Efficient Vision-Language Pretraining with Visual Concepts and Hierarchical Alignment

Mustafa Shukor[1]
mustafa.shukor@sorbonne-universite.fr
Guillaume Couairon[1,2]
gcouairon@fb.com
Matthieu Cord[1,3]
matthieu.cord@sorbonne-universite.fr

[1] Sorbonne University, Paris, France
[2] Meta AI
[3] Valeo.ai, Paris, France

Abstract

Vision and Language Pretraining has become the prevalent approach for tackling multimodal downstream tasks. The current trend is to move towards ever larger models and pretraining datasets. This computational headlong rush does not seem reasonable in the long term to move toward sustainable solutions, and *de facto* excludes academic laboratories with limited resources. In this work, we propose a new framework, dubbed ViCHA, that efficiently exploits the input data to boost the learning by: (a) a new hierarchical cross-modal alignment loss, (b) new self-supervised scheme based on masked image modeling, (c) leveraging image-level annotations, called *Visual Concepts*, obtained with existing foundation models such as CLIP to boost the performance of the image encoder. Although pretrained on four times less data, our ViCHA strategy outperforms other approaches on several downstream tasks such as Image-Text Retrieval, VQA, Visual Reasoning, Visual Entailment and Visual Grounding. The code is available here: https://github.com/mshukor/ViCHA.

1 Introduction

Vision and Language Pretraining (VLP) [11, 21] consists of training a vision and language model with simple pretraining tasks, like image-text alignment or masked language modelling, usually on large datasets. It is becoming the prominent paradigm to solve multimodal vision-language tasks (*e.g.* VQA [4], NLVR^2 [68]), outperforming other task-customised approaches [7, 31]. The representation learned by these models have shown to be useful also for unimodal tasks, paving the way for more general or foundational models [55, 65, 86, 87].

Due to the abundance of image-text pairs data on the internet [10, 17, 60], scaling these models has gained a lot of attention. The current most efficient approaches are those beyond 1B parameters [1, 86] and trained on hundred of millions to several billions examples [30, 55, 77]. This race for bigger models/data, is most often done to the detriment of the quality of learning strategies, which becomes more and more difficult to control. We argue that a lot of improvements can be done by designing more efficient learning schemes. For example, while Dosovitskiy et al. [19] train Vision Transformers with a huge pre-training

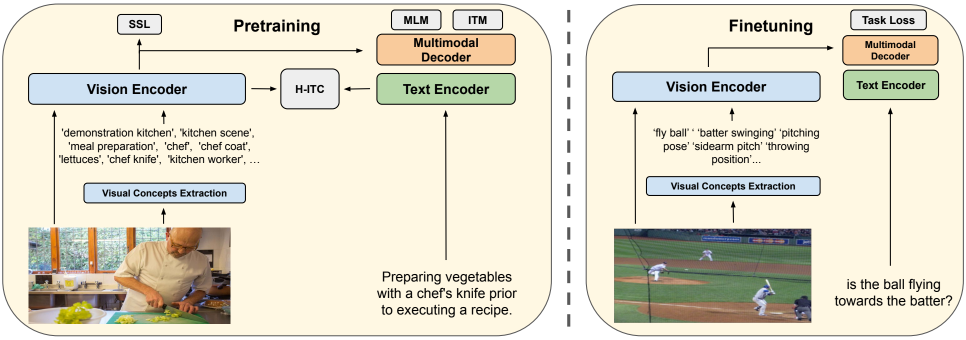

Figure 1: Illustration of our ViCHA approach; during pretraining (left) and finetuning (right).

dataset, Touvron et al. [71] obtained state-of-the art results using the same architecture with bag of learning recipes, without pre-training.

To this end, we propose a new VLP strategy with carefully designed training procedures and losses dedicated to efficiently align both modalities. Specifically, we adopt an early fusion architecture, but instead of having only one alignment loss on top of the dual encoders, we also align both representations at several layers. This hierarchical alignment strategy is extended with a multimodal fusion decoder in order to capture more complex Vision-Language interaction. We also exploit more efficiently the input data with Masked Image Modeling, based on the recently proposed Masked Auto Encoder (MAE) [26]. Moreover, we leverage image-level annotations (*i.e.*, objects and high level concepts describing the image) using existing foundation models (*i.e*, CLIP [55]) due their generalization ability and low computational cost at inference time. Thus departing from classical use of such models (*i.e.*, initialization and finetuning), and leveraging them without any retraining. We complete our scheme with a CLIP-based filtering technique. An illustration of our approach, called ViCHA for Efficient Vision-Language Pretraining with **Vi**sual **C**oncepts and **H**ierarchical **A**lignment, is presented in Fig. 1. We will show the effectiveness of our approach on classical downstream tasks used for VLP evaluation while drastically limiting the size of the training datasets, contrary to other methods.

2 Related Work

Vision and Language Pretraining (VLP): VLP methods can be categorized into 3 main categories based on their model architectures. First, dual stream approaches have two separate encoders for images and texts (e.g. CLIP [55], ALIGN [30]) that are usually trained using a contrastive loss [51] on a global similarity between their output embeddings. These models are efficient during inference, but can not exploit finegrained cross modal interaction that is useful for multimodal tasks (*e.g.*, VQA). Several attempts based on self supervision [41, 50], teacher-student distillation [3, 75] or finegrained interaction [82] have been proposed to alleviate the massive amount of data usually used for training (*e.g.*, CLIP 400M). Second, unified approaches try to simplify the model and adopt one transformer that can process the two modalities [73, 74]. These approaches are simpler, but still under perform other approaches with modality specific modules. Third, most of the work concentrated recently on hybrid approaches due to their success on multimodal tasks, where the models have separate encoders as well as a multimodal module that can exploit the cross-modal interaction more effectively. Early methods have relied on object detection models to extract image regions and tags [12, 35, 37, 48, 66, 89], however, the visual representation is limited by the performance of the pretrained object detector, which is expensive to scale on large annotated datasets, besides being slow at inference time. To remedy this, many methods have proposed to replace the object detector with vision transformers [20, 32, 80] or CNNs [29, 77]. These models are usually trained with image-text matching (ITM [12]) or masked language modeling (MLM [18]) losses, and recently with image text contrastive (ITC) to align the vision and language representation before feeding them to the fusion module [22, 36, 81].

