# Capture Concept through Comparison: Vision-and-Language Representation Learning with Intrinsic Information Mining

Yun-Zhu Song[1], Yi-Syuan Chen[1], Tzu-Ling Lin[1], Bei Liu[2], Jianlong Fu[2], and Hong-Han Shuai[1]

[1] National Yang Ming Chiao Tung University
{yzsong.ee07, yschen.ee09, tzulinglin.11, hhshuai}@nycu.edu.tw
[2] Microsoft Research
{Bei.Liu, jianf}@microsoft.com

**Abstract.** Achieving alignment between vision and language semantics poses a critical challenge. Prior works have sought to enhance alignment by incorporating additional supervision, such as tags or object bounding boxes, as anchors between modalities. However, these methods predominantly concentrate on aligning tangible entities, disregarding other crucial abstract *concepts* that elude perception, such as *side by side*. To overcome this limitation, we propose a novel approach to **C**apture various **C**oncepts through data **C**omparison (C3) for learning cross-modal representations. Specifically, we devise a data mining procedure to uncover intrinsic information within the database, avoiding the need for external annotations. Furthermore, we distinctly frame model inputs as triplets to better elucidate abstract semantics in images. Building upon this formulation, we propose two concept-centric pre-training objectives to signify concept learning. Extensive experiments show that models trained within the C3 framework consistently achieve significant enhancements across a wide range of comprehension and reasoning benchmarks, whether starting from scratch or fine-tuning from an existing model.

**Keywords:** Vision-and-Language Learning · Concept Learning · Information Mining

## 1 Introduction

Semantic alignment between the domains of vision and language emerges as a crucial concern for various vision-language (VL) tasks. Consequently, numerous pre-training objectives have been meticulously designed to investigate the pairing relations between images and texts using large-scale datasets [29,54,58]. However, the information in the two modalities is often inequivalent for most existing datasets. In other words, textual descriptions often fall short of providing a comprehensive account of each image [56]. Such a weakly-aligned relation hinders the effective learning of cross-modal representations. Meanwhile, fine-grained alignments across modalities cannot be naturally achieved due to the lack of explicit annotations between entities and regions.

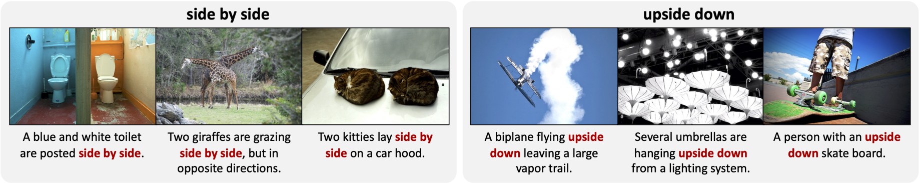

side by side
A blue and white toilet are posted side by side.
Two giraffes are grazing side by side, but in opposite directions.
Two kitties lay side by side on a car hood.
upside down
A biplane flying upside down leaving a large vapor trail.
Several umbrellas are hanging upside down from a lighting system.
A person with an upside down skate board.

**Fig. 1:** Examples of concepts mined from the database [4]. The concepts could be abstract and shared across different scenes and subjects. The semantics of the concepts become clear when comparing multiple images carrying the same concept.

To alleviate this problem, prior works have sought to leverage additional supervision to bridge the gap between images and texts. For example, pre-trained object detectors are widely adopted. The detectors can be used to extract region-based features as visual inputs [21,30,46,53,63], provide detected tags as additional inputs to enhance alignments [32,61], or create learning targets for knowledge distillations [36]. In addition to object detectors, other works attempt to obtain visual attributes in linguistic form through entity prompter [28] or noun phrases of captions [10]. However, prior efforts leveraging additional supervision still have limitations. Notably, these approaches focus on aligning data with concrete entities such as objects, regions, or attributes, which lack clear indications for aligning complex concepts that are challenging to precisely depict in the visual domain, such as “side by side” or “upside down” as shown in Fig. 1.

Another line of research, instead of relying on additional supervision, focuses on enhancing alignments through modifications to pre-training objectives or architectures [10,18,25,29,53,55]. For instance, ALBEF [29] adopts an intermediate image-text contrastive loss to align the image and text features before performing cross-modal interactions in later layers. In addition to cross-modal alignments, TCL [56] further applies contrastive learning for intra-modal alignment by image or text augmentation. Nevertheless, most VL pretrained models still suffer from two issues. First, the supervision for alignments is limited in terms of diversity and quantity, often relying on the use of external models or predefined categories. Second, there is a lack of clear indications for learning concepts with abstract semantics, a critical requirement for tasks that demand comprehension and reasoning.

To tackle these challenges, we present a novel approach called **Capture Concept through Comparison (C3)**. The term “concept”, rooted in psychology, is defined as *“the label of a set of things that have something in common”* [1]. Inspired by the definition, we posit that a concept shall become more evident as more examples are provided. Therefore, the core idea of C3 is to leverage the data comparison to achieve concept-level alignments, thereby enhancing the comprehension of abstract semantics. To this end, we first propose a mining procedure to discover the concepts that are intrinsically shared among the database. Specifically, given an image-text pair, we extract text fragments and compare them with other texts in the training data. A fragment is identified as a concept if the same fragment appears in other texts. As such, this mining approach enables us to discover a broader spectrum of concepts without being constrained by external detectors or linguistic grammar.

Equipped with the mined concepts, the next challenge is to harness them for enhancing cross-modal alignment. An intuitive approach is to employ the concepts directly as language input and adhere to the conventional VL pre-training pipeline. However, we have discovered that employing an image-image-text triplet, i.e., two images with a concept text, can further assist models in discerning the abstract concept intertwined within images. Such a triplet formulation enables our model to learn a concept by pinpointing the "intersection" between two images, thereby streamlining the information to be focused in the visual domain and refining the alignment of concepts. With this input formulation, we design two concept-centric learning objectives, *Matched Concept Prediction (MCP)* and *Matched Concept Contrastive (MCC)*, to enhance alignments for both cross- and uni-modal representations. These objectives offer a direct learning mechanism for the mined concepts.

Finally, we assess our method under two configurations: continual pre-training and pre-training from scratch. Experimental results demonstrate that our approach can effectively leverage existing models without full re-training and significantly improve general VL behavioral testing. Furthermore, the experiments of pre-training from scratch highlight the benefits of concept-centric learning on various downstream tasks. Our main contributions can be summarized as follows: (i) We propose a novel mining procedure to discover the concepts intrinsic to the database, which is general and could potentially be leveraged in other studies as the immediate supervision for fine-grained alignments; (ii) We reformulate image-text learning scheme by considering image-image-text triplets, which facilitates models to identify and learn the abstract semantics in both modalities; (iii) We design two novel concept-centric objectives, i.e., Matched Concept Prediction (MCP) and Matched Concept Contrastive (MCC), to learn the matching of triples for better concept-level alignment; (iv) Extensive experiments and analysis demonstrate that the proposed concept-centric learning can improve both model capacity and downstream task performances.

