrieval by calculating dot product between features. The contrastive learning [25] is leveraged to distinguish paired image-text data from unpaired data. However, imposing contrastive objectives only on the global features leads to the under-exploitation of local semantics of images and texts.

Association Enhancement for Dual-encoder. Recent works [21, 34] introduce Masked Language Modeling (MLM) [12] to facilitate image-text association of dual-encoder, where a proportion of words are randomly masked and the model is trained to recover the masked words with global visual features. These works ignore the association between local text semantics and local visual information, hindering the learning of fine-grained image-text alignment. Moreover, the MLM task only considers a proportion of words (*e.g.*, 15%) and may ignore visual features to predict the masked words with only textual context, affecting the efficiency and effectiveness for the learning of image-text association. On the contrary, our ViLEM task enforces the model to fully exploit detailed image and text semantics to determine the correctness of each word in the text and correct the wrong words. Furthermore, we perform ViLEM by

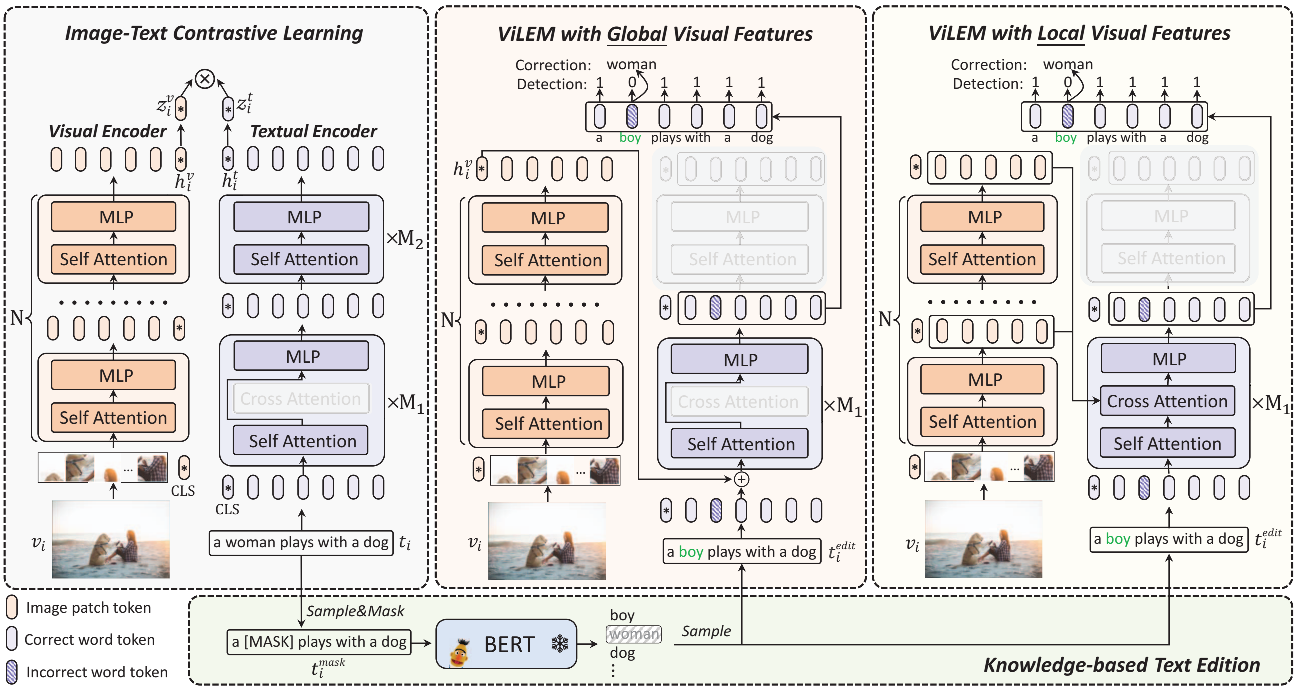


Figure 2. The illustration of ViLEM and multi-granularity interaction framework. ViLEM is performed via interacting text features with global and local visual features respectively. The model is trained with three objectives: image-text contrastive learning, ViLEM with high-level global visual features, and ViLEM with multi-level local visual features. We adopt a pre-trained language model (BERT) to generate plausible negative texts with local errors. We only show the edition process of one word and omit momentum encoders for simplicity.

interacting text features with both global and local visual features, associating local text semantics with both high-level visual context and multi-level local visual information.

Text Error Correction. Text error correction has an important application area named Grammatical Error Correction (GEC) [24, 30, 35]. GEC task takes a potentially erroneous sentence as input and is expected to correct different kinds of linguistic errors in text such as spelling, punctuation, grammatical, *etc.* Another work [5] pre-trains language model with the task of word detection. They train a generator with MLM task to corrupt natural sentences and a discriminator to detect whether the words are corrupted. The generator and discriminator both learn rich linguistic knowledge through the adversarial training procedure. On the contrary, we adopt a pre-trained language model to generate plausible but visual-incorrect texts, which serve as training samples for ViLEM. Moreover, we detect and correct text errors via resorting to visual features, aiming at facilitating image-text association and further improving retrieval performance.