Multilevel Alignment: Using multiscale representation extracted from different stages of the model have shown to be successful for vision tasks, in particular, some work in representation learning have proposed to do contrastive learning [53] or mutual information maximization [5] between features extracted at different scales, or between local and global features [28]. In the context of VLP, Li et al. [42] propose a multilevel semantic alignment by aligning local and global representations. Gao et al. [24] propose an intra and cross level alignment based on features extracted from cropped images and image regions on one side, and the caption and its summarization on the other side. Li et al. [34] use both single and dual stream encoders to align the two modalities at multiple levels. Our approach use simpler strategy to align the features of both modalities at different transformer blocks.

Object tags: Several works have explicitly used high level concepts for vision-language tasks (*e.g.* for image captioning and VQA [79, 84]). In the context of VLP, the concepts are usually extracted from an off-the-shelf object detectors [58] which are fed to the multimodal transformer, alongside region features and text tokens [91]. Other than using object detectors, some approaches extract the concepts from the captions using Scene Graph Parser [2]. These concepts are used in several ways; predicting these concepts [47, 70, 85], using them as a bridge between non aligned image and text data [38], adopting a more efficient masking strategies [15] or leveraging them for fine-grained alignment [78]. In particular, OSCAR [40] uses the object tags to ease the alignment process by considering them as anchors for masked tokens prediction and contrastive learning. The concepts extracted from pretrained object detection models are limited to the training classes, in addition, scene graph parser are noisy and there is no guaranty that all the concepts in the caption are extracted. Our approach leverages a more general model to capture a large set of diverse concepts and using them to enhance the visual representation.

Leveraging Foundational Models: In recent years the focus have been shifted from task-customized models to more general or foundational models [55, 65, 87]. The motivation is to have a holistic model that can be later used or adapted to many downstream tasks. Due to their generalization capabilities, models such as CLIP [55], have been successfully leveraged in many tasks and domains, such as image editing [14], generation [57], segmentation [76], captioning [67], food retrieval [64], explainability [59] and beyond. In the context of VLP, these models are used as initialization [62] and recently, for dataset filtering [60]. In this work, we leveraged CLIP for the extraction of VCs and data selection/filtering.

3 ViCHA framework

Illustrated in Figure 2, our scheme presents Vision and Text encoders with three original components: a Visual Concepts Module to enrich the image encoder with relevant VCs, a new cross-modal interaction to align visual and langage feature representations at multiple levels, and a self-supervised component based on the recently proposed Masked Auto Encoder.

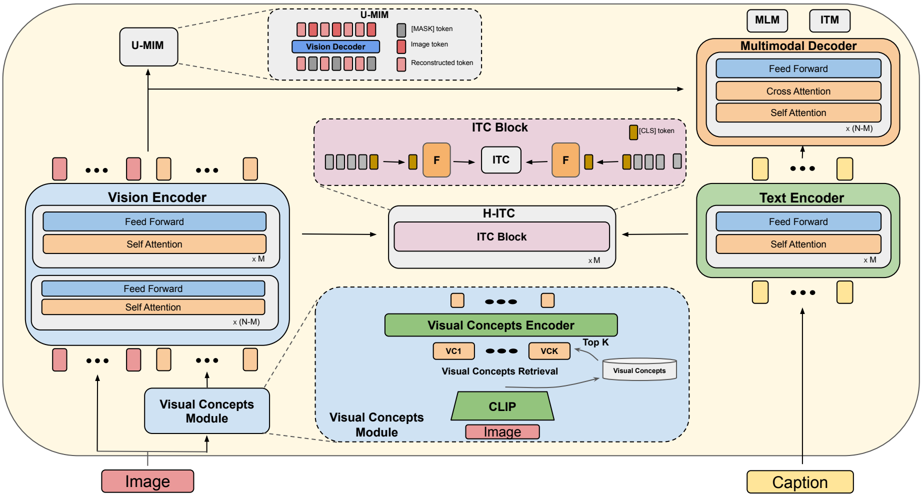

Figure 2: ViCHA framework: our Vision-Language transformer model consists of the vision encoder (left) and the text encoder (right), both aligned through our Hierarchical Image-Text Contrastive block (H-ITC). Importantly, the input visual tokens are completed through the Visual Concepts Module providing extra visual semantic tokens. Finally, a multimodal decoder (top right) allows to learn complex multimodal relationship thanks to MLM and ITM objectives. We also introduce a unimodal Masked Image Modeling (U-MIM) on top of the vision encoder.

