Semantics-enhanced Cross-modal Masked Image Modeling for Vision-Language Pre-training

Haowei Liu[1,2*]**, Yaya Shi**[3*]**, Haiyang Xu**[4†]**, Chunfeng Yuan**[1†]**, Qinghao Ye**[4]
Chenliang Li[4]**, Ming Yan**[4]**, Ji Zhang**[4]**, Fei Huang**[4]**, Bing Li**[1]**, Weiming Hu**[1,2,5]

[1]MAIS, Institute of Automation, Chinese Academy of Sciences, China
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, China
[3]University of Science and Technology of China [4]Alibaba Group
[5]School of Information Science and Technology, ShanghaiTech University, China
liuhaowei2019@ia.ac.cn, shiyaya@mail.ustc.edu.cn
{shuofeng.xhy, ym119608}@alibaba-inc.com, {cfyuan, bli, wmhu}@nlpr.ia.ac.cn

Abstract

In vision-language pre-training (VLP), masked image modeling (MIM) has recently been introduced for fine-grained cross-modal alignment. However, in most existing methods, the reconstruction targets for MIM lack high-level semantics, and text is not sufficiently involved in masked modeling. These two drawbacks limit the effect of MIM in facilitating cross-modal semantic alignment. In this work, we propose a semantics-enhanced cross-modal MIM framework (SemMIM) for vision-language representation learning. Specifically, to provide more semantically meaningful supervision for MIM, we propose a local semantics enhancing approach, which harvest high-level semantics from global image features via self-supervised agreement learning and transfer them to local patch encodings by sharing the encoding space. Moreover, to achieve deep involvement of text during the entire MIM process, we propose a text-guided masking strategy and devise an efficient way of injecting textual information in both masked modeling and reconstruction target acquisition. Experimental results validate that our method improves the effectiveness of the MIM task in facilitating cross-modal semantic alignment. Compared to previous VLP models with similar model size and data scale, our SemMIM model achieves state-of-the-art or competitive performance on multiple downstream vision-language tasks.

Keywords: masked image modeling, local semantics enhancing, cross-modal alignment

1. Introduction

Learning transferable vision-language representations for various downstream multimodal tasks (*e.g.* image-text retrieval, visual question answering and image captioning) by pre-training on large-scale image-text datasets has attracted extensive attention in recent years (Tan and Bansal, 2019; Huang et al., 2020; Li et al., 2021, 2022a; Ji et al., 2023a). For vision-language pre-training (VLP), the most crucial thing is to align the representation spaces of the two modalities. To this end, contrastive learning between global image and text features is commonly adopted. However, such coarse-grained alignment usually ignores detailed information in image and text. To achieve finer-grained cross-modal alignment, recent work (Ge et al., 2022; He et al., 2022b; Geng et al., 2022; Bao et al., 2022) has introduced the masked image modeling (MIM) task into VLP.

Despite the achievements of MIM in the vision domain, leveraging MIM to facilitate cross-modal semantic alignment and promote vision-language representation learning still faces significant challenges. **1) Semantic issue.** Different from text

whose words are naturally discrete and abstract, the region information of image is continuous and ambiguous. Therefore, without semantically meaningful supervision for MIM, the inconsistency in semantic levels between vision and language will impair cross-modal alignment. However, in most existing methods, the reconstruction targets provided for MIM lack high-level semantics. For instance, VL-BEiT (Bao et al., 2022) uses a discrete variational autoencoder (dVAE) to encode image patches into discrete codes as MIM's supervision. As dVAE is trained through image reconstruction which simply minimizes pixel-level differences between original and reconstructed images, its patch encodings lack high-level visual semantics. VLMAE (He et al., 2022b) and M3AE (Geng et al., 2022) take the raw pixels of the masked regions as the reconstruction targets of MIM, thus facing even more severe semantic issue. **2) Insufficient text involvement.** Most previous methods directly introduce MIM into VLP without elaborate designs for deep involvement of text. For example, in MILES (Ge et al., 2022) and MaskCLIP (Dong et al., 2023b), text features do not directly participate in the MIM process, which inevitably limits their contribution to cross-modal semantic interaction. As for the selection of masked patches, most methods simply adopt a

---

[*]Equal contribution.
[†]Corresponding authors.

14664

random masking strategy without considering the interaction between image and text.

To address these issues, we propose a semantics-enhanced cross-modal MIM framework (dubbed as **SemMIM**) to improve the effectiveness of MIM in facilitating vision-language semantic alignment. In this framework, we first propose to inject high-level semantics into the local encodings of image patches via self-supervised agreement learning and sharing encoding space, which can thus provide more semantically meaningful supervision for MIM. Moreover, we make elaborate designs for deep involvement of text during the MIM process (*i.e.* masking strategy, masked modeling and reconstruction target acquisition) to further promote cross-modal interaction. Specifically, we adopt a fusion architecture which includes an image encoder, a text encoder and a fusion encoder in the model. For masked image modeling, we set an extra encoding head on top of the image encoder, which projects the masked visual tokens into an encoding space.

Due to the small granularity of image patches, it's challenging to directly acquire high-level semantics from them. To tackle this, we propose to harvest high-level semantics from global visual features via self-supervised agreement learning, and transfer them into patch encodings by sharing the same encoding space. Specifically, we derive a momentum image encoder from the original image encoder, whose parameters are obtained via the exponential moving average approach. Similarly, a momentum encoding head is also derived on top of the momentum image encoder. Given an input image, we apply random augmentation to obtain two different views of it, which are fed into the original and the momentum image encoders respectively. Then we employ the encoding heads to map the obtained global visual features into the encoding space. By learning the agreement between different views, we can acquire high-level visual semantics and shape the semantic structure of the encoding space. In the meantime, we share the encoding space for global and local features, *i.e.*, using the same momentum image encoder and encoding head to transform patch features into visual encodings. The learned high-level semantics are thus transferred to the local patch encodings. In this manner, we obtain semantics-enhanced local encodings and provide more semantically meaningful reconstruction targets for MIM.