3. Method

In this work, we propose a novel proxy task ViLEM and a multi-granularity interaction framework to effectively inject detailed image-text association into the "dual-encoder" architecture. We first revisit the image-text pre-training for dual-encoder in Sec. 3.1, then introduce the proposed proxy task ViLEM with multi-granularity interaction in Sec. 3.2 and the learning objectives in Sec. 3.3.

3.1. Revisiting Pre-training for Dual-encoder

As shown in Figure 2, the dual-encoder contains a visual encoder $f^v(\cdot)$ and a textual encoder $f^t(\cdot)$. Both encoders consist of multiple transformer blocks [37] and each block mainly contains a multi-head self-attention and a feed-forward network. We additionally employ cross attention modules in the first M_1 layers of the textual encoder to enable local image-text interaction for ViLEM. But the cross attention modules are deactivated during the image-text contrastive learning for maintaining high retrieval efficiency. Given an input image $\boldsymbol{v}_i$ and its paired text $\boldsymbol{t}_i$, the [CLS] token is concatenated with inputs for feature aggregating, and the global representations $\boldsymbol{h}_i^v$ and $\boldsymbol{h}_i^t$ are encoded by the visual encoder and textual encoder respectively. Then the global representations are projected into a shared semantic embedding space as $\boldsymbol{z}_i^v$ and $\boldsymbol{z}_i^t$ with two linear transformations. The similarity between image and text is measured with dot product between $\boldsymbol{z}_i^v$ and $\boldsymbol{z}_i^t$.

The momentum contrastive learning [9, 21] is adapted for global feature alignment between images and texts. Two momentum updated encoders $\hat{f}^v(\cdot)$ and $\hat{f}^t(\cdot)$ are maintained to produce consistent momentum features $\hat{z}^v$, $\hat{z}^t$, which serve as negative samples for current input images and texts. The parameters of momentum encoders are updated as:

$$\hat{\boldsymbol{\theta}}^v = m \cdot \hat{\boldsymbol{\theta}}^v + (1-m) \cdot \boldsymbol{\theta}^v, \tag{1}$$

$$\hat{\boldsymbol{\theta}}^t = m \cdot \hat{\boldsymbol{\theta}}^t + (1-m) \cdot \boldsymbol{\theta}^t, \tag{2}$$

where m is momentum coefficient. $\boldsymbol{\theta}^v$, $\boldsymbol{\theta}^t$, $\hat{\boldsymbol{\theta}}^v$ and $\hat{\boldsymbol{\theta}}^t$ denote the parameters of $f^v(\cdot)$, $f^t(\cdot)$, $\hat{f}^v(\cdot)$, $\hat{f}^t(\cdot)$ respectively.

Moreover, we maintain two queues $\mathcal{Q}^v = \{\hat{\boldsymbol{z}}_j^v\}_{j=1}^{Nq}$ and $\mathcal{Q}^t = \{\hat{\boldsymbol{z}}_j^t\}_{j=1}^{Nq}$ to keep the momentum features $\hat{\boldsymbol{z}}^v$ and $\hat{\boldsymbol{z}}^t$ from previous iterations. The introduction of queues dramatically increases the number of negative samples, which is vital for contrastive learning. Given each image in the current mini-batch, its paired text is regarded as a positive sample. Its unpaired texts in the mini-batch and all samples in the $\mathcal{Q}^t$ are regarded as negative samples. The InfoNCE loss [25] is utilized to maximize the similarity between positive image-text pairs and minimize the similarity between negative pairs, which is defined as follows:

$$\mathcal{L}_{\mathrm{I2T}} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(\boldsymbol{z}_i^v, \hat{\boldsymbol{z}}_i^t, \tau)}{\sum_{j=1}^{B+N_q} \exp(\boldsymbol{z}_i^v, \hat{\boldsymbol{z}}_j^t, \tau)}, \tag{3}$$

where $\exp(\boldsymbol{x}, \boldsymbol{y}, \tau) = e^{\boldsymbol{x}^T \boldsymbol{y}/\tau}$, τ is the temperature hyperparameter, and B is the batch size.

Similarly, given each text in the current mini-batch, the contrastive loss is defined as:

$$\mathcal{L}_{\mathrm{T2I}=} - \frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(\boldsymbol{z}_i^t, \hat{\boldsymbol{z}}_i^v, \tau)}{\sum_{j=1}^{B+N_q} \exp(\boldsymbol{z}_i^t, \hat{\boldsymbol{z}}_j^v, \tau)}. \tag{4}$$

The total loss for image-text contrastive learning is defined as:

$$\mathcal{L}_{\mathrm{align}} = (\mathcal{L}_{\mathrm{I2T}} + \mathcal{L}_{\mathrm{T2I}})/2. \tag{5}$$

3.2. Visual-Language Error Modeling

3.2.1 Knowledge-based Text Edition

ViLEM facilitates the learning of local semantic association by detecting and correcting local text errors from plausible negative texts. However, it is difficult to find corresponding plausible negative texts for a given image, and obtaining training labels requires expensive human annotation to locate and correct the wrong words that do not match with image content. To automatically construct training samples for ViLEM, we propose to leverage the rich linguistic knowledge of the pre-trained language model BERT [12] to edit the image-paired text and generate local text errors.