Notations: Given a dataset of image-text pairs $\{I_i, T_i\}_i^K$, each image I_i is associated with several Visual Concepts (VCs) $\{C_i^1, ..., C_i^P\}$ extracted using CLIP and exploited by our vision encoder E_v. The VCs are projected to the image patches embedding space using a Visual Concept Encoder E_{vc}. E_v is a vision transformer that takes the image and the concepts as input and extracts the image tokens alongside a special class token $\{v_{cls}, v_1, ..., v_M\}$. Similarly, the text transformer encoder E_t takes the text or image caption T and extracts the text tokens $\{t_{cls}, t_1, ..., t_N\}$. The class tokens are aligned using a hierarchical contrastive loss before feeding the image and text tokens to a multimodal transformer decoder E_{vl} that takes the text tokens as query and the image ones as keys and values.

3.1 Enhancing Visual Representation with Visual Concepts

We propose to enhance the visual representation by explicitly injecting visual concepts. These concepts are extracted for each image, projected using a visual concepts encoder E_{vc} and then concatenated to the patch tokens before feeding them to the vision transformer E_v.

The motivation behind VCs is two-fold; VCs (a) might guide the vision encoder to focus on important objects/aspects of the image, (b) facilitate the alignment with the textual modality, first, because the visual tokens are already fused with the textual tokens of the VCs, second, it is easier to align the caption/text with the VCs tokens than the image tokens, especially at the beginning of the training.

Also we can see this technique from a visual prompt perspective. Text prompts are usually used with large language models to steer their output based on a given task/input [8, 44, 63, 83]. We derive the analogy between these prompts and the prepend of VCs, as they may also guide the model to focus on different aspects of the image during training.

Next, we detail how we extract, project and incorporate the VCs in our framework;

Visual Concepts Extraction: Different from other approaches (*e.g.*, OSCAR [40]) that extract these concepts (*i.e.*, object tags) using pretrained object detectors, we use an off-the-shelf foundational model to extract more diverse, larger and semantically meaningful set of visual concepts. We show in section 4, that this approach helps to capture high level and global concepts describing the scenes, which are hard to obtain with other approaches (*e.g.* object detectors). The extraction of the visual concepts is done as follows;

1. Concepts extraction: we extract the objects using Scene Graph Parsers [61] from all the captions of a given dataset. We use only simple text filters (transform to lower case and select objects repeated more than once). However, more complicated filters might improve the results even further.

2. Concepts and image embeddings: We use a pretrained CLIP (ViT-B/16) model to obtain the embeddings of all the images and the extracted concepts.

3. Concepts selection: for each image, we compute its cosine similarity with all the embedded concepts and select the top k similar concepts.

Visual Concepts Encoder (E_{vc}): To encode Visual Concepts, we use a small text encoder E_{vc} and then concatenate the output token embeddings with the image patch tokens before feeding E_v.
Visual Concepts Augmentation (VCA): Here we propose a simple yet effective VCs augmentation technique. Having a set of concepts extracted for each image, instead of considering all the concepts at once, we sample randomly a fraction of these concepts (*i.e.*, p_{vc}%) at each iteration step. The benefit of VCA is three-fold; (1) it may prevent the model from overfitting on specific concepts and potentially disregard the image or other concepts during training, (2) the model sees different combinations of concepts, which helps to have more diversity. (3) the model is exposed to more information during finetuning or test, as it will see all the concepts at once.

3.2 Training scheme

We pretrain the model using several objectives; Hierarchical Image-Text Contrastive Learning (H-ITC), ITM, MLM and Masked Image Modeling (MIM).
Hierarchical Image-Text Contrastive Learning (H-ITC) Here we propose to exploit the hierarhical representations capture by the vision and language encoders. CNN based vision encoders learn hierarchical representations, starting from local features to more abstract ones at later stages [88]. Vision transformers have also been shown to display some level of hierarchy in learned representations [19, 56].

On the text side, recent works show that the attention heads specialize in particular part-of-speech tags, which differs across heads and layers [72] and reflect different aspects of the syntax [13], while the concepts learned by BERT differ significantly across layers and evolve into representing a linguistic hierarchy [16].

Motivated by this, and different from other work [22, 36, 81] that align the two modalities only at the last layer, we propose to align them at different layers of the vision and text transformers. We argue that doing the alignment at early stages facilitates the process at the subsequent layers, while allowing to align the representation at different semantic levels.

There is an asymmetry between the visual and textual information, as the image contains more diverse and detailed information while the caption usually contains more abstract one, thus we align the textual features only with the visual ones from the last layers of E_v.

Method	# Pre-train Images	VQA test-dev	VQA test-std	NLVR^2 dev	NLVR^2 test-P	SNLI-VE val	SNLI-VE test
ViLBERT [48]	3M	70.55	70.92	-	-	-	-
12-in-1 [49]	3M	73.15	-	-	78.87	-	76.95
ERNIE-ViL [85]	3.8M	73.18	73.36	-	-	-	-
ImageBERT [54]	6M	-	-	-	-	-	-
Unicoder-VL [35]	3.8M	-	-	-	-	-	-
UNITER [12]	4M	72.70	72.91	77.18	77.85	78.59	78.28
OSCAR [40]	4M	73.16	73.44	78.07	78.36	-	-
VILLA [23]	4M	73.59	73.67	78.39	79.30	79.47	79.03
UNIMO$_B$[39]	4M+1.7M	73.79	74.02	-	-	80.00	79.10
ViLT [32]	4M	70.94	-	75.24	76.21	-	-
ALBEF [36]	4M	74.54	74.70	80.24	80.50	80.14	80.30
ALBEF*	1.1M	72.51	72.69	75.72	76.31	78.08	78.02
ViCHA	1.1M	**73.55**	**73.52**	**77.27**	**77.08**	**78.96**	**78.22**
ViCHA†	800K	73.23	-	78.14	77.00	79.02	78.65

Table 1: Comparison with SOTA; we report the accuracy on VQA, NLVR^2 and VE. ViCHA outperforms other approaches trained on much more data ($\sim$ 4M images) and SOTA trained on the same setup (ALBEF*). Our model trained on the filtered dataset (ViCHA†) is very competitive while using much less data ($\sim$ 800K images). Results on 4M images are in the supp.