Moreover, to further enhance MIM's ability of facilitating cross-modal semantic alignment, we make three elaborate designs for deep text involvement throughout the entire MIM process. 1) We leverage the semantic relevance between image patches and the paired text to select masked patches, which can encourage the utilization of textual information during masked modeling. 2) We devise an efficient way of fusing textual information while masked modeling. This allows recovering the masked regions by reasoning both contextual visual information and paired textual information, and thus promotes cross-modal semantic interaction. 3) We fuse text features when transforming image patches into visual encodings with the momentum image encoder, so that high-level textual semantics can also be injected into the encodings.

Visualization results demonstrate that the proposed SemMIM framework can learn high-level visual semantics effectively. And extensive experiments show our model achieves state-of-the-art or competitive performance on multiple downstream vision-language tasks, which validates the superiority of our method.

Our contributions can be summarized as follows:

- We propose a semantics-enhanced, text-deeply-involved MIM framework to facilitate fine-grained cross-modal semantic alignment in vision-language pre-training.

- High-level visual semantics are injected into image patch encodings via self-supervised learning and shared encoding space, thus providing more semantically meaningful reconstruction targets for MIM.

- Our elaborate designs for deep text involvement throughout the entire MIM process further promote cross-modal interaction, including masking strategy, masked modeling and reconstruction target acquisition.

2. Related Work

2.1. Vision-Language Pre-training

Large-scale vision-language pre-training (Lu et al., 2019; Huang et al., 2020; Zhou et al., 2020; Li et al., 2021; Xu et al., 2021; Li et al., 2022a; Xu et al., 2023; Shi et al., 2023) has shown promising performance on various downstream vision-language tasks. To mine the associations between image and text, various pre-training objectives are proposed, which can be divided into two categories: discriminative tasks and generative tasks. Two commonly-used discriminative tasks are image-text contrastive learning (ITC) and image-text matching (ITM). ITC is used to align the global representations of image and text. ITM aims to predict whether an image and a text match with each other in the cross-modal representation space. Generative tasks aim to reconstruct the corrupted text (or image) with the help of vision (or text) modality. For instance, masked language modeling (MLM) predicts the masked tokens by reasoning contextual text information and the

paired image information (Wu et al., 2022; Chen et al., 2023). In this work, we focus on improving the effectiveness of the Masked Image Modeling (MIM) task, so as to better facilitate fine-grained cross-modal alignment.

2.2. Masked Image Modeling in VLP

Inspired by the success of MLM in multimodal learning, recent methods (Ge et al., 2022; He et al., 2022b; Geng et al., 2022; Bao et al., 2022) have introduced MIM into VLP for fine-grained cross-modal alignment. VLMAE (He et al., 2022b) and M3AE (Geng et al., 2022) follow MAE (He et al., 2022a) to recover the raw pixels of the masked image regions with an autoencoder architecture. Following BEiT (Bao et al., 2021), VL-BEiT (Bao et al., 2022) and BEiT-v3 (Wang et al., 2022b) utilize a pre-trained discrete variational autoencoder (dVAE) to encode image patches into discrete codes as the supervision for MIM. As dVAE is trained through image reconstruction which minimizes pixel-level differences between original and reconstructed images, its encodings concentrate more on low-level visual information. Therefore, the above methods can't provide high-level semantic supervision for MIM. Such inconsistency in semantic levels between vision and language impairs MIM's effectiveness in facilitating cross-modal alignment. Another issue of existing methods is insufficient text involvement in MIM. For instance, in MILES (Ge et al., 2022) and MaskCLIP (Dong et al., 2023b), text features do not directly participate in the MIM process. Thus during masked modeling, the model only reasons visual information to recover masked regions, without direct vision-language interaction. This inevitably limits MIM's benefit to cross-modal semantic alignment. In this work, we propose a semantics-enhanced cross-modal MIM framework, which tackles the above stated issues by local semantics enhancing and text-deeply-involved cross-modal MIM.

3. Method

3.1. Model Architecture

As shown in Figure 1(a), our SemMIM framework contains an image encoder $E_I(\cdot)$, a text encoder $E_T(\cdot)$, a multi-modal fusion encoder $E_F(\cdot)$, a decoder $D(\cdot)$ and an additional momentum image encoder $\hat{E}_I(\cdot)$. All the encoders are Transformer models. The image encoder splits the input image into patches and encodes them as a sequence of visual features $\{v_{cls}, v_1, v_2, \ldots, v_N\}$. The text encoder tokenizes the input text into tokens and projects them into a sequence of word features $\{w_{cls}, w_1, w_2, \ldots, w_L\}$. v_{cls} and w_{cls} are the features of the [CLS] token for image and text respec-

tively. N is the number of image patches, and L is the number of text tokens. The multimodal fusion encoder fuses the visual features and the word features through cross attention. The obtained cross-modal representations are fed into the decoder for sequence to sequence learning, which equips our model with both understanding and generation abilities.

3.2. Semantics-enhanced Cross-modal Masked Image Modeling

3.2.1. Local Semantics Enhancing

As pointed in PeCo (Dong et al., 2023a) and BEiT v2 (Peng et al., 2022), a significant drawback of the offline visual tokenizers (*e.g.* dVAE) adopted by previous methods is that their encodings lack high-level semantics. Instead, they concentrate more on low-level visual information such as color and texture. Taking such visual encodings as supervision inevitably hinders MIM's objective of cross-modal semantic alignment. However, as the granularity of image patches is rather small, it's challenging to directly acquire high-level semantics from them. Therefore, we propose to harvest high-level semantics from global image features via self-supervised learning, which are further transferred to local patch features by sharing the same encoding space.