Given a text composed of n tokens $\boldsymbol{t}_i = [t_{i1}, t_{i2}, ..., t_{in}]$, we first randomly select a set of positions to edit $\boldsymbol{e}_i = [e_{i1}, ...e_{ik}]$ and replace the tokens in the selected position with [MASK] to obtain the masked text $\boldsymbol{t}_i^{mask}$. BERT then takes the $\boldsymbol{t}_i^{mask}$ as input and reasons with textual context to predict possible candidate words for each masked position. To ensure the reasonableness and semantic richness of the predicted words, we randomly sample words from the top-k candidates to generate the final edited text $\boldsymbol{t}_i^{edit}$. We also avoid sampling the original words to ensure that text errors are generated at each mask position. Through the text editing process, the ground-truth error detection label $\boldsymbol{y}_i^{det}$ and error correction label $\boldsymbol{y}_i^{cor}$ can be inherently obtained as follows:

$$y_{ij}^{det} = \begin{cases} 0, & \text{if } j \in \boldsymbol{e}_i, \\ 1, & \text{if } j \notin \boldsymbol{e}_i. \end{cases} \tag{6}$$

$$y_{ij}^{cor} = \begin{cases} t_{ij}, & \text{if } j \in \boldsymbol{e}_i, \\ \text{none}, & \text{if } j \notin \boldsymbol{e}_i, \end{cases} \tag{7}$$

where $y_{ij}^{cor} = \text{none}$ indicates we don't calculate loss on the j-th token of $\boldsymbol{t}_i^{edit}$. It is worth noting that the synonyms may be sampled to replace the original word, which introduces noise in the training process. Fortunately, the number of synonyms per word is relatively small, and most of the sampled words have different semantics from the original word, ensuring the effectiveness of our method.

3.2.2 ViLEM with Global Visual Feature

We first perform ViLEM with global visual feature to associate local text information with high-level visual context and enhance the discriminativeness to fine-grained text semantics. It is worth noting that the textual encoder also serves as a textual decoder to predict correct words, which may interfere with the encoding of text features. Thus, we perform ViLEM with only the first M_1 layers of the textual encoder, which essentially divides the textual encoder into a sub-decoder and a sub-encoder, and decouples the encoding and decoding functions of the textual encoder to mitigate interference.

As shown in Figure 2, given an image $\boldsymbol{v}_i$ and its corresponding edited text $\boldsymbol{t}_i^{edit}$, we extract the global image feature $\boldsymbol{h}_i^v$ and add it to the word embeddings of $\boldsymbol{t}_i^{edit}$, providing visual condition for text error detection and correction. Then we feed word embeddings into the first M_1 layers of the textual encoder to discriminate the correctness of each word and predict the corresponding correct words. We take the output features from the textual encoder's M_1 layer to compute error detection loss $\mathcal{L}_{\mathrm{det}}$ and correction loss $\mathcal{L}_{\mathrm{cor}}$, which are formulated as:

$$\mathcal{L}_{\mathrm{det}}(\boldsymbol{h}^v) = \mathbb{E}\Big(\sum_{j=1}^{n} -\log P_{det}^j(y_{ij}^{det}|\boldsymbol{t}_i^{edit}, \boldsymbol{h}_i^v)\Big), \tag{8}$$

$$\mathcal{L}_{\mathrm{cor}}(\boldsymbol{h}^v) = \mathbb{E}\Big(-\log P_{cor}^j(y_{ij}^{cor}|\boldsymbol{t}_i^{edit}, \boldsymbol{h}_i^v)\Big), \tag{9}$$

where $\mathcal{L}_{\mathrm{det}}(\boldsymbol{h}^v)$ and $\mathcal{L}_{\mathrm{cor}}(\boldsymbol{h}^v)$ indicates the text error detection and correction are performed under the condition of global visual feature $\boldsymbol{h}^v$. P_{det}^j and P_{cor}^j are predicted probability distributions of error detection and error correction for j-th token. The final loss for ViLEM with global visual features is formulated as:

$$\mathcal{L}_{\mathrm{EMG}} = \mathcal{L}_{\mathrm{det}}(\boldsymbol{h}^v) + \mathcal{L}_{\mathrm{cor}}(\boldsymbol{h}^v). \tag{10}$$

3.2.3 ViLEM with Local Visual Feature

We also perform ViLEM with local visual features to associate local text semantics with multi-level local visual information and facilitate fine-grained image-text alignment. To enable interactions between local image and text features, we activate cross-attention modules in the first M_1 layer of the textual encoder. Given the image $\boldsymbol{v}_i$ and its corresponding edited text $\boldsymbol{t}_i^{edit}$, we extract local image patch features from all intermediate layers $\mathcal{H}_i^v = \{\boldsymbol{H}_{il}^v\}_{l=1}^{\mathrm{N}}$, where l is the layer index of visual encoder. Then we feed edited text $\boldsymbol{t}_i^{edit}$ into a textual encoder. In the m-th layer ($m \in \{1, 2, ..., \mathrm{M}_1\}$), the cross attention module takes intermediate word features as queries and image patch features from the l_m-th layer as keys and values to aggregate word-related visual concept. The l_m is calculated as follows:

$$l_m = \lfloor \frac{\mathrm{N}}{\mathrm{M}_1} \rfloor (m-1) + 1, \tag{11}$$

which ensures that image patch features of all levels are uniformly utilized for ViLEM.