## 2 Related Work

Aligning vision and language representations is a critical challenge in VL pre-training. Recent works have attempted to address this challenge by leveraging additional supervision beyond traditional image-text pairs. For instance, [46] uses Faster R-CNN to extract region of interest (RoI) features, while [32] introduces object tags as the anchors for alignment. [61] improves on this approach by enhancing the visual representations via better pre-training of object detectors. Similarly, [53] incorporates object detection objectives [2] into a sequence-to-sequence VL model, and [36] relies on external detectors for object knowledge distillation. Other approaches include leveraging object detectors and phrase generators to learn hierarchical alignment [33] and performing contrastive learning with patch features and bounding boxes [60]. Additionally, [12] proposes aligning vision and language at different semantic levels, i.e., global image, local region, and ROI features for vision; summarization, caption, and attributes for language. [8] proposes to teach VL models structure concepts by manipulating the text input based on pre-defined rules. [3] proposes leveraging synthetic data to improve alignment.

Another research line focuses on aligning solely with image-text pairs through modifications in objectives or architectures. [29] introduces an intermediate image-text contrastive loss on unimodal features to facilitate subsequent cross-modal alignments. Other approaches suggest additional pre-training objectives, including word-region contrastive loss [19,57], pseudo-labeled keyword prediction [24], weakly-supervised phrase grounding [34], token-wise maximum similarity [58], and visual dictionary as pixel-level supervision [17]. [9] encodes features into a shared coding space defined by a dictionary of cluster centers for alignment. [56] introduces intra-modal contrastive objectives to complement the cross-modal objectives. The integration of learning across vision, language, and multimodal tasks has been studied in [45]. A two-stage pre-training strategy has also been suggested in [6], involving initial coarse-grained training based on image-text data, followed by fine-grained training on image-text-box data.

In summary, the aforementioned approaches, aimed at enhancing alignments between vision and language, have showcased significant successes across a range of downstream tasks. Nevertheless, challenges persist in learning intricate concepts that resist easy specification through perceptual features. Although prior research has utilized additional supervision with respect to tangible entities like objects, regions, or attributes, these strategies are constrained when dealing with more abstract concepts. To overcome these limitations, we propose a framework that mines hidden concepts in the dataset and reformulates input based on the philosophy of the mining procedures, thereby enabling more effective alignment between vision and language.

## 3 Method

In this section, we propose a learning procedure for enhancing the fundamental abilities of the VL models, which can be effectively applied to both continual pre-training and pre-training from scratch. First, Sec. 3.1 describes the model architectures for better illustration. Sec. 3.2 and Sec. 3.3 introduce concept mining strategy and concept-centric objectives.

### 3.1 Overall Framework

Fig. 2 depicts an overview of C3, which comprises a text encoder $\mathcal{E}_t$, an image encoder $\mathcal{E}_v$, and a cross-modal encoder $\mathcal{E}_{cross}$ as contemporary vision-language models [25,29,61]. We adopt such a succinct architecture and focus on studying the proposed concept-centric pre-training. A text $T$ is tokenized into a sequence of subwords $[t_1, t_2, ...]$, and two special tokens $t_{cls}$ and $t_{sep}$ are respectively prepended and appended to the sequence. The sequence is then passed

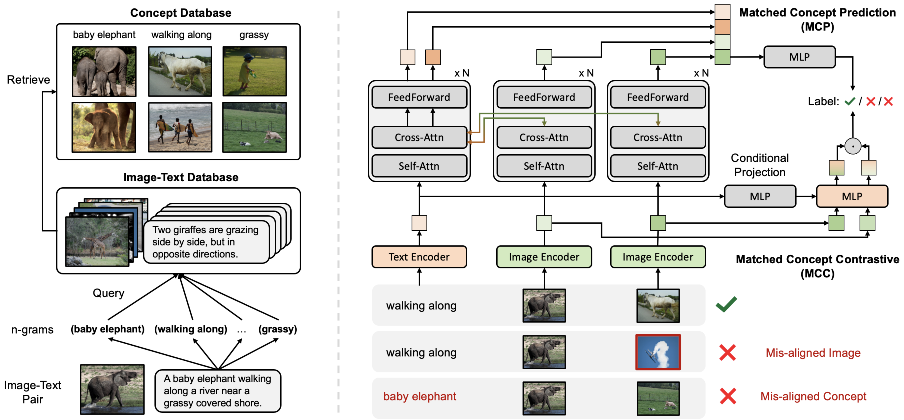


**Fig. 2:** The proposed concept mining procedure (left) and concept-centric pre-training architecture and objectives (right). As shown on the left side, we use the n-grams from the text as the queries to retrieve images also containing the n-grams in their texts. With the mined n-gram concepts, the inputs are formulated as triplets for pre-training, shown on the right side.

through $\mathcal{E}_t$ to obtain the unimodal features. An image $I$ is first divided into several patches and processed by a convolutional layer to extract patch features $[v_1, v_2, ...]$. These patch features are then flattened and fed into the $\mathcal{E}_v$ for further feature extraction. We also add a learnable vector $v_{cls}$ to aggregate global information for the vision modality. For fusing the features from unimodal encoders, we apply co-attention modules [7] as cross-modal encoders $\mathcal{E}_{cross}$ for both vision and language. Finally, for an image-text pair, the vision-language joint representation $z$ is obtained as follows:

$$\begin{aligned} H_t = [h^t_{cls}, h^t_1, ...] = \mathcal{E}_t([t_{cls}, t_1, ...]), H_v = [h^v_{cls}, h^v_1, ...] = \mathcal{E}_v([v_{cls}, v_1, ...]), \\ Z = [z^t_{cls}, z^t_1, ..., z^v_{cls}, z^v_1, ...] = \mathcal{E}_{cross}([H_t, H_v]), z = [z^t_{cls}, z^v_{cls}]. \end{aligned} \tag{1}$$

We pre-train C3 from scratch with the proposed concept-centric objectives (Section 3.3), i.e., Matched Concept Prediction (MCP) or Matched Concept Contrastive (MCC), and the widely-used pair-centric objectives, Image Text Matching (ITM) and Masked Language Modeling (MLM). For the continual pre-training, we insert trainable low-rank residual adapters (LoRA [16]) into existing models and learn with the MCC objective to enhance the capacity.