Specifically, at layer i, we compute the image-to-text and text-to-image similarities and then apply a Softmax with a learned temperature parameter τ:

$$p^i_{i2t} = \frac{exp(s(I,T)^i/\tau)}{\sum_{m=1}^{Q} exp(s(I,T_m)^i/\tau)}, \quad p^i_{t2i} = \frac{exp(s(T,I)^i/\tau)}{\sum_{m=1}^{Q} exp(s(T,I_m)^i/\tau)}, \tag{1}$$

T_m and I_m are the negative text and image examples. $s(\cdot,\cdot)^i$ is the cosine similarity and is obtained after linearly projecting and normalizing the class tokens at layer i into a shared latent space:

$$s(I,T)^i = g_{vi}(v^i_{cls})^T g'_{ti}(t^i_{cls}), \quad s(T,I)^i = g_{ti}(t^i_{cls})^T g'_{vi}(v^i_{cls}), \tag{2}$$

where $g_{ti}()$ and $g_{vi}()$ are the linear projection layers at layer i. We maintain two queues of size Q that store the normalized features from the momentum models $g'_{li}()$ and $g'_{vi}()$. We didn't see a significant improvement when using queues for the other layers, thus we use the queues only for the last layer for simplicity (on select in-batch negative examples for other layers). We then compute the cross entropy loss between $\boldsymbol{p}$ and the one-hot ground truth and sum across all layers.

Image-Text Matching (ITM) ITM loss is applied on top of the multimodal trasformer decoder for more finegrained fusion. It is a binary cross entropy loss, where the model tries to classify if the image-text pairs are positive or negative. The image and its corresponding caption are considered as positive pairs while the other examples in the batch are considered as negative. We sample a hard text and hard image from the batch based on the global cosine similarity on top of the dual encoders [36].

Masked Language Modeling (MLM) MLM consists of predicting a masked token given other contextual tokens. In our work, the model has also access to the visual tokens, which helps to learn a cross modal representation. Following BERT [18], we mask 15% of the tokens and replace them with [MASK] token, random token or we keep them unchanged with probabilities of 80%, 10% and 10% respectively. The task is a classification loss where the model should predict the token id from a list of vocabulary.

Method	# Pre-train Images	Flickr30K (1K test set)						MSCOCO (5K 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
ViLBERT [48]	3M	-	-	-	58.2	84.9	91.5	-	-	-	-	-	-
l2-in-1 [49]	3M	-	-	-	67.90	-	-	-	-	-	68.00	-	-
ERNIE-ViL [85]	3.8M	86.7	97.80	99.00	74.44	92.72	95.54	-	-	-	-	-	-
ImageBERT [64]	6M	87.0	97.6	99.2	73.1	92.6	96.0	66.4	89.8	94.4	50.5	78.7	87.1
Unicoder-VL [35]	3.8M	-	-	-	-	-	-	62.3	87.1	92.8	46.7	76.0	85.3
UNITER [12]	4M	87.3	98.0	99.2	75.6	94.1	96.8	65.7	88.6	93.8	52.9	79.9	88.0
OSCAR [40]	4M	-	-	-	-	-	-	70.0	91.1	95.5	54.0	80.8	88.5
VILLA [23]	4M	87.9	97.5	98.8	76.3	94.2	96.8	-	-	-	-	-	-
UNIMO$_B$ [39]	4M+1.7M	89.7	98.4	99.1	74.7	93.4	96.1	-	-	-	-	-	-
ViLT [32]	4M	83.5	96.7	98.6	64.4	88.7	93.8	61.5	86.3	92.7	42.7	72.9	83.1
ALBEF [36]	4M	94.3	99.4	99.8	82.8	96.7	98.4	73.1	91.4	96.0	56.8	81.5	89.2
ALBEF*	1.1M	88.1	97.8	99.0	73.6	92.0	95.6	71.2	91.3	95.6	54.2	80.7	88.6
ViCHA	1.1M	**91.7**	**98.7**	**99.5**	**77.2**	**94.2**	**96.8**	**73.6**	**92.4**	**96.2**	**56.8**	**82.2**	**89.5**
ViCHA†	800K	90.0	98.4	99.8	77.4	94.3	96.7	73.3	92.1	96.2	55.8	81.8	89.1

Table 2: Comparison with SOTA; we report R@K for finetuning on Image-Text Retrieval. ViCHA outperforms other approaches trained on much more data ($\sim$ 4M images) and SOTA trained on the same setup (ALBEF*). Our model trained on the filtered dataset (ViCHA†) is very competitive while using much less data ($\sim$ 800K images). Results on 4M images are in the supp.