At last, we take the output features of the M_1-th layer to perform binary classification on each word feature to detect the correctness of each word and predict the corresponding correct word for each wrong word. The loss for ViLEM with local visual features is formulated as:

$$\mathcal{L}_{\mathrm{EML}} = \mathcal{L}_{\mathrm{det}}(\mathcal{H}^v) + \mathcal{L}_{\mathrm{cor}}(\mathcal{H}^v), \tag{12}$$

where $\mathcal{L}_{\mathrm{det}}(\mathcal{H}^v)$ and $\mathcal{L}_{\mathrm{cor}}(\mathcal{H}^v)$ are computed following Equations 8 and 9 but text error detection and correction are performed with multi-level local visual features $\mathcal{H}^v$ instead of global visual feature $\boldsymbol{h}^v$.

3.3. Pre-training Objectives

We train the network with three losses jointly to facilitate global image-text alignment, and establish detailed associations between local text semantics and multi-granularity visual features. The total loss is formulated as:

$$\mathcal{L} = \mathcal{L}_{\mathrm{align}} + \lambda_1 \mathcal{L}_{\mathrm{EML}} + \lambda_2 \mathcal{L}_{\mathrm{EMG}}, \tag{13}$$

where λ_1 and λ_2 are hyper-parameters to adjust the effect of ViLEM losses.

4. Experiments

4.1. Datasets

Pre-training Datasets. We pre-train our model with two image-text datasets: (1) **CC4M** contains 4 million images and 5.1 million captions from Conceptual Captions (CC3M) [32], SBU [26], MSCOCO [19] and Visual Genome [14]. (2) **CC13M** consists of CC4M and CC12M [2] (about 3.3 million image URLs are now invalid for us), which contains 13M images and 14.1M captions in total. Details are shown in the supplementary materials.

Downstream Datasets. We conduct downstream image-text retrieval evaluation on two widely used datasets: MSCOCO [19] and Flick30K [28]. In addition, we validate the effectiveness of ViLEM on improving the discriminativeness to local text semantics with Winoground [36] dataset. To further demonstrate the generalization ability of our model to video-text tasks, we conduct experiments on a public video-text retrieval dataset MSR-VTT [39]. The details of these downstream datasets and the evaluation metrics can be found in the supplemental material.

4.2. Implementation Details

Our model adopts $\mathrm{BERT}_{\mathrm{base}}$ [12] as textual encoder and a ViT-B/16 [6] initialized with weights pre-trained on ImageNet-1k as the visual encoder. We randomly replace word tokens with 15% probability for the knowledge-based text edition. We use the AdamW [20] optimizer with a weight decay of 0.02. The learning rate is warmed-up to $3e^{-4}$ in the first 2000 iterations and decays to $1e^{-5}$ following a cosine schedule. We pre-train the model for 20 epochs with a batch size of 2048 on 32 NVIDIA A100 GPUs. We take the image resolution of 256× 256 for pre-training and increase the image resolution to 384× 384 for fine-tuning. The momentum coefficient for updating momentum encoders is set as 0.995, and the queue size N_q is set as 65536. The learnable temperature hyper-parameter for contrastive loss is initialized to 0.07. The loss weight λ_1 and λ_2 are set as 0.8 and 0.2 respectively. More implementation details can be found in the supplementary materials.

4.3. Image-Text Retrieval

Comparison with the State-of-the-Art. We compare with state-of-the-art methods on Flickr30K and MSCOCO datasets. As shown in Table 1, under a fair comparison experimental setting (excluding VSE$\infty^{*\dagger}$ and COOKIE*† as they use 940M tagged images for visual-encoder pre-training), our method surpasses all dual-encoder methods by a large margin under all evaluation metrics. Specifically, compared with the current state-of-the-art dual-encoder method COTS [21] with 5.3M pre-training data, our method with 5.1M pre-training data achieves higher performance by 4.2% and 2.9% on the R@1 of image-to-text and text-to-image retrieval of Flickr30K dataset. On the MSCOCO dataset, we also surpass COTS (5.3M) by 2.1% on the R@1 of both image-to-text and text-to-image retrieval. Moreover, our method (5.1M) outperforms COTS pre-trained on 15.3M image-text pairs with only 1/3 data. The performance of our method is further improved when leveraging a larger pre-training dataset CC13M, even outperforming VSE$\infty^{*\dagger}$ and COOKIE*†. Furthermore, our method

Table 1. Comparative results for fine-tuned image-text retrieval results on the Flickr30K (1K) test set and MSCOCO (5K) test set. We make comparisons with both dual-encoder methods and joint-encoder methods. Our method surpasses previous state-of-the-art dual-encoder methods by a large margin and achieves comparable performance but much faster inference speed *w.r.t.* latest joint-encoder methods. (64× and 7240× faster than ALBEF and VinVL-base.) **Higher** R@K indicates better performance. **PT Pairs**: the number of image-text pairs for pre-training. † is ensemble result of two models. ∗ models use 940M tagged images for visual encoder pre-training.