Unlike using pre-trained dVAE or raw pixels as reconstruction targets, we adopt an online approach and utilize the model itself to provide supervision for MIM. As shown in Figure 1(a), we derive a momentum image encoder $\hat{E}_I$ from the original image encoder E_I. $\hat{E}_I$ shares the same structure with E_I, and its parameters are obtained via exponential moving averaging the parameters of E_I during the iterative updating process of the model:

$$\theta_{\hat{E}_I} = m\theta_{\hat{E}_I} + (1-m)\theta_{E_I}, \tag{1}$$

where m is a momentum coefficient. Besides, we add two encoding heads H_I and $\hat{H}_I$ on top of the original and the momentum image encoders respectively. Each encoding head is an MLP module, which transforms the extracted image features into categorical distributions in a K-dimensional encoding space. Here the parameters of $\hat{H}_I$ are also obtained via exponential moving averaging the parameters of H_I. Previous methods such as dVAE encode image patches into discrete codes, *i.e.*, their categorical distributions are one-hot. Instead, to preserve more information, we adopt a softmax form of categorical distributions. In other words, the reconstruction target of each masked image patch is a soft label.

To obtain more semantically meaningful visual encodings, we first harvest high-level semantic information from global image features. Specifically,

(a) Model architecture and pre-training objectives

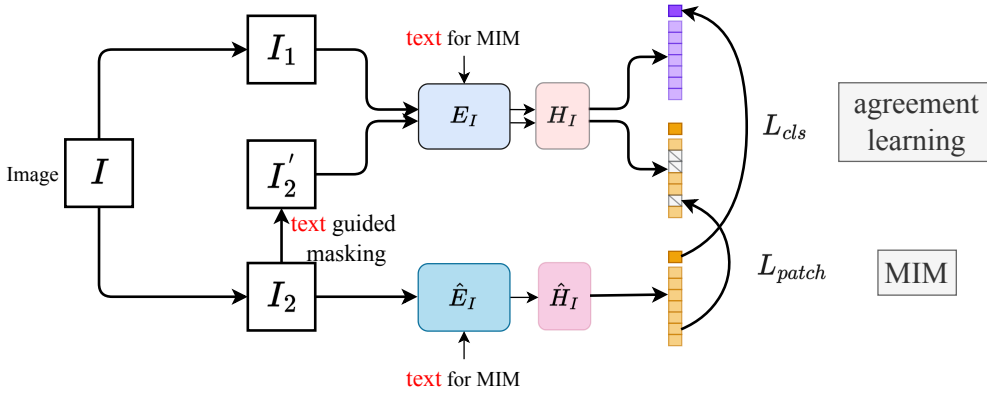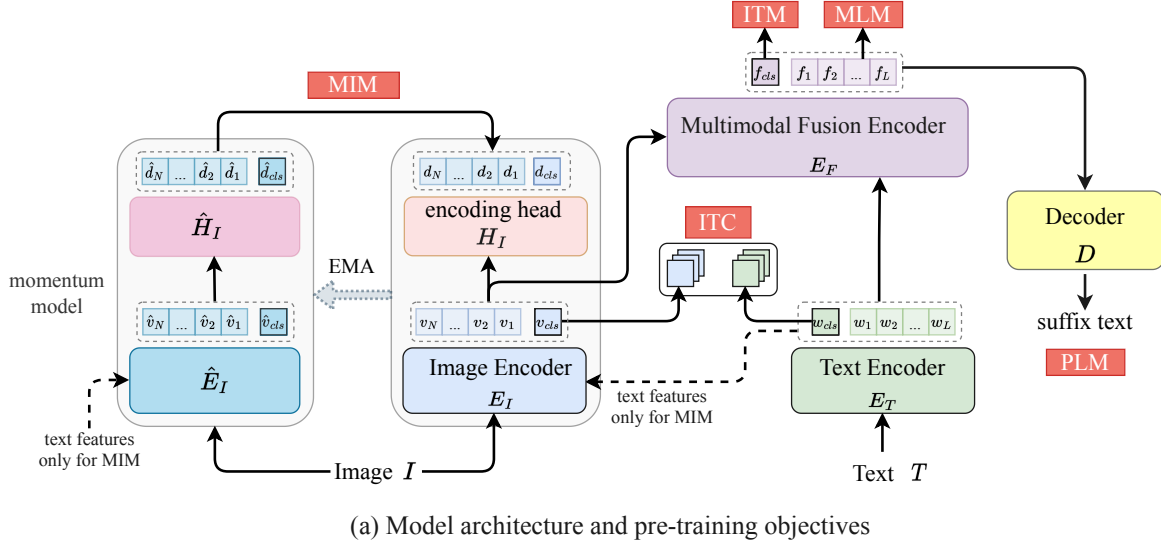(b) Local semantics enhancing and text-deeply-involved cross-modal MIM

Figure 1: Overview of our method. (a) shows the model architecture and pre-training objectives of our SemMIM framework. (b) illustrates the proposed local semantics enhancing approach, which harvests high-level semantics from global visual features via agreement learning, and transfer them into patch encodings by sharing the same encoding space. And (b) also shows our designs for deep involvement of text during MIM, including text-guided masking strategy and injecting textual information into both masked modeling and reconstruction target acquisition.

as illustrated in Figure 1(b), given an input image I, we first generate two different distorted views of it I_1 and I_2 through random augmentation. The two views are then fed into the momentum and original image encoders respectively, and get the corresponding encodings of their [CLS] tokens d_1^{cls} and $\hat{d}_2^{cls}$. Inspired by DINO (Caron et al., 2021), we learn the agreement between view 1 and view 2 as follows:

$$\mathcal{L}_{cls} = \mathbb{E}(\text{CE}(d_1^{cls}, \hat{d}_2^{cls})), \qquad (2)$$

where CE denotes cross entropy loss. Through this bootstrapping process on the [CLS] tokens, we acquire high-level semantics from the global image features and shape the semantic structure of the encoding space. In the meantime, we use the same momentum image encoder and encoding head to transform local patch features into visual

encodings, *i.e.*, the [CLS] tokens and the patch tokens share the same encoding space. Therefore, the learned high-level semantics can be naturally transferred to the local patch encodings. In this way, we achieve local semantics enhancing and can provide more semantically meaningful reconstruction targets for the MIM task.