Masked Image Modeling (MIM): Improving the visual representation have shown to affect significantly the performance on vision-language tasks [89]. Many work in self supervised learning (SSL) have been proposed [9, 25], in particular, the methods based on masked image reconstruction have achieved SOTA results [6, 26]. However, it is not clear how much MIM is useful for VLP, as recent work show that MIM based on approaches such as BEiT, masked region regression or classification degrades the performance [20, 27, 32]. Motivated by these findings, here we propose two approaches to investigate whether MIM could help VLP. The two methods are based on the recent Masked Auto Encoder (MAE) approach [26], where we randomly mask the image (*e.g.*, 75%) and pass only the unmasked tokens to E_v, however, they are different in how they reconstruct the image:

Unimodal MIM (U-MIM): Here we propose to improve the visual representation by reconstructing the masked tokens given only the unmasked image ones (without access to VCs). Specifically, we add a small decoder D_v (*e.g.*, 2-layer transformer) that takes the output tokens, concatenated to the masked ones ($\hat{\boldsymbol{v}}$) and reconstruct the input image as $\hat{I}$. The unimodal masked image modeling (U-MIM) loss can be written as:

$$\mathcal{L}_{U-MIM} = MSE(I, D_v(\hat{\boldsymbol{v}})). \tag{3}$$

Multimodal MIM (M-MIM): Here the decoder has access also to the textual tokens, which helps to provide more informative context to reconstruct the masked tokens. The decoder is the same as our E_{vl} (shared weights) but is provided with the visual tokens ($\hat{\boldsymbol{v}}$) as queries and the textual ones ($\boldsymbol{t}$) as keys and values (hence, the queries, keys and values are different when computing MLM and ITM). The loss can be written as:

$$\mathcal{L}_{M-MIM} = MSE(I, F(E_{vl}(\hat{\boldsymbol{v}}, \boldsymbol{t}))), \tag{4}$$

where F is a linear projection. The total loss can be written as ($\lambda_{H-ITC} = 0.1$ and $\lambda_{MIM} = 1$):

$$\mathcal{L} = \mathcal{L}_{ITM} + \mathcal{L}_{MLM} + \lambda_{H-ITC}\mathcal{L}_{H-ITC} + \lambda_{MIM}\mathcal{L}_{MIM}. \tag{5}$$

4 Experiments

We follow other works [36] and evaluate the model on four downstream tasks: Image-Text Retrieval, VQA, NLVR^2, Visual Entailment and Visual Grounding. More details about

Method	# Pre-train Images	Flickr30K (1K test set) TR			Flickr30K IR			MSCOCO (5K test set) TR			MSCOCO 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
ImageBERT [54]	6M	70.7	90.2	94.0	54.3	79.6	87.5	44.0	71.2	80.4	32.3	59.0	70.2
UNITER [12]	4M	80.7	95.7	98.0	66.2	88.4	92.9	-	-	-	-	-	-
ViLT [32]	4M	73.2	93.6	96.5	55.0	82.5	89.8	56.5	82.6	89.6	40.4	70.0	81.1
ALBEF [36]	4M	90.5	98.8	99.7	76.8	93.7	96.7	68.7	89.5	94.7	50.1	76.4	84.5
CLIP [55]	400M	88.0	98.7	99.4	68.7	90.6	95.2	58.4	81.5	88.1	37.8	62.4	72.2
ALIGN [30]	1.2B	88.6	98.7	99.7	75.7	93.8	96.8	58.6	83.0	89.7	45.6	69.8	78.6
ALBEF*	1.1M	81.3	94.9	98.2	64.7	87.2	92.0	59.5	84.4	91.2	38.3	68.0	79.1
ViCHA	1.1M	**86.0**	**97.1**	**99.0**	**72.6**	**91.1**	**95.0**	**63.1**	**86.9**	**92.7**	**47.1**	**73.8**	**82.8**
ViCHA†	800K	85.0	96.8	98.7	71.6	91.0	94.9	-	-	-	-	-	-

Table 3: Zero-shot Comparison with SOTA on Flickr30K (after fine-tuning ViCHA on COCO) and COCO (pretrained model only).

downstream tasks, implementation details, comparability and dataset filtering can be found in the Appendix.

Our setup: We favor low data and compute regimes. Specifically, we pretrain only on 3 public datasets; COCO [43], Visual Genome [33] and SBU [52], which account to $\sim 1.1M$ images in total, thus, 4 times lower than other approaches (*e.g.* 4M images [12, 20, 36]). In addition, we train for 10 epochs (in contrast to 30 epochs [22, 36, 81]) with relatively small batch size of 128 (32 per GPU) using 4 GPUs.

Filtered dataset: We go further and apply a CLIP-based filtering technique to reduce the number of images to $\sim 800k$, and train our approach on this dataset (ViCHA†). This dataset consists of; COCO, 50% of VG captions and 70 % of SBU.

Implementation details: We follow the implementation of ALBEF [36], including the same architecture, with our setup. The visual encoder is ViT-B/16 [19, 71], the text encoder is the first 6 layers of BERT-base [18] and the multimodal encoder is the last 6 layers of BERT-base. E_{vc} is the first 2 layers of BERT-base. We extract 15 concepts for each image and we set p_{vc}=30% for VCA. For H-ITC loss, we use the last 6 layers of E_v and the all 6 layers of E_l. For U-MIM, we use 2-layer transformer encoder.

To have a fair comparison, we compare with ALBEF trained on the same setup (called ALBEF*). Even though it is common to compare among all approaches in the literature, we think that it is hard to assess the methods as they follow different training setups.

Comparison with SOTA on standard tasks: Table 1 shows a comparison with other approaches, on *VQA, NLVR*^{2}*, VE* following the finetuning setup. We outperform significantly ALBEF* (+1.04% VQA, +1.55% NLVR^2 and +0.88% VE) and other approaches trained on more data (*i.e.*, 4M images), such as ViLT (+2.61% VQA, +2.03% NLVR^2), UNITER (+0.85% VQA, and +0.37% VE) and OSCAR (+0.39% VQA). For *Image-Text Retrieval*, The model is evaluated on COCO and Flickr30K (F30K) following 2 setups; finetuning and zero-shot. Table 2 shows the finetuning results. Compared to ALBEF*, our approach achieves significant improvements, especially on R@1 with absolute improvement of 3.74% IR and 3.6% TR on Flickr30K, and 2.59% IR and 2.42% TR on the more challenging COCO. Interestingly, we outperform other approaches trained on more data such as ViLT, UNITER and UNIMO (+2.58% R@1 IR F30K). Compared to the SOTA trained with 4M images, we outperform ALBEF [36] on COCO (+1.72% TR RSUM and +0.88% IR RSUM). On zero-shot F30K, we follow other approaches [22, 36, 81] and use the model finetuned on COCO. While for zero-shot COCO, we directly use the model after pretraining (without VCA). Note that, by zero-shot, we mean the model is not explicitly trained on the target dataset after pretraining.