Model	PT Pairs	Flickr30K (1K test set)							MSCOCO (5K test set)						
		image→text			text→image			R@S	image→text			text→image			R@S
		R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
Joint-Encoder:															
Pixel-BERT-X152 [10]	5.6M	87.0	98.9	99.5	71.5	92.1	95.8	544.8	63.6	87.5	93.6	50.1	77.6	86.2	458.6
Unicoder-VL [15]	3.8M	86.2	96.3	99.0	71.5	91.2	95.2	539.4	62.3	87.1	92.8	48.4	76.7	85.9	453.2
UNITER-base [4]	9.6M	85.9	97.1	98.8	72.5	92.4	96.1	542.8	64.4	87.4	93.1	50.3	78.5	87.2	460.9
ERNIE-ViL-base [41]	3.8M	86.7	97.8	99.0	74.4	92.7	95.9	546.5	–	–	–	–	–	–	–
VILLA-base [7]	9.6M	86.6	97.9	99.2	74.7	92.9	95.8	547.1	–	–	–	–	–	–	–
Oscar-base [18]	6.5M	–	–	–	–	–	–	–	70.0	91.1	95.5	54.0	80.8	88.5	479.9
ViLT [13]	9.9M	83.5	96.7	98.6	64.4	88.7	93.8	525.7	61.5	86.3	92.7	42.7	72.9	83.1	439.2
VinVL-base [42]	8.9M	–	–	–	–	–	–	–	**74.6**	**92.6**	**96.3**	**58.1**	**83.2**	**90.1**	**494.9**
ALBEF [16]	5.1M	**94.3**	**99.4**	**99.8**	**82.8**	**96.7**	**98.4**	**571.4**	73.1	91.4	96.0	56.8	81.5	89.2	488.0
Dual-Encoder:															
VSE$\infty^{*\dagger}$ [3]	–	88.7	98.9	99.8	76.1	94.5	97.1	555.1	68.1	90.2	95.2	52.7	80.2	88.3	474.7
COOKIE*† [38]	5.9M	89.0	98.9	99.7	75.6	94.6	97.2	555.0	71.6	90.9	95.4	54.5	81.0	88.2	481.6
LightningDOT [34]	9.5M	83.9	97.2	98.6	69.9	91.1	95.2	535.9	60.1	85.1	91.8	45.8	74.6	83.8	441.2
COOKIE [38]	5.9M	84.7	96.9	98.3	68.3	91.1	95.2	534.5	61.7	86.7	92.3	46.6	75.2	84.1	446.6
COTS [21]	5.3M	88.2	98.5	99.7	75.2	93.6	96.5	551.7	66.9	88.8	94.0	50.5	77.6	86.1	463.9
COTS [21]	15.3M	90.6	98.7	99.7	76.5	93.9	96.6	556.0	69.0	90.4	94.9	52.4	79.0	86.9	472.6
Ours	5.1M	92.4	**99.2**	99.7	78.1	94.6	97.0	561.0	69.0	90.7	95.1	52.6	79.4	87.2	474.0
Ours	14.1M	**93.6**	99.0	**99.7**	**80.5**	**96.0**	**98.0**	**566.8**	**73.2**	**91.8**	**95.9**	**54.5**	**80.6**	**88.2**	**484.2**

Table 2. Comparison for image-text retrieval results (without fine-tuning) on the MSCOCO (5K) test set.

Model	image→text			text→image			R@S
	R@1	R@5	R@10	R@1	R@5	R@10	
CLIP [29]	58.4	81.5	88.1	37.8	62.4	72.2	400.4
ALIGN [11]	58.6	83.0	87.9	45.6	69.8	78.6	423.5
COTS [21]	60.4	84.7	91.7	43.8	71.6	81.3	433.5
Ours	**65.6**	**88.0**	**93.8**	**47.7**	**75.2**	**84.5**	**454.8**

also achieves comparable performance with the latest joint-encoder methods VinVL-base and ALBEF while having much higher retrieval efficiency. Specifically, we measure the inference time for performing image-text retrieval on the MSCOCO 5K test set. Our method is 64× and 7240× faster than ALBEF and VinVL-base. More details of inference time measurement are shown in the suppl. materials.

Comparison of Retrieval Results without Fine-tuning. Following previous works [13, 21], we report the retrieval performance without fine-tuning on the MSCOCO dataset and make comparisons with recent powerful dual-encoder methods. As shown in Table 2, with a similar pre-training data size, we surpass the COTS [21] by 5.2% and 3.9% on the R@1 of image-to-text retrieval and text-to-image retrieval. Moreover, our method also outperforms CLIP [29] and ALIGN [11], which utilize 28× and 128× pre-training data than our method respectively.

Zero-shot Text-to-Video Retrieval. We perform zero-shot text-to-video retrieval to validate the generalization ability of our image-text model to the video-text task. Specifically, we uniformly sample 8 frames per video and use the mean frame features as global video features. The video-text similarity scores can be calculated by the dot product between global video features and global text features. Text-to-video retrieval results on the MSR-VTT dataset are reported in Table 3. It can be seen that our pure image-text model outperforms previous state-of-the-art video-text methods even without complex temporal modeling of video.

Table 3. Zero-shot text-to-video retrieval results on the MSRVTT (1K) test set. **Lower** MedR indicates better performance.