3.2.2. Text-deeply-involved Cross-modal MIM

By enhancing local semantics, we provide more semantically meaningful reconstruction targets for MIM. However, lacking text involvement would lead MIM to degrade into learning purely visual information, which hinders facilitating cross-modal semantic alignment. Therefore, we further make three elaborate designs within our framework for deep involvement of text during the entire MIM process.

14667

Firstly, we propose an text-guided masking strategy. Instead of masking a proportion of image patches randomly, we leverage semantic similarities between image patches and text for masked patch selection, so as to encourage the utilization of textual information during MIM. Specifically, we first compute the similarity S_i between each local patch feature and the global feature of the paired text. Then we normalize the similarity scores of all patches using softmax, and take them as sampling probabilities to select a predefined proportion of patches. In this way, image regions with higher relevance to text are more likely to be selected, thereby encouraging the model to recover masked regions by reasoning textual information.

Secondly, we devise an efficient way of fusing textual information while masked modeling, which enables the model to recover masked regions by reasoning both contextual visual information and paired textual information. In cross-modal MLM, an extra fusion encoder is often required. Instead, we concatenate the output of the text encoder with visual tokens at the last N layers of the image encoder. Thus the masked tokens can interact with the paired text features via self-attention. We also investigate the effect of the start visual layer used to inject textual information for MIM. Please refer to the ablation study in Section 4.4.

Thirdly, we also fuse text features when using the momentum encoder to transform image patches into encodings, which injects high-level textual semantics into patch encodings. Note that the bootstrapping on [CLS] tokens in Section 3.2.1 would fail if textual information is leaked into [CLS] tokens. Therefore, we infer twice at the last N layers of the momentum encoder $\hat{E}_I$. One for regular visual modeling to extract [CLS] tokens, the other for injecting textual information into patch encodings.

In Equation 2, we feed the two views I_1 and I_2 of the input image into the momentum and original image encoders respectively. Here, to perform cross-modal MIM, we first apply the proposed masking strategy to replace the selected image patches with a learnable [MASK] token. Then we input the corrupted image I_2' into the image encoder E_I to recover the information of the masked regions. We denote $d_{2,i}^{patch}$ as the categorical distribution of the i-th masked patch of I_2', which is predicted by the encoding head H_I. And $\hat{d}_{2,i}^{patch}$ for that generated by the momentum encoding head $\hat{H}_I$. The loss of the MIM task can be formulated as follows:

$$\mathcal{L}_{patch} = \mathbb{E}(\frac{1}{M}\sum_{i=1}^{M}\mathrm{CE}(d_{2,i}^{patch}, \hat{d}_{2,i}^{patch})), \quad (3)$$

where CE denotes cross-entropy loss. M is the number of masked patches.

3.2.3. Pre-training Objectives

In addition to the cross-modal MIM task, we also jointly train the pre-training tasks below to learn cross-modal alignment following existing methods (Li et al., 2021, 2022a; Xu et al., 2023):
Image-Text Contrastive Learning (ITC, $\mathcal{L}_{\mathrm{itc}}$) aims to learn better uni-modal representations before fusion. **Image-Text Matching** (ITM, $\mathcal{L}_{\mathrm{itm}}$) uses a binary classification task to predict whether a pair of image and text is matched with each other. **Masked Language Modeling** (MLM, $\mathcal{L}_{\mathrm{mlm}}$) aims to predict the randomly masked words based on cross-modal representations. **Prefix Language Modeling** (PLM, $\mathcal{L}_{\mathrm{plm}}$) aims to equip the model with generation capability. Given an image and the prefix text T^p, we use the decoder to predict the suffix text T^s. Following previous methods (Li et al., 2021, 2022a), we assign equal loss weights to all these pre-training objectives, and thus the full pre-training objective is:

$$\mathcal{L} = \mathcal{L}_{\mathrm{cls}} + \mathcal{L}_{\mathrm{patch}} + \mathcal{L}_{\mathrm{itc}} + \mathcal{L}_{\mathrm{itm}} + \mathcal{L}_{\mathrm{mlm}} + \mathcal{L}_{\mathrm{plm}} \quad (4)$$

4. Experiments

4.1. Pre-training Datasets

Following ALBEF (Li et al., 2021), we use a hybrid pre-training dataset with 4 million images and 5.1 million captions, including two in-domain datasets (MS COCO (Lin et al., 2014), Visual Genome (Krishna et al., 2017)), and two web-crawled datasets (Conceptual Captions (Sharma et al., 2018), SBU Captions (Ordonez et al., 2011)).

4.2. Implementation Detail

We adopt ViT-B/16 as the visual encoder, and initialize it with CLIP-ViT (Radford et al., 2021) which is pre-trained on 400 million noisy image-text pairs. Besides, we use a 6-layer Transformer architecture for both the text encoder and the fusion encoder, and a 12-layer Transformer architecture for the decoder. The text encoder and the fusion encoder are initialized using the first 6 layers and the last 6 layers of the BERT_base model respectively. For pre-training, we use the AdamW (Loshchilov and Hutter, 2018) optimizer with a weight decay of 0.02. We use different learning rates for the visual encoder and the rest parts of our model, which are warmed-up to 1e-5 (ViT-B/16) and 1e-4 (BERT_base) in the first 1000 iterations, and decay to 1e-6 following a cosine schedule. We pre-train our model for 20 epochs on 8 NVIDIA A100 GPUs, with a total batch size of 1024. The image resolution is 224×224 for pre-training and is increased to 384×384 for fine-tuning. For the encoding head,