## 3.2 Concept-centric Learning Formulation

**Mutual Information Maximization.** Existing works [5, 7, 25, 29, 32, 46, 61] commonly adopt the combinations of Masked Language Modeling (MLM), Masked Vision Modeling (MVM), and Image-Text Matching/Contrastive (ITM/ITC) for VL pre-training. Previous works [29, 47] have shown that many existing objectives can be interpreted as the maximization of the mutual information (MI) between different views of an image-text pair. For example, ITC treats the image and text as two different views; MLM treats the masked tokens as one view and other tokens with the image as another view. In other words, these approaches aim to learn the multimodal representations invariant across different views for improving downstream tasks. It is noteworthy that the aforementioned approaches maximize the MI for each image-text pair independently. Besides, the views are considered either at the instance-level (ITC) or token-/patch-level (MLM/MVM). However, we argue that models can be better learned by considering *concepts* that are cross-data and diverse in granularity. Therefore, instead of considering only a single image-text pair, we construct a novel learning formulation by centering a concept that is built from multiple image-text pairs.

Specifically, we define two random variables $c_1$ and $c_2$ as two different views of a concept, where the views correlate to *a concept text* and *multiple images*. We could maximize a lower bound on $\mathrm{MI}(c_1, c_2)$ by minimizing the InfoNCE loss [29,38] defined as follows.

$$\mathcal{L}_{\mathrm{NCE}} = -\mathbb{E}_{p(c_1,c_2)} \log \frac{\exp(f(c_1, c_2))}{\sum_{c_2' \in B} \exp(f(c_1, c_2'))}, \tag{2}$$

where $f(\cdot)$ is a scoring function and $B$ is a batch containing one positive sample with other negative samples. To realize this learning framework, we next elaborate on the proposed methods for mining concepts in the database.

**Concept Mining**. Drawing inspiration from the field of psychology, which defines a concept "as the label of a set of things that have something in common" [1], we propose to mine the concepts by exploring the commonality between pairs of data. Our approach is based on the identification of *overlapping n-grams* [20, 50] between pairs of data, specifically image-text pairs in a database $\mathcal{D} = \{(I_i, T_i = \{t_{i1}, t_{i2}, ...\})\}_{i=1}^{N_\mathcal{D}}$. In each iteration $i$, we extract all $n$-grams in the associated text $T_i$ as $\{(t_{i1}, ..., t_{in}), (t_{i2}, ..., t_{in+1}), ...\}$. Next, we treat each $n$-gram as a query to retrieve images carrying the same $n$-gram in their texts. If there is any matching, the $n$-gram is defined as a concept shared across these data. We retrieve at most $K_1$ pairs for a concept and early terminate the current iteration if $K_2$ pairs are obtained for $T_i$. To allow for concepts of varying granularity, we consider different $n \in [1, N]$ and mine the concepts with a descending order of $n$ since the longer concept covers the shorter one. In the end, each image may involve multiple concepts,

**Algorithm 1:** Concept Mining

```
Data: Image-text database
      $\mathcal{D} = \{(I_i, T_i)\}_{i=1}^{N_\mathcal{D}}$.
Result: Concept database
      $\mathcal{D}^c = \{(I_i, \mathcal{I}_i, \mathcal{C}_i)\}_{i=1}^{N_\mathcal{D}}$.
for $(I_i, T_i) \in \mathcal{D}$ do
    Initial $\mathcal{C}_i$ and $\mathcal{I}_i$ as empty sets;
      $k = 0$;
    Obtain $n$-grams $\mathcal{G}$ from $T_i$ for
      $n \in [N, ..., 1]$;
    for $G \in \mathcal{G}$ do
        Random sample a subset $\mathcal{D}_s$
          from $\mathcal{D}$;
        Initial $\mathcal{I}_k$ as empty sets;
        for $(I_j, T_j) \in \mathcal{D}_s$ do
            if $G \in T_j$ and $|\mathcal{I}_k| \leq K_1$
              then
                Assign $G$ to $C_k$; Add
                  $I_j$ to $\mathcal{I}_k$;
        end
        if $\mathcal{I}_k$ is not empty and
          $|\mathcal{I}_i| \leq K_2$ then
            Add $C_k$ to $\mathcal{C}_i$; Add $\mathcal{I}_k$ to
              $\mathcal{I}_i$; $k$ += 1;
    end
end
```

and each concept correlates to multiple different images. Let $\mathcal{C}_i$ and $\mathcal{I}_{ik} = \{I_{ikj}\}$ respectively denote the matched concept set for $i$-th sample and the matched image set of $k$-th concept $\mathcal{C}_{ik}$. Accordingly, the concept database is constructed as follows:

$$\mathcal{D}^c = \{\{(I_i, \mathcal{I}_{ik}, \mathcal{C}_{ik})\}_{k=1}^{|\mathcal{C}_i|}\}_{i=1}^{N_\mathcal{D}}. \tag{3}$$

The mining procedure is further presented in Algorithm 1.

### 3.3 Concept-centric Pre-training

Based on the concept database, we propose to maximize the mutual information across an n-gram concept and two corresponding images. The input formulation is therefore transformed from pair-based (i.e., image-text) to triplet-based (i.e., image-image-text). In the following, we present two concept-centric objectives to explore this formulation.

**Matched Concept Prediction (MCP).** Different from prior works, MCP takes a concept text and a pair of images as the input. This objective aims to predict whether the concept $C$ is shared between the two images $(I_i, I_j)$, which provides explicit supervision to learn the semantics of concepts across modalities. An image could encapsulate numerous concepts in different granularity, and the contrasts of two images help capture and identify the specified concept more efficiently. For a triplet $(I_i, I_j, C)$, we divide it into $(I_i, C)$ and $(I_j, C)$ to encode them respectively. Let $z_{ij}^c$ denote the concatenation of joint representations from $(I_i, C)$ and $(I_j, C)$. To obtain the negative examples for learning, we investigate two strategies. The first one is to replace one of the images in a positive triplet with a mismatched image $I'_j$, i.e., $(I_i, I'_j, C)$, while the other is to replace the concept with another concept $C'$, i.e., $(I_i, I_j, C')$. As such, we could define the objective as:

$$\widehat{\mathcal{L}}_{\text{MCP}} = -\mathbb{E}_{p(I_i,I_j,C)}[\log \frac{\exp(\psi^\top z_{ij}^c)}{\sum_{(I'_j,C')\in B} \exp(\psi^\top z_{ij'}^{c'})}], \tag{4}$$

where $\psi$ is a learnable matrix. However, since the objective utilizes multimodal representations, it requires forwarding all triplets in a batch independently despite some images or concepts being shared, making the optimization memory-intensive in practice. Therefore, we adopt the local NCE [14,15] to approximate the loss [26,35] as:

$$\mathcal{L}_{\text{MCP}} = -\mathbb{E}_{p(I_i,I_j,C)}[y_{ij}^c \log \phi_{\text{MCP}}(z_{ij}^c) + (1 - y_{ij}^c) \log(1 - \phi_{\text{MCP}}(z_{ij}^c))], \tag{5}$$

where $y_{ij}^c$ is the label and $\phi_{\text{MCP}}$ is a network producing a value as probability. This formulation in another way leverages a binary classifier to distinguish matched samples from the noisy ones.