Model	PT Pairs	R@1	R@5	R@10	MedR↓
MIL-NCE [23]	Video 120M	9.9	24.0	32.4	29.6
TACo [40]	Video 120M	9.8	25.0	33.4	29.0
SupportSet [27]	Video 120M	12.7	27.5	36.2	24.0
Frozen [1]	Image 3M+Video 2.5M	18.7	39.5	51.6	10.0
BridgeFormer [8]	Image 3M+Video 2.5M	26.0	46.4	56.4	7.0
Ours	Image 14.1M	**27.6**	**49.8**	**60.7**	**6.0**

4.4. Vision-linguistic Stress Testing

To validate the effectiveness of ViLEM on improving discriminativeness to local text semantics, we perform vision-linguistic stress testing on the Winoground dataset. Each sample in the Winoground dataset consists of two image-text pairs with only minor differences between them. The model needs to correctly match the two image-text pairs, which requires a powerful discrminativeness to local image and text semantics. We report the text score in Table 4 following [36], which reflects the proportion of samples where both images are correctly matched with their paired texts. Compared with vanilla dual-encoder without ViLEM, our method achieves 3.4%, 6.4%, and 15.4% improvement in recognizing object differences, relational differences, and co-occurrence of both differences. In addition, the overall

Table 4. Comparison with vanilla dual-encoder and state-of-the-art methods on the Winoground dataset. **Object**, **Relation**, and **Both** indicate the matching accuracy for samples with object difference, relation difference, and both differences. **1 Pred** and **2 Preds** indicate the matching accuracy for samples with one predicate and two predicates respectively. **All** reflects the overall performance.

Model	Object	Relation	Both	1 Pred	2 Preds	All
Joint-Encoder:						
UNITER [4]	34.0	30.0	42.3	35.3	24.1	32.3
ViLBERT [22]	29.1	19.3	34.6	24.0	23.2	23.8
ViLLA [7]	33.3	27.0	38.5	33.2	21.3	30.0
ViLT [13]	31.9	36.9	30.8	35.3	33.3	34.8
FLAVA$_{itm}$ [33]	31.9	30.0	53.8	36.3	21.3	32.3
VinVL [42]	**36.9**	**37.8**	**42.3**	**39.4**	**33.3**	**37.8**
Dual-Encoder:						
FLAVA$_{constrastive}$ [33]	23.4	23.6	50.0	26.4	22.2	25.3
CLIP [29]	**34.8**	22.8	**80.8**	35.3	18.5	30.8
w/o ViLEM	30.5	29.1	50.0	33.9	24.1	31.2
Ours	33.9	**35.5**	65.4	**38.7**	**30.6**	**36.5**

performance of our method exceeds all dual-encoder methods and joint-encoder methods except VinVL. Note that CLIP [29] utilizes 28× pre-training data than our method.

4.5. Ablation Studies

In this section, we discuss the effectiveness of our proxy task ViLEM and multi-granularity interaction framework via evaluating different models for zero-shot image-text retrieval on MSCOCO. We sample 1M image-text pairs from CC3M as pre-training dataset due to the limitation of computation resources.

Are ViLEM with local and global visual features effective? Yes. As shown in Table 5, models D and G which perform ViLEM with local and global visual features respectively outperform the baseline model A, indicating that associating local text semantics with high-level global visual features or multi-level local visual features both benefit the global image-text alignment. Moreover, the model H that performs ViLEM with multi-granularity visual features achieves further performance improvement, which shows that the effectiveness of our multi-granularity interaction framework and ViLEM with global and local features are complementary for improving image-text retrieval.

Are error detection and correction effective tasks? Yes. As shown in Table 5, models B and E that perform text error detection outperform baseline model A, indicating the benefits of learning local image-text matching relationship for retrieval. Both models C and F perform text error correction outperform baseline model A, which shows that enforcing the model to reason correct words with visual information also facilitates image-text retrieval. Retrieval performance is further improved when combining text error detection and correction into ViLEM task, *i.e.* models D and G.

Does the position to compute ViLEM losses matter? Yes, we choose to compute ViLEM losses with output features from 6-th layer of textual encoder for the following reasons. (1) Using features from a higher layer for ViLEM,

Table 5. Ablation studies on different components of our method, including text error detection (Det), and correction (Cor) with local and global visual features respectively.

	Local Det	Local Cor	Global Det	Global Cor	image→text R@1	image→text R@5	image→text R@10	text→image R@1	text→image R@5	text→image R@10	R@S
A	–	–	–	–	26.4	53.1	66.2	19.4	42.9	54.8	262.8
B	✓				28.1	54.6	66.8	20.5	43.7	55.6	269.3
C		✓			28.4	55.2	67.0	21.0	43.8	55.2	270.6
D	✓	✓			29.1	**55.5**	67.1	20.7	44.5	55.9	272.8
E			✓		27.3	54.6	66.4	20.5	44.0	55.8	268.6
F				✓	27.4	54.4	66.3	20.6	43.9	55.7	268.3
G			✓	✓	28.0	54.3	66.9	20.9	44.6	56.4	271.1
H	✓	✓	✓	✓	**29.1**	55.3	**68.3**	**22.0**	**45.7**	**57.7**	**278.1**

Table 6. Ablation study on the position to compute ViLEM losses.

Layer Index	image→text R@1	image→text R@5	image→text R@10	text→image R@1	text→image R@5	text→image R@10	R@S
4	27.4	55.0	67.2	21.3	45.2	56.8	272.9
6	**29.1**	55.3	68.3	**22.0**	**45.7**	**57.7**	**278.1**
8	28.1	**56.0**	**68.7**	21.3	45.0	56.8	275.9
10	27.8	55.0	67.4	21.2	44.6	56.6	272.6
12	27.9	55.0	66.8	21.3	45.0	56.8	272.8

Table 7. Comparisons between different sub-module options.