Methods	# Pretrain data	MSCOCO (5K test set)						Flickr30K (1K test set)					
		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
UNITER (Chen et al., 2020)	4M	65.7	88.6	93.8	52.9	79.9	88.0	87.3	98.0	99.2	75.6	94.1	96.8
OSCAR (Li et al., 2020)	4M	70.0	91.1	95.5	54.0	80.8	88.5	-	-	-	-	-	-
ALBEF (Li et al., 2021)	4M	73.1	91.4	96.0	56.8	81.5	89.2	94.3	99.4	99.8	82.8	96.7	98.4
VinVL (Zhang et al., 2021)	4M	74.6	92.6	96.3	58.1	83.2	90.1	-	-	-	-	-	-
ALIGN (Jia et al., 2021)	1.8B	77.0	93.5	96.9	59.9	83.3	89.8	95.3	99.8	100.0	84.9	97.4	98.6
VLMAE (He et al., 2022b)	4M	77.3	93.6	97.4	59.6	83.6	90.3	95.2	99.6	99.9	83.6	96.6	98.5
SCL (Ji et al., 2023a)	4M	77.7	94.1	97.4	60.1	84.5	91.4	95.9	99.8	100.0	84.5	97.4	98.9
MAP (Ji et al., 2023b)	4M	79.3	94.8	97.6	60.9	86.2	93.1	94.9	99.5	99.8	83.8	97.2	98.7
VL-BEiT (Bao et al., 2022)	4M	79.5	-	-	61.5	-	-	95.8	-	-	83.9	-	-
BLIP (Li et al., 2022b)	14M	80.6	95.2	97.6	63.1	85.3	91.1	96.6	99.8	100.0	87.2	97.5	98.8
X-VLM (Zeng et al., 2022)	4M	80.4	95.5	98.2	63.1	85.7	91.6	96.8	99.8	100.0	86.1	97.4	98.7
mPLUG (Li et al., 2022a)	4M	80.5	95.4	97.9	63.3	85.3	91.2	96.7	99.8	100.0	86.5	97.5	98.8
Ours	4M	**81.5**	**96.2**	**98.3**	**64.2**	**86.1**	**91.6**	**97.0**	**99.8**	**100.0**	**86.9**	**97.8**	**99.0**

Table 1: Evaluation results of image-text retrieval on MSCOCO and Flickr30K datasets.

Methods	# Pretrain Data	COCO Caption								NoCaps zero-shot	
		Cross-entropy Optimization				CIDEr Optimization					
		B@4	M	C	S	B@4	M	C	S	C	S
OSCAR (Li et al., 2020)	6.5M	-	-	-	-	41.7	30.6	140.0	24.5	83.4	11.4
VinVL (Zhang et al., 2021)	5.65M	38.5	30.4	130.8	23.4	41.0	31.1	140.9	25.2	97.3	13.8
BLIP (Li et al., 2022b)	14M	38.6	-	129.7	-	-	-	-	-	105.1	14.4
SimVLM$_{base}$ (Wang et al., 2021)	1.8B	39.0	32.9	134.8	24.0	-	-	-	-	-	-
mPLUG (Li et al., 2022a)	4M	39.3	30.1	132.4	23.3	41.2	30.8	140.2	25.2	98.3	12.9
Ours	4M	39.8	30.9	133.5	23.9	41.7	31.3	141.5	25.4	98.8	13.5

Table 2: Evaluation results on COCO Caption "Karpathy" test split and NoCaps validation set. B@4: BLEU@4, M: METEOR, C: CIDEr, S: SPICE.

Methods	# Pretrain Data	VQA	
		Test-dev	Test-std
E2E-VLP (Xu et al., 2021)	4M	73.25	73.67
OSCAR (Li et al., 2020)	6.5M	73.16	73.44
ALBEF (Li et al., 2021)	4M	74.54	74.70
VLMAE (He et al., 2022b)	4M	75.3	75.4
PTP-BLIP (Wang et al., 2023a)	4M	75.5	75.9
ALBEF (Li et al., 2021)	14M	75.84	76.04
VinVL (Zhang et al., 2021)	5.65M	76.52	76.60
VL-BEiT (Bao et al., 2022)	4M	77.53	77.75
BLIP (Li et al., 2022b)	14M	77.54	77.62
METER (Dou et al., 2022)	4M	77.68	77.64
mPLUG (Li et al., 2022a)	4M	77.58	77.73
FLM$_{large}$ (Wang et al., 2023b)	4M	77.80	77.84
SimVLM$_{base}$ (Wang et al., 2021)	1.8B	77.87	78.14
Ours	4M	**78.12**	**78.18**

Table 3: Evaluation results on VQA.

we set the dimension of the categorical distribution (*i.e.* the K value) to 1024.

4.3. Evaluation on Downstream Vision-Language Tasks

Image-Text Retrieval. We conduct experiments on MSCOCO and Flickr30K (Young et al., 2014) datasets and adopt a two-stage retrieval strategy following ALBEF (Li et al., 2021), which takes a fast retrieval via uni-modal encoders and then reranks via the fusion encoder. As shown in Table 1, our model outperforms previous methods on both datasets. Note that VLMAE (He et al., 2022b) and VL-BEiT (Bao et al., 2022) are also MIM-based methods. VL-BEiT adopts an offline tokenizer (*i.e.* dVAE) to provide reconstruction targets for MIM, while VLMAE directly recovers the raw pixels of the masked regions. We outperform these two methods by a large margin on both MSCOCO and Flickr30K. This shows that, by injecting high-level semantics into local patch encodings via self-supervised learning and shared encoding space, our model can provide more semantically meaningful supervision for MIM and improves the effectiveness of MIM for cross-modal alignment. Besides, our model also surpasses BLIP (Li et al., 2022b), which uses a significantly bigger dataset of 14 million image-text pairs for pre-training.