VCs	H-ITC	U-MIM	VCA	Flickr30K (1K test set) TR RSUM	R@1	R@5	R@10	IR R@1	R@5	R@10	VQA test-dev Acc.
Baseline (ALBEF*)				535.8	85.8	97.4	98.5	70.2	89.9	94.0	71.1
✓				545.4	87.6	97.6	99.3	74.0	91.6	95.3	71.6
	✓			545.1	87.3	98.1	99.3	73.2	92.0	95.2	72.0
		✓		543.6	88.6	97.6	99.2	72.3	91.1	94.8	71.8
✓	✓			547.6	88.3	97.6	99.3	73.7	92.7	96.0	71.8
✓	✓	✓		550.0	89.8	97.6	99.3	74.9	92.8	95.6	71.7
✓	✓	✓	✓	550.7	88.2	97.8	99.4	75.7	93.4	96.2	72.6

Table 4: Ablation study: The models are finetuned on Flickr30K and VQA v2 (as in [20, 81]).

Table 3 shows that our model significantly outperforms ALBEF* (+7.88% F30K and +8.76% COCO R@1 IR) and other approaches trained on more data such as UNITER, ViLT, CLIP on F30K IR and COCO, as well as ALIGN on COCO.

Ablation Study: Table 4 shows the contribution of the different components of our ViCHA strategy. Using VCs seems to give significant improvements over the baseline (+9.6% RSUM and +0.5% Acc. VQA), similarly for H-ITC and MIM. Then we add our H-ITC loss, which also brings additional points (+2.2% RSUM and +0.2% Acc. VQA), showing the importance of a stronger pre-alignment loss. We favor the MAE based, unimodal MIM loss which improves the results by 2.4% RSUM. We choose U-MIM over M-MIM due to its superior performance, however, both objectives gives better result than the baseline; +7.8% and +3.2% RSUM for U-MIM and M-MIM respectively (detailed results can be found in the appendix). This reveals the importance of MIM as well as using SSL objectives for VLP. We finally add VCA, that significantly help the VQA task (+0.9% Acc.). Overall, we show that all contributions are not antagonistic and can be combined effectively. More ablation can be found in the appendix.

Data and Compute Efficiency: The way the VCs are extracted (using CLIP) is not central to our work, as for the gain coming from 400M pairs of CLIP. This is supported in our experiments in the supplementary material, where we replace CLIP by an object detector. In addition, Table 4 shows that other components contribute significantly, where the gain coming from H-ITC and U-MIM individually, is comparable to VCs. The number of params of ViCHA during inference is 248.5 M compared to 209.9 M for ALBEF and $\sim$ 265 M for METER [20]. The training time of ViCHA (1M) is $\sim$ 70 hours, for ViCHA (800K) the training time is much lower (39h). This is with the paper setup (4 GPUs A100). Compared to other SOTA, using 8GPUs A100; ALBEF [36] takes 2-3 days and METER 8 days [20].

Additional Scores on Visual Grounding: Table 5 shows a comparison between Weakly Supervised Visual Grounding (WSVG) approaches on the RefCOCO+ dataset. We outperform ALBEF* (+0.67 % val, +1.06 % TestA and + 1.19 % TestB) as well other approaches such as DTMR Sun et al. [69] and KPRN Liu et al. [46]. Compared to ALBEF trained on 4M images we obtain better performance on the TestA set (+ 0.68 %). In addition, we show some qualitative results using Grad-CAM by back propagating the ITM loss until the 3rd layer of the multimodal encoder, the results can be found in the appendix.

Data filtering: Interestingly, the model trained on the filtered data (ViCHA†) gives comparable results on image-text retrieval (Table 2 and 3), sometimes better than, training on all data (*e.g.*, NLVR^2 and SNLI-VE, Table 1), while being always better than ALBEF*.

Illustration of VCs: We show (Figure 3) that VCs can capture high level, global and some aspects in the scene that can not be shown explicitly or detected by other techniques (*e.g.*,

Method	Val	TestA	TestB
ARN [45]	32.78	34.35	32.13
CCL [90]	34.29	36.91	33.56
KPRN [46]	35.96	35.24	36.96
DTMR [59]	39.18	40.01	38.08
ALBEF [36]	58.46	65.89	46.25
ALBEF*	57.00	65.51	44.24
ViCHA	57.67	66.57	45.63
ViCHA†	56.40	65.93	45.69

Table 5: Comparison with SOTA on Weakly Supervised Visual Grounding RefCOCO+.

Object detectors as in OSCAR [40]).