	Method	image→text R@1	image→text R@5	image→text R@10	text→image R@1	text→image R@5	text→image R@10	R@S
A	w/o ViLEM	26.4	53.1	66.2	19.4	42.9	54.8	262.8
B	MLM	27.5	55.0	66.4	21.1	44.4	56.3	270.7
C	Edited text cont.	26.6	53.9	66.5	19.8	43.1	55.1	265.0
D	Highest-level	28.3	54.7	67.1	21.1	44.6	56.3	272.1
E	Local-global Unify	28.9	**55.5**	67.6	20.6	44.1	56.2	272.9
F	Random edition	28.4	53.9	66.8	21.7	45.4	57.1	273.3
G	Ours	**29.1**	55.3	**68.3**	**22.0**	**45.7**	**57.7**	**278.1**

such as the 10-th or 12-th layer, degrades the performance. We argue that in this case too many encoder layers undertake the task of text encoding and decoding simultaneously, which interferes the encoding of text features. (2) Computing ViLEM loss with features from a lower layer, such as the 4-th layer, also yields worse results due to the insufficient interaction between visual features and textual features. (3) Computing ViLEM loss with features from the 8-th layer achieves slightly worse performance and requires more computation cost.

ViLEM *vs.* Masked Language Modeling. Different from our ViLEM, Masked Language Modeling (MLM) only considers a proportion of word token and may ignore visual information to recover the masked words. Comparing Model B with G in Table 7, pre-training with ViLEM shows significant advantages over pre-training with Masked Language Model (MLM), which clearly validates the superiority of our method beyond MLM.

ViLEM *vs.* Contrastive learning with edited text. We take edited texts and corresponding images as hard negative pairs for contrastive learning, *i.e.* model C in Table 7. It achieves performance improvement compared to baseline model A but has a large performance gap with our method G. We argue that coarse-grained global alignment is insufficient for capturing fine-grained semantic association.

Multi-level *vs.* Single-level local visual features. Instead

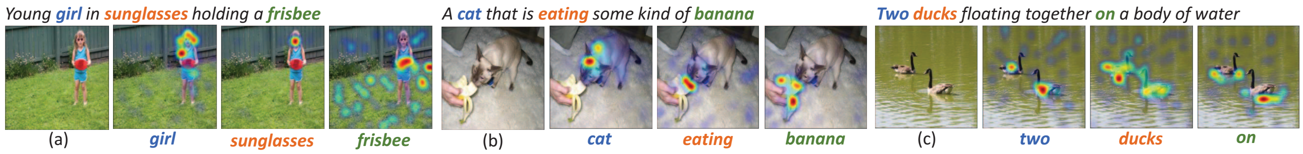
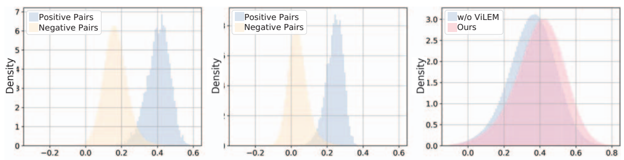


Figure 3. Grad-CAM visualizations on the cross-attention maps corresponding to individual words.



(a) Score w/o ViLEM (b) Score of our method (c) Normalized Margin

Figure 4. Distribution of similarity scores for positive and negative pairs and normalized margins between positive and negative pairs.

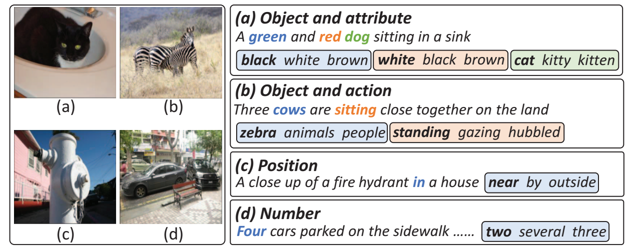


Figure 5. Visualization of text error detection and correction. Different colored words in captions indicate the detected wrong words, and the top-3 candidates for correction are shown in the corresponding colored text boxes.

of using multi-level local visual features, Model D in Table 7 performs ViLEM with global visual features and single-level local visual features. We observe that the performance drops due to the lack of guidance on the intermediate visual features. But model D also outperforms the baseline model A, which validates the effectiveness of our ViLEM task.

Separate *vs.* Joint use of global and local visual features. A straightforward approach to exploit global and local visual features for ViLEM is concatenating them and then feeding them into cross-attention modules. We experiment with this approach, *i.e.*, model E in Table 7, and observe that it achieves worse results than model G. Moreover, model E achieves comparable performance with the model that only uses local visual features (model D in Table 5), indicating model E may only focus on local visual features for ViLEM and lacks the regularization on global visual features.

Knowledge-based *vs.* Random text edition. Replacing words by random sampling from vocabulary (Model F in Table 7) rather than the knowledge-based edition with a pre-trained language model (Model G in Table 7) may generate meaningless texts and reduces the difficulty of ViLEM, leading to performance degradation.

Distribution of similarity scores and normalized margins. We show the distribution of similarity scores and normalized margins in Figure 4 to observe the effect of ViLEM on the image-text embedding space. It can be seen that ViLEM reduces the variance of similarity scores of positive and negative pairs while enlarging the normalized margins between positive and negative pairs from 0.35 to 0.40.