Image Captioning. We evaluate the image caption generation ability of our model on COCO Caption (Lin et al., 2014) and NoCaps (Agrawal et al., 2019) datasets. We fine-tune our model on COCO Caption, and test on the "Karpathy" test split and NoCaps validation set. Following OSCAR (Li et al., 2020), we first fine-tune our model with cross-entropy loss, and then with CIDEr optimization (Rennie et al., 2017) for 5 extra epochs. As shown in Table 2, our method surpasses mPLUG (Li et al., 2022a) which uses the same amount of pre-training data. On COCO Caption, our method even outperforms some methods which use much more data for pre-training. For instance, we sur-

Supervision for MIM	MSCOCO		VQA
	TR	IR	Test-dev
w/o MIM	79.0	62.3	77.20
raw pixels	78.6	62.2	77.08
dVAE	79.8	62.7	77.54
Ours	**81.5**	**64.2**	**78.12**

Table 4: Performance comparison of different supervision for MIM. Retrieval results are reported in terms of the R@1 of TR and IR.

Masking Strategy	MSCOCO		VQA
	TR	IR	Test-dev
random	81.0	63.6	77.96
text-guided	**81.5**	**64.2**	**78.12**

Table 5: Ablation on the masking strategy. R@1 retrieval results are reported.

Mask Ratio	MSCOCO		VQA
	TR	IR	Test-dev
15%	80.8	63.6	77.65
30%	**81.5**	**64.2**	**78.12**
45%	80.6	63.4	77.71

Table 6: Effect of the mask ratio. Retrieval results are reported in terms of the R@1 of TR and IR.

i_{th}	MSCOCO		VQA
	TR	IR	Test-dev
-	80.4	63.2	77.63
1	80.0	62.9	77.35
6	80.9	63.6	77.90
9	81.5	63.9	78.06
12	**81.5**	**64.2**	**78.12**

Table 7: The effect of the start visual layer used to inject textual information for MIM. Results are reported in terms of the R@1 of TR and IR.

pass BLIP which uses 14M image-text pairs by a large margin of 3.8 on the CIDEr metric. Moreover, on the NoCaps dataset which reflects the zero-shot caption generation ability, we also outperform previous methods (Li et al., 2020; Zhang et al., 2021; Li et al., 2022a) which use similar amounts of pre-training data, showing the good generalization ability of our method.

Visual Question Answering. We treat VQA as an answer-generation task and directly apply unconstrained open-vocabulary generation during inference. Following OFA (Wang et al., 2022a), we concatenate the question with the object labels and OCR tokens extracted from images. As Table 3 shows, our model outperforms previous methods using similar amounts of pre-training data. And we even surpass $SimVLM_{base}$ which uses $400\times$ more image-text data for pre-training. We largely outperform VLMAE by 2.82% on test-dev and 2.78% on test-std, and surpass VL-BEiT by 0.6% on the test-dev split. The comparison validates that, with the proposed local semantics enhancing and text-deeply-involved cross-modal MIM, our SemMIM framework equips the model with more powerful semantic reasoning ability.

4.4. Ablation Study

The effect of local semantics enhancing. As shown in Table 4, we compare different supervision for the MIM task. Compared to the baseline model which doesn't perform MIM, in our experiment, using raw pixels as the reconstruction targets for MIM in fact brings a slight performance drop. We analyze this is due to raw pixels don't have any semantic abstraction, and thus taking them as MIM's supervision can't achieve effective cross-modal se-

mantic interaction. Using a pre-trained dVAE to provide supervision improves performance. However, the boost is relatively limited, as the dVAE encodings also lack high-level semantics. In contrast, our method significantly surpasses raw-pixels-based and dVAE-based MIM, showing that our proposed local semantics enhancing approach can largely improve the effectiveness of MIM.

The effect of the text-guided masking strategy. We compare the performance of random masking and the proposed text-guided masking. As Table 5 shows, applying text-guided masking brings significant performance boost. As image regions with higher relevance to text are more likely to be selected, text-guided masking encourages the model to recover masked regions by reasoning textual information, and thus further facilitates cross-modal semantic interaction and fusion.

The effect of the mask ratio in MIM. In Table 6, we experiment with different mask ratios for MIM. The best performance is obtained at 30% mask ratio. We analyze that a smaller mask ratio makes some masked patches can be recovered by surrounding patches without utilizing text information, impairing fine-grained cross-modal alignment. In contrast, too high mask ratio may cause severe information loss, leading to a performance drop.