Dataset: COCO
Caption: "Prepping vegetables with a chef's knife prior to executing a recipe."
VCs: 'demonstration kitchen', 'cooking demonstration', 'kitchen restaurant scene', 'galley kitchen', 'kitchen scene', 'type kitchen', 'meal preparation', 'food preparation', 'chef', 'chef coat', 'lettuces', 'chef knife', 'efficiency kitchen', 'plan kitchen', 'kitchen worker'

Dataset: SBU
Caption: ' Mom opening up a bottle of cider mix from Catherine. Stupid tree was in my way.'
VCs: 'santa juana', 'mama santa', 'bev pelletier', 'pam tree', 'amazon', 'sue veeder', 'christmaspresents', 'christmas woman', 'christmas basket', 'anne peterson', 'santa gifts', 'holiday cheer', 'pamela hinshaw', 'cynthia rybakoff', 'christmas'

Dataset: VG
Caption: 'the hat is pink'
VCs:'birthday celebrate', 'birthday celebration', 'birthday party', 'cake cutting', 'celebration cake', 'cake portion', 'birthay cake', 'bithday cake', 'child cake', 'birth day', 'celebration', 'cutting cake', 'birthday', 'congratulations cake', 'birthday cake'

Figure 3: Illustration of VCs: VCs capture global information, actions and other aspects of the scene that help to give a rich context.

5 Conclusion

We propose a new efficient VLP approach centered on 3 main components; stronger Vision-Language pre-alignment through hierarchical contrastive objective, self supervision via masked image modeling based on MAE, and a new Visual Concepts injection and extraction technique. The approach outperforms state of the models trained on the same setup (+3.6% F30K TR and +1.04 Acc. VQA compared to ALBEF), as well as others trained on much more data (times 4 more data, such as OSCAR and ViLT). Overall, we show that investing in the learning schemes is a very promising approach that helps to exploit more effectively the data, especially at low data regime. We hope that this work will encourage more effort (from a wider range of research laboratories) in this direction, that might lead to have more mature techniques, and then, perhaps, wisely leverage them when going large scale.

6 Acknowledgments

The authors would like to thank Corentin Dancette and Arthur Douillard for fruitful discussion and Christophe Bouder for technical support. This work was partly supported by ANR grant VISA DEEP (ANR-20-CHIA-0022), and HPC resources of IDRIS under the allocation 2022-[AD011013415] and 2022-[A0121012449] made by GENCI.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022.

[2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *European conference on computer vision*, pages 382–398. Springer, 2016.

[3] Alex Andonian, Shixing Chen, and Raffay Hamid. Robust cross-modal representation learning with progressive self-distillation. *arXiv preprint arXiv:2204.04588*, 2022.

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

[5] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. *Advances in neural information processing systems*, 32, 2019.

[6] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEit: BERT pre-training of image transformers. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=p-BhZSz59o4.

[7] Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. Mutan: Multimodal tucker fusion for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2612–2620, 2017.

[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[9] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021.

[10] 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.

[11] Feilong Chen, Duzhen Zhang, Minglun Han, Xiuyi Chen, Jing Shi, Shuang Xu, and Bo Xu. Vlp: A survey on vision-language pre-training. *arXiv preprint arXiv:2202.09061*, 2022.

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

[13] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*, 2019.

[14] Guillaume Couairon, Asya Grechka, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Flexit: Towards flexible semantic image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18270–18279, 2022.

[15] Yuhao Cui, Zhou Yu, Chunqi Wang, Zhongzhou Zhao, Ji Zhang, Meng Wang, and Jun Yu. Rosita: Enhancing vision-and-language semantic alignments via cross-and intra-modal knowledge integration. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 797–806, 2021.

[16] Fahim Dalvi, Abdul Rafae Khan, Firoj Alam, Nadir Durrani, Jia Xu, and Hassan Sajjad. Discovering latent concepts learned in BERT. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=POTMtpYI1xH.

[17] Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. Redcaps: Web-curated image-text data created by the people, for the people. *arXiv preprint arXiv:2111.11431*, 2021.

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

[19] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

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

[21] Yifan Du, Zikang Liu, Junyi Li, and Wayne Xin Zhao. A survey of vision-language pre-trained models. *arXiv preprint arXiv:2202.10936*, 2022.

[22] Jiali Duan, Liqun Chen, Son Tran, Jinyu Yang, Yi Xu, Belinda Zeng, and Trishul Chilimbi. Multi-modal alignment using representation codebook. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15651–15660, 2022.

[23] 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.

[24] Yuting Gao, Jinfeng Liu, Zihan Xu, Jun Zhang, Ke Li, and Chunhua Shen. Pyramidclip: Hierarchical feature alignment for vision-language model pretraining. *arXiv preprint arXiv:2204.14095*, 2022.

[25] 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.

[26] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. *CoRR*, abs/2111.06377, 2021. URL https://arxiv.org/abs/2111.06377.

[27] Lisa Anne Hendricks, John Mellor, Rosalia Schneider, Jean-Baptiste Alayrac, and Aida Nematzadeh. Decoupling the role of data, attention, and losses in multimodal transformers. *Transactions of the Association for Computational Linguistics*, 9:570–585, 2021. doi: 10.1162/tacl_a_00385. URL https://aclanthology.org/2021.tacl-1.35.

[28] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*, 2018.

[29] 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.

[30] 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.

[31] Lei Ke, Wenjie Pei, Ruiyu Li, Xiaoyong Shen, and Yu-Wing Tai. Reflective decoding network for image captioning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8888–8897, 2019.

[32] 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.

[33] 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.

[34] Chenliang Li, Ming Yan, Haiyang Xu, Fuli Luo, Wei Wang, Bin Bi, and Songfang Huang. Semvlp: Vision-language pre-training by aligning semantics at multiple levels. *arXiv preprint arXiv:2103.07829*, 2021.

[35] 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.

[36] 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, 2021.

[37] 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.