4.6. Qualitative Analysis

Fine-grained image-text association. We visualize the word-patch cross-attention maps corresponding to individual words through Grad-CAM [31], which shows that fine-grained association between images and texts is properly established. In Figure 3(a), our model attends to corresponding regions of different objects, even fine-grained ones like "sunglasses" and "frisbee". Figure 3(b) shows that our model correlates action information across visual and language. When recognizing "eating", our model focus on the region where cat's mouth touches the banana. Moreover, our model can capture abstract visual concepts, *i.e.*, number "two" and spatial relation "on" as shown in Figure 3(c).

Proofreading negative texts with ViLEM. Figure 5 visualize examples of our model applied to negative texts with different kinds of local errors. Common types of errors, such as the object error ("dog" in (a) and "cows" in (b)), the attribute error ("green" and "red" in (a)), and the action error ("sitting" in (b)) can be well detected and corrected by the model. Moreover, our model can also deal with position error ("in" in (c)) and counting error ("four" in (d)).

5. Conclusion

In this work, we propose a novel proxy task, Visual-Language Error Modeling (ViLEM) for image-text retrieval, which injects detailed image-text association into "dual-encoder" architecture. A multi-granularity interaction framework is proposed to perform ViLEM via interacting with both high-level visual context and multi-level local visual information while maintaining high efficiency for retrieval. Extensive experiments on image-text retrieval and vision-linguistic stress testing clearly demonstrate the superiority of our method. Our model also shows the generalization capability to video-text data.

Acknowledgments This work is supported by the National Key RD Program of China (Grant No. 2018AAA0102802, 2018AAA0102800), Beijing Natural Science Foundation (Grant No. JQ21017, L223003, M22005), the Natural Science Foundation of China (Grant No. 61972397, 62036011, 62192782, 61721004, U2033210, 62172413, 62192785, 61972071), the Major Projects of Guangdong Education Department for Foundation Research and Applied Research (Grant: 2017KZDXM081, 2018KZDXM066), Guangdong Provincial University Innovation Team Project (Project No.: 2020KCXTD045), Research Fund of ARC Lab, Tencent PCG.

References

[1] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738, 2021. 6

[2] 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

[3] Jiacheng Chen, Hexiang Hu, Hao Wu, Yuning Jiang, and Changhu Wang. Learning the best pooling strategy for visual semantic embedding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15789–15798, 2021. 6

[4] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. 2019. 2, 6, 7

[5] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*, 2019. 3

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 5

[7] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. *Advances in Neural Information Processing Systems*, 33:6616–6628, 2020. 6, 7

[8] Yuying Ge, Yixiao Ge, Xihui Liu, Dian Li, Ying Shan, Xiaohu Qie, and Ping Luo. Bridging video-text retrieval with multiple choice questions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16167–16176, 2022. 6

[9] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 3

[10] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*, 2020. 6

[11] 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, 6

[12] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019. 2, 4, 5

[13] 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. 6, 7

[14] 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

[15] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11336–11344, 2020. 6

[16] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 2, 6

[17] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 2

[18] 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. 2, 6

[19] 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

[20] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018. 5

[21] 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, 3, 5, 6

[22] 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. 2, 7

[23] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9879–9889, 2020. 6

[24] Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhanskyi. Gector–grammatical error correction: Tag, not rewrite. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–170, 2020. 3

[25] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2, 4

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

[27] Mandela Patrick, Po-Yao Huang, Yuki Asano, Florian Metze, Alexander Hauptmann, Joao Henriques, and Andrea Vedaldi. Support-set bottlenecks for video-text representation learning. *arXiv preprint arXiv:2010.02824*, 2020. 6

[28] 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. 5

[29] 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. 1, 2, 6, 7

[30] Sascha Rothe, Jonathan Mallinson, Eric Malmi, Sebastian Krause, and Aliaksei Severyn. A simple recipe for multilingual grammatical error correction. In *ACL/IJCNLP (2)*, 2021. 3

[31] 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

[32] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 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 (Volume 1: Long Papers)*, pages 2556–2565, 2018. 5

[33] 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. 7

[34] Siqi Sun, Yen-Chun Chen, Linjie Li, Shuohang Wang, Yuwei Fang, and Jingjing Liu. Lightningdot: Pre-training visual-semantic embeddings for real-time image-text retrieval. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 982–997, 2021. 2, 6

[35] Maksym Tarnavskyi, Artem Chernodub, and Kostiantyn Omelianchuk. Ensembling and knowledge distilling of large sequence taggers for grammatical error correction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3842–3852, 2022. 3

[36] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248, 2022. 5, 6

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[38] Keyu Wen, Jin Xia, Yuanyuan Huang, Linyang Li, Jiayan Xu, and Jie Shao. Cookie: Contrastive cross-modal knowledge sharing pre-training for vision-language representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2208–2217, 2021. 1, 2, 6

[39] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016. 5

[40] Jianwei Yang, Yonatan Bisk, and Jianfeng Gao. Taco: Token-aware cascade contrastive learning for video-text alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11562–11572, 2021. 6

[41] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3208–3216, 2021. 6

[42] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2021. 2, 6, 7