The effect of the start visual layer used to inject textual information for MIM. Table 7 shows that fusing textual information starting from different layers of the image encoder has varying effects. The first line is MIM without using textual information. As can be seen, injecting text features into shallow layers causes performance drop. We analyze this
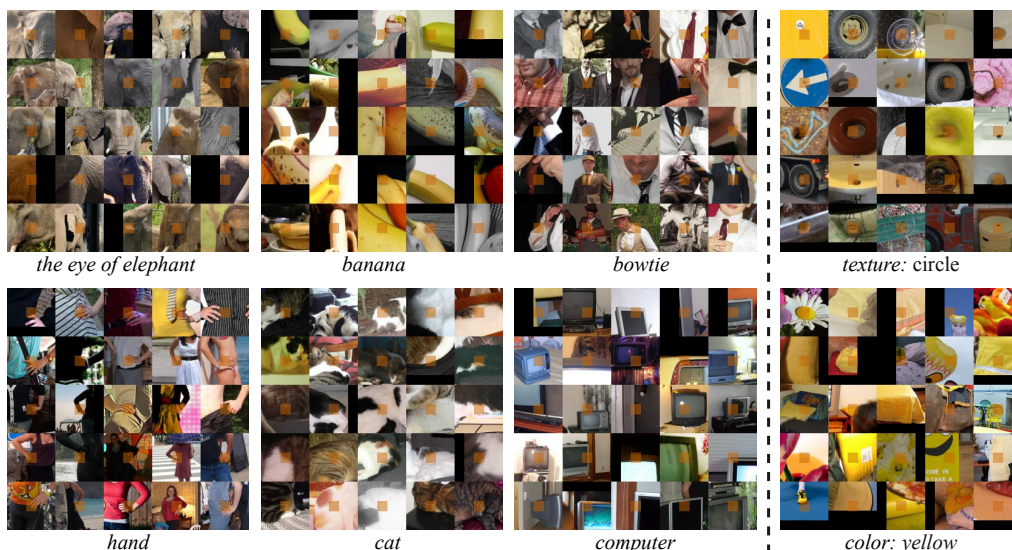
Figure 2: Pattern clusters of image patch encodings. The left six figures (our model) showcase high-level semantic patterns *the eye of elephant*, *banana*, *bowtie*, *hand*, *cat* and *computer*. The right two figures (dVAE) showcase relatively low-level visual patterns *circle texture* and *yellow color*.
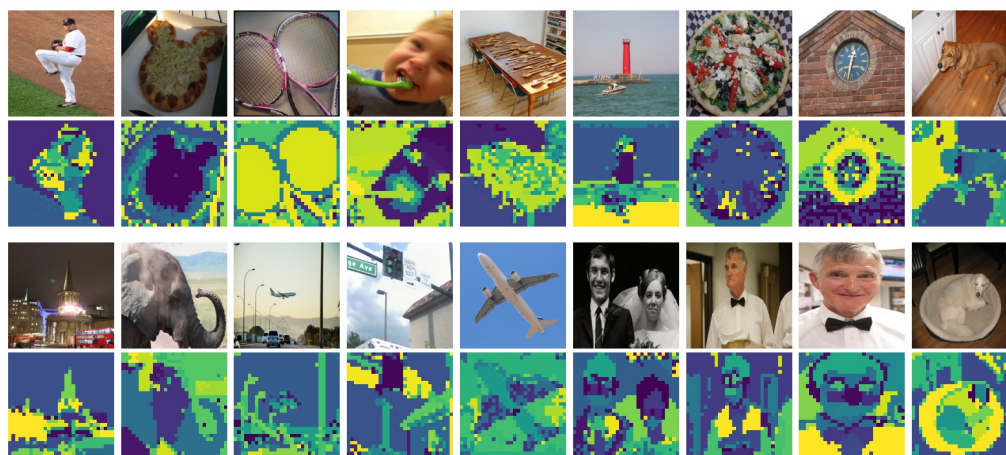

Figure 3: Visualization of pattern layout of full images. Different patterns are shown in different colors.

is due to high-level textual semantics interfere the modeling of low-level visual information. Starting from deeper layers brings significant performance improvement. This shows that the injected textual semantics can help high-level visual features to better recover masked regions, thereby facilitating cross-modal semantic alignment.

4.5. Visualization

Pattern Cluster. As shown in Figure 2, we visualize the pattern clusters of the image patch encodings obtained by our model and dVAE (Ramesh et al., 2021). For case(a), we can find that the patches are grouped with certain semantic meanings, *e.g., hand* despite their complicated background. This demonstrates that our proposed local semantics enhancing approach can effectively inject high-level semantics into local patch encodings, thus providing more semantically meaningful supervision for cross-modal MIM. In contrast, as case (b) shows, the patches grouped by dVAE share similar textures or colors but have little in common in semantics. This is because dVAE is trained through image reconstruction which simply minimizes pixel-level differences between original and reconstructed images. Therefore, it concentrates more on relatively low-level visual information rather than high-level semantics.

Pattern Layout of Full Images. In Figure 3, we visualize all the patterns of the image patch encodings in full images, where different patterns are distinguished by different colors. From the outline of the patterns, we can find that the extracted patterns are discriminative. This shows that, benefiting from the proposed local seman-

Methods	#Param	Pre-train Time (GPU Days))	MSCOCO TR@1	IR@1
ALBEF (Li et al., 2021)	210M	28 (A100)	73.1	56.8
FLM$_{large}$ (Wang et al., 2023b)	-	57 (V100)	73.5	56.6
VinVL (Zhang et al., 2021)	-	320 (V100)	74.6	58.1
METER (Dou et al., 2022)	-	64 (A100)	76.2	57.1
BLIP (Li et al., 2022b)	446M	112 (A100)	80.6	63.1
mPLUG (Li et al., 2022a)	354M	40 (A100)	80.5	63.3
Ours	360M	32 (A100)	**81.5**	**64.2**

Table 8: Comparison of the number of trainable parameters and the pre-training time.

tics enhancing approach, our model can effectively obtain high-level semantic information in image patches. Therefore, our SemMIM framework can provide MIM with more semantically meaningful reconstruction targets, thereby more effectively promoting cross-modal semantic alignment.

4.6. Efficiency

The number of parameters. Table 8 shows the comparison of the number of trainable parameters and the pre-training time with some existing methods. ALBEF (Li et al., 2021) has fewer parameters than ours, mainly because it hasn't decoder and thus hasn't caption generation ability. Compared to mPLUG (Li et al., 2022a), with a mere 1.7% increase in the number of parameters, our method achieves significantly better performance. The reason why our method doesn't substantially increase the number of parameters is that, the parameters of the momentum model are acquired via Exponential Moving Average (EMA), rather than trainable parameters. Therefore, our method in fact only adds the parameters of the encoding head, an MLP module with less than 6M parameters. Besides, our masking strategy and injecting textual information do not add any learnable parameters.

Training efficiency. As Table 8 shows, compared to existing methods, our method requires almost the shortest pre-training time (except for AL-BEF), while achieving state-of-the-art performance across various vision-language tasks. On one hand, compared to the baseline, most operations we add are computationally efficient. For example, the momentum model only needs a forward pass to extract image features without requiring back-propagation. Injecting textual information only increases the sequence length slightly, as the text length is much shorter than the number of image patches. On the other hand, our SemMIM framework effectively promotes cross-modal semantic alignment, and thus accelerates the convergence of training, requiring fewer training epochs.