**Matched Concept Contrastive (MCC)**. Apart from utilizing cross-modal representations for learning the alignment of concepts, we could extend such an idea with the unimodal ones. Specifically, we use the outputs of the image and text encoders to learn the matching of triplets before the cross-modal layers.

This strategy could be beneficial for the cooperated objectives to leverage the aligned representation in an early stage. The objective is presented as:

$$\mathcal{L}_{\text{MCC}} = -\mathbb{E}_{p(I_i,I_j,C)}[\log \frac{\exp(s(\psi_c^\top h_i, \psi_c^\top h_j))}{\sum_{(I'_j,C')\in B}\exp(s(\psi_{c'}^\top h_i, \psi_{c'}^\top h_{j'}))}], \tag{6}$$

where $s(\cdot)$ is cosine similarity, $h_i = h_{cls}^{I_i}$ ,$\psi_c = \phi_{\text{MCC}}(h_{cls}^{C})$, and $\phi_{\text{MCC}}$ is an MLP-based network. In this formulation, we use the concept to generate a projection matrix $\psi_c$ that transforms the two images into a space that is conditioned on the concept.

## 4 Experiments

### 4.1 Experimental Settings

**Datasets and Benchmarks**. We conduct pre-training on four image-caption datasets: COCO [4], Visual Genome (VG) [27], Conceptual Captions (CC) [43], and SBU Captions [39].[3] Our model evaluations encompass a range of vision-language benchmarks, including vision-language behavior assessment (VL-Checklist [62]), visual entailment (SNLI-VE [52]), natural language visual reasoning (NLVR$^2$ [48]), and image-text retrieval (Flickr30k [40]).

**Training Configurations.**

We evaluate our methods in two configurations: *continual pre-training* and *pre-training from scratch*. In the case of continual pre-training, our goal is to assess the advantages of applying concept-centric learning to existing models without necessitating a full model re-training. Leveraging pre-trained knowledge can prove both effective and cost-efficient. Specifically, we select CLIP [41] as our base model, which has undergone pre-training on 400 million image-text pairs and has been applied to a wide range of tasks. In this context, we introduce LoRA [16] to enhance the base model's capacity while keeping all base model parameters fixed, allowing only the parameters of LoRA to be trainable. For the pre-training from scratch scenario, we follow the setup of METER [7], given its relatively manageable pre-training scale, utilizing 4.0 million images and 5.1 million image-text pairs for pre-training.

**Implementation Details.** For continual pre-training, we initialize the model with CLIP-ViT-B/32 or CLIP-ViT-B/16 and train it for 1 epoch using COCO, VG, CC, and SBU, respectively. The trainable parameters constitute approximately 1.2% of the entire model. The concept mining procedure is executed within each dataset, considering $n$-grams with $n$ ranging from 1 to 5 as concept candidates. To ensure comprehensive coverage, we set the hyperparameters $K_1$ and $K_2$ to 5 and 80, respectively. When pre-training from scratch, our model undergoes training for 50k steps on COCO, VG, CC, and SBU, which is half the number of learning steps compared to our baseline METER [7] (pre-trained

[3] Notably, VQAv2 [13] is omitted due to its lower abstractness relative to other benchmarks, diverging from our paper's primary focus.

**Table 1:** Performance comparison on VL-Checklist [62]. *C/V/O/S* refers to the CC/VG/COCO/SBU dataset. The arrow → indicates the performance change from models w/o C3 to models w/ C3. For CLIP architecture, the model w/o C3 is an ablation of MCC.

| Base Model | Dataset | VL-Checklist | | | | |
|---|---|---|---|---|---|---|
| | | Attribute | Object | Relation | Average | $\Delta$ |
| Continual Pre-training | | | | | | |
| CLIP-ViT-B/32 | - | 69.09 | 81.94 | 63.30 | 71.44 | - |
| | S | 69.09 → 70.87 | 81.24 → 81.99 | 60.62 → 63.68 | 70.32 → 72.18 | -1.12 → 0.74 |
| | C | 68.63 → 71.18 | 80.32 → 82.04 | 53.75 → 55.72 | 67.57 → 69.65 | -3.87 → -1.79 |
| | V | 71.59 → 72.31 | 87.28 → 87.16 | 62.86 → **64.77** | 73.91 → **74.75** | 2.47 → 3.31 |
| | O | 71.43 → **75.26** | 85.38 → **87.36** | 57.66 → 61.11 | 71.49 → 74.58 | 0.05 → 3.14 |
| CLIP-ViT-B/16 | - | 70.37 | 82.94 | 61.98 | 71.76 | - |
| | S | 70.68 → 71.13 | 82.85 → 83.40 | 61.52 → 62.64 | 71.68 → 72.39 | -0.08 → 0.63 |
| | C | 69.66 → 70.18 | 81.54 → 87.56 | 55.86 → 62.86 | 69.02 → 73.53 | -2.74 → 1.77 |
| | V | 68.40 → 69.98 | 87.34 → 87.75 | 60.80 → **62.98** | 72.18 → 73.57 | 0.42 → 1.81 |
| | O | 71.30 → **75.87** | 86.94 → **88.48** | 53.98 → 61.75 | 70.74 → **75.37** | -1.02 → 3.61 |
| Pre-training from Scratch | | | | | | |
| METER | C+V+O+S | 81.65 → **84.28** | 84.72 → **89.04** | 71.94 → **73.90** | 79.44 → **82.41** | 2.97 |

with 100k steps). For ablation and analysis purposes, we train the models from scratch for 2.6k steps on COCO, enabling extensive experimentation. All images are resized to 224×224 through center-cropping during the pre-training process.

### 4.2 Vision-Language Behavioral Testing

We first assess the fundamental vision-language capability of C3 from different angles with the VL-Checklist benchmark. The comparison with the base model under different configurations is shown in Table 1. The results reveal that directly continuing the pre-training may deteriorate performance, while our methods can significantly and consistently improve the VL capability across diverse aspects. Notably, the enhancement in the object aspect is evident across various data sources. This could result from the fact that most concepts would naturally involve objects, contributing more to this aspect. Besides, the improvement in the attribute aspect is particularly pronounced in COCO, signifying that COCO includes more attributed-related descriptions, such as the size, color, and material, which C3 can effectively identify and leverage. VG includes abundant region descriptions, which enables C3 to grab the concept of objects and their relations by comparing data even without using the structured annotations in this dataset, which is hard to learn by image-caption pairs. The diverse characteristics of data sources also underscore the idea that different data sources may cover different concepts, which can be successfully exploited by C3 to enhance the model's overall capabilities. Importantly, our methods prove beneficial for both continual pre-training and pre-training from scratch settings, highlighting the generalizability of the proposed approach.