[38] Liunian Harold Li, Haoxuan You, Zhecan Wang, Alireza Zareian, Shih-Fu Chang, and Kai-Wei Chang. Unsupervised vision-and-language pre-training without parallel images and captions. *arXiv preprint arXiv:2010.12831*, 2020.

[39] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. *arXiv preprint arXiv:2012.15409*, 2020.

[40] 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.

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

[42] Zejun Li, Zhihao Fan, Huaixiao Tou, and Zhongyu Wei. Mvp: Multi-stage vision-language pre-training via multi-level semantic alignment. *arXiv preprint arXiv:2201.12596*, 2022.

[43] 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.

[44] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*, 2021.

[45] Xuejing Liu, Liang Li, Shuhui Wang, Zheng-Jun Zha, Dechao Meng, and Qingming Huang. Adaptive reconstruction network for weakly supervised referring expression grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[46] Xuejing Liu, Liang Li, Shuhui Wang, Zheng-Jun Zha, Li Su, and Qingming Huang. Knowledge-guided pairwise reconstruction network for weakly supervised referring expression grounding. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 539–547, 2019.

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

[48] 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.

[49] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10437–10446, 2020.

[50] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. *arXiv preprint arXiv:2112.12750*, 2021.

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

[52] Vicente Ordonez, Girish Kulkarni, and Tamara L Berg. Im2text: Describing images using 1 million captioned photographs. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, NIPS'11, page 1143–1151, Red Hook, NY, USA, 2011. Curran Associates Inc. ISBN 9781618395993.

[53] Theodoros Pissas, Claudio S Ravasio, Lyndon Da Cruz, and Christos Bergeles. Multi-scale and cross-scale contrastive learning for semantic segmentation. *arXiv preprint arXiv:2203.13409*, 2022.

[54] Di Qi, Lin Su, Jia Song, Edward Cui, Taroon Bharti, and Arun Sacheti. Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data. *arXiv preprint arXiv:2001.07966*, 2020.

[55] 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.

[56] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems*, 34, 2021.

[57] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[58] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[59] Fawaz Sammani, Tanmoy Mukherjee, and Nikos Deligiannis. Nlx-gpt: A model for natural language explanations in vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8322–8332, 2022.

[60] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

[61] Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D. Manning. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Proceedings of the Fourth Workshop on Vision and Language*, pages 70–80, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-2812. URL https://aclanthology.org/W15-2812.

[62] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How much can CLIP benefit vision-and-language tasks? In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=zf_Ll3HZWgy.

[63] Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, 2020.

[64] Mustafa Shukor, Guillaume Couairon, Asya Grechka, and Matthieu Cord. Transformer decoders with multimodal regularization for cross-modal food retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4567–4578, 2022.

[65] 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.

[66] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*, 2019.

[67] Yixuan Su, Tian Lan, Yahui Liu, Fangyu Liu, Dani Yogatama, Yan Wang, Lingpeng Kong, and Nigel Collier. Language models can see: Plugging visual controls in text generation. *arXiv preprint arXiv:2205.02655*, 2022.

[68] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1644. URL https://aclanthology.org/P19-1644.

[69] Mingjie Sun, Jimin Xiao, Eng Gee Lim, Si Liu, and John Y Goulermas. Discriminative triad matching and reconstruction for weakly referring expression grounding. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):4189–4195, 2021.

[70] Hao Tan and Mohit Bansal. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1514. URL https://aclanthology.org/D19-1514.

[71] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021.

[72] Jesse Vig and Yonatan Belinkov. Analyzing the structure of attention in a transformer language model. *arXiv preprint arXiv:1906.04284*, 2019.

[73] Jianfeng Wang, Xiaowei Hu, Zhe Gan, Zhengyuan Yang, Xiyang Dai, Zicheng Liu, Yumao Lu, and Lijuan Wang. Ufo: A unified transformer for vision-language representation learning. *arXiv preprint arXiv:2111.10023*, 2021.

[74] Wenhui Wang, Hangbo Bao, Li Dong, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *arXiv preprint arXiv:2111.02358*, 2021.

[75] Zekun Wang, Wenhui Wang, Haichao Zhu, Ming Liu, Bing Qin, and Furu Wei. Distilled dual-encoder model for vision-language understanding. *arXiv preprint arXiv:2112.08723*, 2021.

[76] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11686–11695, 2022.

[77] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. SimVLM: Simple visual language model pretraining with weak supervision. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=GUrhfTuf_3.

[78] Hao Wu, Jiayuan Mao, Yufeng Zhang, Yuning Jiang, Lei Li, Weiwei Sun, and Wei-Ying Ma. Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[79] Qi Wu, Chunhua Shen, Lingqiao Liu, Anthony Dick, and Anton Van Den Hengel. What value do explicit high level concepts have in vision to language problems? In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 203–212, 2016.

[80] Hongwei Xue, Yupan Huang, Bei Liu, Houwen Peng, Jianlong Fu, Houqiang Li, and Jiebo Luo. Probing inter-modality: Visual parsing with self-attention for vision-and-language pre-training. *Advances in Neural Information Processing Systems*, 34, 2021.

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

[82] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021.

[83] Yuan Yao, Bowen Dong, Ao Zhang, Zhengyan Zhang, Ruobing Xie, Zhiyuan Liu, Leyu Lin, Maosong Sun, and Jianyong Wang. Prompt tuning for discriminative pre-trained language models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3468–3473, 2022.

[84] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4651–4659, 2016.

[85] 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.

[86] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

[87] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021.

[88] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[89] 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.

[90] Zhu Zhang, Zhou Zhao, Zhijie Lin, Jieming Zhu, and Xiuqiang He. Counterfactual contrastive learning for weakly-supervised vision-language grounding. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

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