Inference efficiency. It's also worth noting that, although we add the cross-modal MIM task during pre-training, this only affects the pre-training time and has no impact on the inference speed. After training, our model maintains similar inference efficiency to other models with comparable size and architecture such as mPLUG (Li et al., 2022a).

5. Conclusion

In this work, we have proposed a semantics-enhanced cross-modal MIM framework (SemMIM) for fine-grained vision-language semantic alignment. Firstly, we propose to inject high-level semantics into local patch encodings via self-supervised agreement learning and sharing encoding space, which can provide more semantically meaningful supervision for MIM. And we propose a text-guided masking strategy and devise an efficient way of injecting textual information in both masked modeling and reconstruction target acquisition, which achieves deep involvement of text during the entire MIM process. Extensive experiments on various vision-language tasks and the visualization results verify the effectiveness of the proposed SemMIM method.

6. Acknowledgements

7. Bibliographical References

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8948–8957.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. 2021. Beit: Bert pre-training of image transformers. In *International Conference on Learning Representations*.

Hangbo Bao, Wenhui Wang, Li Dong, and Furu Wei. 2022. Vl-beit: Generative vision-language pretraining. *arXiv preprint arXiv:2206.01127*.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-

supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660.

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

Zhuo Chen, Yufeng Huang, Jiaoyan Chen, Yuxia Geng, Wen Zhang, Yin Fang, Jeff Z Pan, and Huajun Chen. 2023. Duet: Cross-modal semantic grounding for contrastive zero-shot learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 405–413.

Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, Nenghai Yu, and Baining Guo. 2023a. Peco: Perceptual codebook for bert pre-training of vision transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 552–560.

Xiaoyi Dong, Jianmin Bao, Yinglin Zheng, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, et al. 2023b. Maskclip: Masked self-distillation advances contrastive language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10995–11005.

Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, et al. 2022. An empirical study of training end-to-end vision-and-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18166–18176.

Yuying Ge, Yixiao Ge, Xihui Liu, Jinpeng Wang, Jianping Wu, Ying Shan, Xiaohu Qie, and Ping Luo. 2022. Miles: Visual bert pre-training with injected language semantics for video-text retrieval. In *European Conference on Computer Vision*, pages 691–708. Springer.

Xinyang Geng, Hao Liu, Lisa Lee, Dale Schuurams, Sergey Levine, and P. Abbeel. 2022. Multimodal masked autoencoders learn transferable representations. *ArXiv*, abs/2205.14204.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022a. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009.

Sunan He, Taian Guo, Tao Dai, Ruizhi Qiao, Chen Wu, Xiujun Shu, and Bo Ren. 2022b. Vlmae: Vision-language masked autoencoder. *arXiv preprint arXiv:2208.09374*.

Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. 2020. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *ArXiv*, abs/2004.00849.

Yatai Ji, Rongcheng Tu, Jie Jiang, Weijie Kong, Chengfei Cai, Wenzhe Zhao, Hongfa Wang, Yujiu Yang, and Wei Liu. 2023a. Seeing what you miss: Vision-language pre-training with semantic completion learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6789–6798.

Yatai Ji, Junjie Wang, Yuan Gong, Lin Zhang, Yanru Zhu, Hongfa Wang, Jiaxing Zhang, Tetsuya Sakai, and Yujiu Yang. 2023b. Map: Multimodal uncertainty-aware vision-language pre-training model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23262–23271.

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

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

Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, He Chen, Guohai Xu, Zheng Cao, et al. 2022a. mplug: Effective and efficient vision-language learning by cross-modal skip-connections. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7241–7259.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022b. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision

and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.

Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *ECCV*.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*.

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

Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. 2022. Beit v2: Masked image modeling with vector-quantized visual tokenizers. *arXiv preprint arXiv:2208.06366*.

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

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8821–8831. PMLR.

Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1179–1195.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. 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.

Yaya Shi, Haowei Liu, Haiyang Xu, Zongyang Ma, Qinghao Ye, Anwen Hu, Ming Yan, Ji Zhang, Fei Huang, Chunfeng Yuan, et al. 2023. Learning semantics-grounded vocabulary representation for video-text retrieval. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4460–4470.

Hao Tan and Mohit Bansal. 2019. 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)*. Association for Computational Linguistics.

Jinpeng Wang, Pan Zhou, Mike Zheng Shou, and Shuicheng Yan. 2023a. Position-guided text prompt for vision-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23242–23251.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022a. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR.

Teng Wang, Yixiao Ge, Feng Zheng, Ran Cheng, Ying Shan, Xiaohu Qie, and Ping Luo. 2023b. Accelerating vision-language pretraining with free language modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23161–23170.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. 2022b. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *ArXiv*, abs/2208.10442.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2021. Simvlm: Simple visual language model pretraining with weak supervision. In *International Conference on Learning Representations*.

Siying Wu, Xueyang Fu, Feng Wu, and Zheng-Jun Zha. 2022. Cross-modal semantic alignment pre-training for vision-and-language navigation.

In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4233–4241.

Haiyang Xu, Ming Yan, Chenliang Li, Bin Bi, Songfang Huang, Wenming Xiao, and Fei Huang. 2021. E2e-vlp: End-to-end vision-language pretraining enhanced by visual learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 503–513.

Haiyang Xu, Qinghao Ye, Ming Yan, Yaya Shi, Jiabo Ye, Yuanhong Xu, Chenliang Li, Bin Bi, Qi Qian, Wei Wang, et al. 2023. mplug-2: A modularized multi-modal foundation model across text, image and video. *arXiv preprint arXiv:2302.00402*.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Yan Zeng, Xinsong Zhang, and Hang Li. 2022. Multi-grained vision language pre-training: Aligning texts with visual concepts. In *International Conference on Machine Learning*, pages 25994–26009. PMLR.

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

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. 2020. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13041–13049.