**Table 2:** Performance comparison on SNLI-VE [52] and NLVR$^2$ [48].

| Model | Images | Iters | Params | SNLI-VE | | NLVR$^2$ | |
|---|---|---|---|---|---|---|---|
| | | | | dev | test | dev | test |
| ALBEF(14M) [29] | 14M | 420M | 500M | 80.80 | 80.91 | 82.55 | 83.14 |
| SimVLM$_{\text{HUGE}}$ [51] | 1.8B | 4.1B | 632M | 86.21 | 86.32 | 84.53 | 85.15 |
| BEIT-3 [49] | 36M | 6.1B | 1.9B | - | - | 91.51 | 92.58 |
| PixelBERT [18] | 207K | 8.3M | 170M | - | - | 76.5 | 77.2 |
| Visual Parsing [54] | 221K | 8.8M | 308M | - | - | 77.61 | 78.05 |
| OSCAR$_{\text{LARGE}}$ [32] | 4M | 512M | 380M | - | - | 79.12 | 80.37 |
| ViLT [25] | 4M | 819M | 114M | - | - | 75.70 | 76.13 |
| UNITER$_{\text{LARGE}}$ [5] | 4M | - | 343M | 79.39 | 79.38 | 79.12 | 79.98 |
| VILLA$_{\text{LARGE}}$ [11] | 4M | - | 343M | 80.18 | 80.02 | 79.76 | 81.47 |
| UNIMO$_{\text{BASE}}$ [31] | 5.7M | 1.5B | 165M | 80.00 | 79.10 | - | - |
| VinVL$_{\text{BASE}}$ [61] | 5.7M | 1B | 290M | - | - | 82.05 | 83.08 |
| CLIP-ViL [44] | 4M | 184M | 330M | 80.61 | 80.20 | - | - |
| ALBEF(4M) [29] | 4M | 154M | 500M | 80.14 | 80.30 | 80.24 | 80.50 |
| METER [7] | 4M | 410M | 384M | 80.86 | 81.19 | 82.33 | 83.05 |
| C3 (our) | 4M | 205M | 384M | **81.30** | **81.34** | **82.36** | **83.35** |

### 4.3 Vision-Language Reasoning

To compare with prior works, C3 uses the same framework as our primary baseline model, METER. Given the resource-intensive nature of VL pre-training, learning efficiency is of utmost importance. Therefore, we compare C3 with models of similar data scales and learning steps. Our goal is to achieve state-of-the-art performance with fewer training steps, as the model tends to improve with increased data and learning steps [42]. As shown in Table 2, C3 attains superior performance in NLVR$^2$ and SNLI-VE, requiring fewer training iterations than METER, and surpasses ALBEF. These results suggest that learning concepts through image comparison can enhance a model's inference capabilities.[4]

### 4.4 Ablation Study

**Pre-training Objectives.** We investigate diverse pre-training settings by restricting the learning steps to 2.6k and using CLIP-ViT-224/32 as the vision encoder. Table 3 demonstrates that incorporating MCP (row 4) enhances all tasks compared to the METER baseline (row 1), particularly for image-text retrieval and the VL-Checklist, owing to the improved representations learned with multi-grained concept alignment. Furthermore, our approach naturally meets the requirements of NLVR$^2$, where models must assess the accuracy of descriptions between two images, bringing in additional benefits. Notably, the ablation study demonstrates that ITM is critical for retrieval tasks and the VL-Checklist but not for semantic inference tasks (row 3), while MLM greatly impacts the performance of the reasoning task (row 2). Thus, each objective covers distinct aspects,

[4] Additionally, we evaluate C3 on the zero-shot and fine-tuned image-text retrieval tasks to assess cross-modal representation quality, as shown in the appendix.

**Table 3:** Ablation study of C3. The first row is METER [7]. All models are trained for 2.6k steps on 224×224 images with patch size 32. $I/C$ refers to constructing negative samples by misaligned images/concepts. $PW$ refers to the pairwise formulation for the MCP objective.

| ITM | MLM | MCP | VL-Checklist<br>Att / Obj / Rel | $\Delta$ | Flickr30K-ZS<br>IR / TR | $\Delta$ | NLVR$^2$<br>dev / test | $\Delta$ | SNLI-VE<br>dev / test | $\Delta$ |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | n-gram | | | | | | | |
| ✓ | ✓ | - | 54.13 / 73.45 / 54.80 | -6.2 | 55.82 / 62.17 | -5.2 | 72.83 / 74.35 | -1.2 | 76.98 / 77.31 | -0.5 |
| ✓ | - | ✓ | 62.57 / **77.70** / 53.20 | -2.5 | **60.55** / 67.49 | -0.1 | 69.47 / 71.47 | -4.4 | 76.41 / 76.58 | -1.1 |
| - | ✓ | ✓ | 50.70 / 46.77 / 45.38 | -19.4 | 2.95 / 6.23 | -59.6 | 73.79 / 74.78 | -0.6 | **77.62** / 77.57 | 0.0 |
| ✓ | ✓ | ✓ | **64.52** / 76.53 / **60.05** | 0.0 | 60.45 / **67.87** | 0.0 | **74.27** / **75.40** | 0.0 | 77.55 / **77.67** | 0.0 |
| ✓ | ✓ | I | 61.21 / 76.98 / 57.86 | -1.7 | 58.98 / 64.87 | -2.2 | 74.01 / 74.49 | -0.6 | 77.31 / 77.53 | -0.2 |
| ✓ | ✓ | C | 61.66 / 77.30 / 57.26 | -1.6 | 60.07 / 66.20 | -1.0 | 73.33 / 74.52 | -0.9 | 77.47 / 77.66 | 0.0 |
| ✓ | ✓ | PW | 61.44 / 77.48 / 58.23 | -1.3 | 58.87 / 65.67 | -1.9 | 73.91 / 74.90 | -0.4 | 77.16 / 77.46 | -0.1 |
| | | | noun phrase | | | | | | | |
| ✓ | ✓ | ✓ | 62.03 / 77.22 / 53.74 | -2.7 | 58.50 / 64.53 | -2.6 | 73.97 / 75.25 | -0.2 | 77.23 / 77.52 | -0.2 |

and the best performance can be achieved by combining them.

**Training Sample Formation.** We investigate the impacts of two methods for constructing negative samples of MCP (row 4 & row 5-6). We refer to the misaligned image method as *type-1-negative* and the misaligned concept method as *type-2-negative*. Results indicate that type-2-negative is relatively effective since it is more challenging, forcing models to learn semantics without relying on spurious clues. Comparatively, type-1-negative is formed by replacing the matched image with a random one, where the image pairs mostly do not have clearly shared concepts. Therefore, models might be able to make predictions solely by comparing visual features. Nevertheless, the combination of both types still yields the best performance. Furthermore, to understand the effectiveness of the triplet input formulation, we compare it with pairwise input. By altering the input from triplet to pairwise, the MCP objective aims to predict whether a concept is present in an image. The results (row 7) show that triplet training still outperforms pairwise training across all metrics. This is likely because an image can contain multiple concepts, making direct alignment via pairwise training ambiguous and inefficient. In contrast, triplet training explicitly provides two references for each mined concept, reducing the number of potential concepts to be considered in the visual domain and enabling more precise alignment. Additionally, pairwise training still improves upon the baseline model (row 1), highlighting the efficacy of learning with concepts.

**Concept Mining Strategy.** We propose extracting overlapped n-grams to form concepts, which can identify a wider range of concepts compared to previous works limited to specific scopes such as object tags [32] or verb-/adj-nouns [22]. To evaluate the benefits, we learn a baseline model on a restricted concept database by performing the proposed mining procedure but only considering noun phrases. As shown in the bottom row of Table 3, the results demonstrate that our approach (row 4) outperforms the baseline across all tasks, indicating that the proposed mining procedure is critical and can serve as a general method for exploring concepts in learning.

## 4.5 Analysis for Different Semantic Complexity

The proposed C3 model aligns concepts to enhance inter-modality semantic relationships, thereby improving its reasoning capabilities for complex semantics. To validate the claim, we analyze the models' performance under various levels of semantic complexity in the challenging NLVR$^2$ visual reasoning task. We define the semantic complexity of data from three perspectives, i.e., the constituency parsing tree, character length, and token length. The maximum depth of the constituency parsing tree is selected as the measure of semantic complexity since deeper trees generally indicate more intricate text structures, while longer text or word lengths suggest richer contexts. Fig. 3 consistently demonstrates that the C3 model's performance is negatively correlated with semantic complexity across all three definitions and experimental settings. Furthermore, our proposed matched concept prediction shows to be particularly effective for challenging instances, highlighting its practical value.

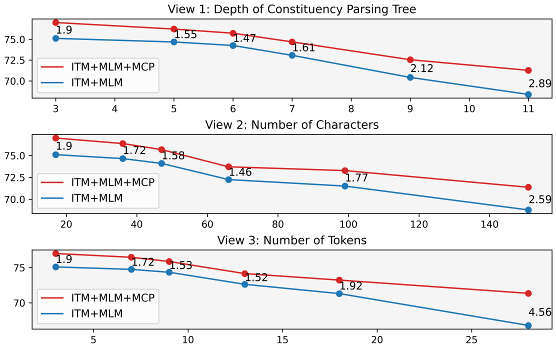


**Fig. 3:** The performance of NLVR$^2$ under varying levels of semantic complexity.

## 4.6 Visualization

To provide a clearer understanding, we visualize the attention maps from the final attention layer of the cross-modal encoder. To gain insights into the models' comprehension of challenging concepts, we randomly select concepts with 4- or 5-gram attributes from the validation dataset. Figure 4 displays these visualizations for both C3 and our baseline model, METER. Notably, the results illustrate that C3 exhibits the ability to focus on specific regions in accordance with the text fragments, while METER tends to distribute attention more evenly across the entire space. This difference may stem from the distribution gap between concept texts and captions. Concept-specific learning helps models better identify meaningful regions, improving their overall capacity.

Examples a & b, as well as examples c & d, demonstrate that C3 adapts its attention to different regions according to the input concepts. Furthermore, C3 outperforms METER in appropriately attending to regions for ambiguous phrases, as exemplified in examples e, f, and g. These visualizations suggest that C3 holds significant potential for tasks involving visual-linguistic grounding [23, 37, 59], opening up promising avenues for future research. Additional examples can be found in the appendix.

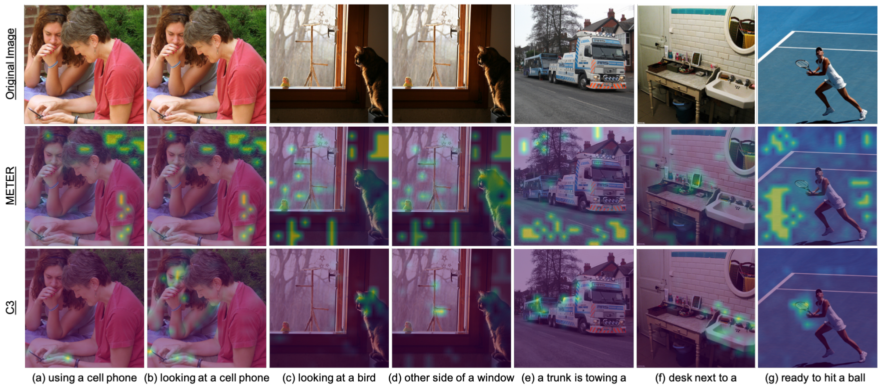


**Fig. 4:** Visualization of attention maps for abstract concepts randomly selected from examples containing 4- or 5-gram concepts, providing insight into the model's understanding of abstract semantic relationships. Examples a & b and examples c & d show that C3 can attend to different regions depending on the input concepts. Examples e, f, and g show that C3 can process ambiguous sentence fragments.

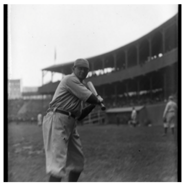
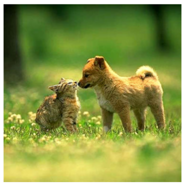
### 4.7 Examples of Concepts

In this section, we present the findings of our investigations into mined concepts using the proposed $n$-gram strategy and the baseline noun-phrase strategy, which are showcased in Table 4. Our analysis reveals that the $n$-gram approach provides a higher degree of diversity in captured concepts as compared to the baseline noun-phrase strategy. Specifically, in the first example (row 1), the $n$-gram approach identifies a wider range of concepts, including "to swing a baseball bat" and "while standing on top of," whereas the baseline strategy is more limited. Similarly, in the second example (row 2), the $n$-gram approach identifies the concept of "a grassy field," not captured by the baseline approach. The greater diversity of concepts captured through our $n$-gram approach enhances our model's ability to learn concept-level alignments without being confined to predefined categories or part-of-speech. These examples underscore the potential of the $n$-gram strategy as a powerful tool for mining concepts in the vision-language domain.

## 5 Conclusion

This paper presents Capture Concepts through Comparison (C3), a novel framework designed to enhance the core capabilities of vision-language (VL) models by

**Table 4:** Concept examples from COCO [4]. An example includes five captions for each image.

| Image | Captions | Concepts by n-gram mining | Concepts by noun phrase mining |
|---|---|---|---|
|  | (1) An old picture of a baseball player holding a baseball bat. (2) A black and white image depicting a man preparing to swing a baseball bat. (3) A man holding a bat while standing on top of a field. (4) An old fashioned picture shows a baseball batter in uniform. (5)A black and white picture of a baseball player. | black and white picture of, a black and white picture, old picture of a baseball, an old fashioned picture shows, a bat while standing on, holding a bat while standing, a black and white image, picture of a baseball player, to swing a baseball bat, baseball player holding a baseball, player holding a baseball bat, a man preparing to swing, man holding a bat while, man preparing to swing a, a man holding a bat, while standing on top of | a black and white picture, a black and white image, a baseball bat, an old picture, a bat |
|  | (1) In a grassy field is a puppy and a cat who are rubbing noses. (2) A small puppy standing next to a small kitten. (3) The puppy and kitten are in a field of grass. (4) A dog and a cat that are standing in the grass. (5) A kitten is touching noses with a puppy outside. | that are standing in the, standing next to a small, are standing in the grass, in a grassy field, field is a, are in a field of, in a field of grass, a grassy field is, a small kitten, a small puppy, a cat that are, a dog and a cat, and a cat that | a small kitten, a puppy, who, the puppy, noses |

strengthening the semantic alignment between the realms of vision and language. To begin, we introduce a mining procedure aimed at uncovering latent concepts within the database, all without the need for predefined scopes or external annotations. Building upon these mined concepts, we put forth two innovative learning objectives tailored for different architectural choices for the model, where inputs are formulated as triplets comprising a concept and images. These settings align with the psychological insight that concepts are shaped by shared characteristics. Finally, our comprehensive experiments conclusively demonstrate that C3 effectively boosts model capacity for multi-modality and enhances performance on downstream tasks, both in the context of continual pre-training and pre-training from scratch. These findings underscore the efficacy and versatility of our concept-centric learning approach. Building upon this research, we plan to extend our approach to more complicated foundation models and additional modalities in future endeavors.

## Acknowledgments

This work is partially supported by the National Science and Technology Council, Taiwan under Grants NSTC-112-2221-E-A49-059-MY3 and NSTC-112-2221-E-A49-094-MY3.

## References

1. Archer, E.J.: Chapter 3 - the psychological nature of concepts. In: Analyses of Concept Learning. Academic Press (1966)
2. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020)
3. Cascante-Bonilla, P., Shehada, K., Smith, J.S., Doveh, S., Kim, D., Panda, R., Varol, G., Oliva, A., Ordonez, V., Feris, R., Karlinsky, L.: Going beyond nouns with vision & language models using synthetic data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2023)
4. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325 (2015)
5. Chen, Y.C., Li, L., Yu, L., Kholy, A.E., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: Uniter: Universal image-text representation learning. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020)
6. Dou, Z.Y., Kamath, A., Gan, Z., Zhang, P., Wang, J., Li, L., Liu, Z., Liu, C., LeCun, Y., Peng, N., Gao, J., Wang, L.: Coarse-to-fine vision-language pre-training with fusion in the backbone. In: Advances in Neural Information Processing Systems (NeurIPS) (2022)
7. Dou, Z.Y., Xu, Y., Gan, Z., Wang, J., Wang, S., Wang, L., Zhu, C., Zhang, P., Yuan, L., Peng, N., Liu, Z., Zeng, M.: 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 (CVPR) (2022)
8. Doveh, S., Arbelle, A., Harary, S., Schwartz, E., Herzig, R., Giryes, R., Feris, R., Panda, R., Ullman, S., Karlinsky, L.: Teaching structured vision & language concepts to vision & language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
9. Duan, J., Chen, L., Tran, S., Yang, J., Xu, Y., Zeng, B., Chilimbi, T.: Multimodal alignment using representation codebook. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
10. Fang, Z., Wang, J., Hu, X., Liang, L., Gan, Z., Wang, L., Yang, Y., Liu, Z.: Injecting semantic concepts into end-to-end image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
11. Gan, Z., Chen, Y.C., Li, L., Zhu, C., Cheng, Y., Liu, J.: Large-scale adversarial training for vision-and-language representation learning. In: Advances in Neural Information Processing Systems (NeurIPS) (2020)
12. Gao, Y., Liu, J., Xu, Z., Zhang, J., Li, K., Ji, R., Shen, C.: Pyramidclip: Hierarchical feature alignment for vision-language model pretraining. In: Advances in Neural Information Processing Systems (NeurIPS) (2022)
13. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
14. Gutmann, M., Hyvärinen, A.: Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS). PMLR (2010)

15. Gutmann, M.U., Hyvärinen, A.: Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. Journal of Machine Learning Research (JMLR) (2012)
16. Hu, E.J., yelong shen, Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (ICLR) (2022)
17. Huang, Z., Zeng, Z., Huang, Y., Liu, B., Fu, D., Fu, J.: Seeing out of the box: End-to-end pre-training for vision-language representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
18. Huang, Z., Zeng, Z., Liu, B., Fu, D., Fu, J.: Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. arXiv preprint arXiv:2004.00849 (2020)
19. Jiang, C., Xu, H., Ye, W., Ye, Q., Li, C., Yan, M., Bi, B., Zhang, S., Zhang, J., Huang, F.: Copa: Efficient vision-language pre-training through collaborative object-and patch-text alignment. In: Proceedings of the 31th ACM International Conference on Multimedia (2023)
20. Järvelin, A., Järvelin, A., Järvelin, K.: s-grams: Defining generalized n-grams for information retrieval. Information Processing & Management (IPM) (2007)
21. Kamath, A., Singh, M., LeCun, Y., Synnaeve, G., Misra, I., Carion, N.: Mdetr-modulated detection for end-to-end multi-modal understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021)
22. Kamath, A., Clark, C., Gupta, T., Kolve, E., Hoiem, D., Kembhavi, A.: Webly supervised concept expansion for general purpose vision models. In: Proceedings of the European Conference on Computer Vision (ECCV) (2022)
23. Kazemzadeh, S., Ordonez, V., Matten, M., Berg, T.: ReferItGame: Referring to objects in photographs of natural scenes. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (2014)
24. Khan, Z., Vijay Kumar, B.G., Yu, X., Schulter, S., Chandraker, M., Fu, Y.: Single-stream multi-level alignment for vision-language pretraining. In: Proceedings of the European Conference on Computer Vision (ECCV) (2022)
25. Kim, W., Son, B., Kim, I.: Vilt: Vision-and-language transformer without convolution or region supervision. In: International Conference on Machine Learning (ICML) (2021)
26. Kong, L., de Masson d'Autume, C., Yu, L., Ling, W., Dai, Z., Yogatama, D.: A mutual information maximization perspective of language representation learning. In: International Conference on Learning Representations (ICLR) (2020)
27. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision (IJCV) (2017)
28. Li, D., Li, J., Li, H., Niebles, J.C., Hoi, S.C.: Align and prompt: Video-and-language pre-training with entity prompts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
29. Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H.: Align before fuse: Vision and language representation learning with momentum distillation. In: Advances in Neural Information Processing Systems (NeurIPS) (2021)
30. Li*, L.H., Zhang*, P., Zhang*, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., Chang, K.W., Gao, J.: Grounded language-image pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)

31. Li, W., Gao, C., Niu, G., Xiao, X., Liu, H., Liu, J., Wu, H., Wang, H.: UNIMO: Towards unified-modal understanding and generation via cross-modal contrastive learning. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (2021)
32. Li, X., Yin, X., Li, C., Hu, X., Zhang, P., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., Choi, Y., Gao, J.: Oscar: Object-semantics aligned pre-training for vision-language tasks. Proceedings of the European Conference on Computer Vision (ECCV) (2020)
33. Li, Y., Fan, J., Pan, Y., Yao, T., Lin, W., Mei, T.: Uni-eden: Universal encoder-decoder network by multi-granular vision-language pre-training. ACM Transactions on Multimedia Computing, Communications, and Applications (2022)
34. Li, Z., Fan, Z., Tou, H., Chen, J., Wei, Z., Huang, X.: Mvptr: Multi-level semantic alignment for vision-language pre-training via multi-stage learning. In: Proceedings of the 30th ACM International Conference on Multimedia (2022)
35. Liu, B., Rosenfeld, E., Ravikumar, P.K., Risteski, A.: Analyzing and improving the optimization landscape of noise-contrastive estimation. In: International Conference on Learning Representations (ICLR) (2022)
36. Liu, Y., Wu, C., Tseng, S.Y., Lal, V., He, X., Duan, N.: KD-VLP: Improving end-to-end vision-and-language pretraining with object knowledge distillation. In: Findings of the Association for Computational Linguistics: NAACL 2022 (2022)
37. Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A.L., Murphy, K.: Generation and comprehension of unambiguous object descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
38. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
39. Ordonez, V., Kulkarni, G., Berg, T.: Im2text: Describing images using 1 million captioned photographs. In: Advances in Neural Information Processing Systems (NeurIPS) (2011)
40. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2015)
41. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (ICML) (2021)
42. Rae, J.W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., Aslanides, J., Henderson, S., Ring, R., Young, S., et al.: Scaling language models: Methods, analysis & insights from training gopher. arXiv preprint arXiv:2112.11446 (2021)
43. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (2018)
44. Shen, S., Li, L.H., Tan, H., Bansal, M., Rohrbach, A., Chang, K.W., Yao, Z., Keutzer, K.: How much can CLIP benefit vision-and-language tasks? In: International Conference on Learning Representations (ICLR) (2022)
45. Singh, A., Hu, R., Goswami, V., Couairon, G., Galuba, W., Rohrbach, M., Kiela, D.: Flava: A foundational language and vision alignment model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)

46. Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., Dai, J.: Vl-bert: Pre-training of generic visual-linguistic representations. In: International Conference on Learning Representations (ICLR) (2020)
47. Su, W., Zhu, X., Tao, C., Lu, L., Li, B., Huang, G., Qiao, Y., Wang, X., Zhou, J., Dai, J.: Towards all-in-one pre-training via maximizing multi-modal mutual information. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
48. Suhr, A., Zhou, S., Zhang, A., Zhang, I., Bai, H., Artzi, Y.: A corpus for reasoning about natural language grounded in photographs. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (2019)
49. Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O.K., Singhal, S., Som, S., Wei, F.: Image as a foreign language: BEiT pretraining for vision and vision-language tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
50. Wang, X., McCallum, A., Wei, X.: Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In: Seventh IEEE International Conference on Data Mining (ICDM 2007) (2007)
51. Wang, Z., Yu, J., Yu, A.W., Dai, Z., Tsvetkov, Y., Cao, Y.: SimVLM: Simple visual language model pretraining with weak supervision. In: International Conference on Learning Representations (ICLR) (2022)
52. Xie, N., Lai, F., Doran, D., Kadav, A.: Visual entailment: A novel task for fine-grained image understanding. arXiv preprint arXiv:1901.06706 (2019)
53. Xu, H., Yan, M., Li, C., Bi, B., Huang, S., Xiao, W., Huang, F.: E2E-VLP: End-to-end vision-language pre-training 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 (Volume 1: Long Papers) (2021)
54. Xue, H., Huang, Y., Liu, B., Peng, H., Fu, J., Li, H., Luo, J.: Probing inter-modality: Visual parsing with self-attention for vision-and-language pre-training. In: Advances in Neural Information Processing Systems (NeurIPS) (2021)
55. Yan, M., Xu, H., Li, C., Bi, B., Tian, J., Gui, M., Wang, W.: Grid-vlp: Revisiting grid features for vision-language pre-training. arXiv preprint arXiv:2108.09479 (2021)
56. Yang, J., Duan, J., Tran, S., Xu, Y., Chanda, S., Chen, L., Zeng, B., Chilimbi, T., Huang, J.: Vision-language pre-training with triple contrastive learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
57. Yao, L., Han, J., Liang, X., Xu, D., Zhang, W., Li, Z., Xu, H.: Detclipv2: Scalable open-vocabulary object detection pre-training via word-region alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
58. Yao, L., Huang, R., Hou, L., Lu, G., Niu, M., Xu, H., Liang, X., Li, Z., Jiang, X., Xu, C.: FILIP: Fine-grained interactive language-image pre-training. In: International Conference on Learning Representations (ICLR) (2022)
59. Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L.: Modeling context in referring expressions. In: Proceedings of the European Conference on Computer Vision (ECCV) (2016)
60. Zeng, Y., Zhang, X., Li, H.: Multi-grained vision language pre-training: Aligning texts with visual concepts. In: International Conference on Machine Learning (ICML) (2022)

61. Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., Gao, J.: Vinvl: Revisiting visual representations in vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
62. Zhao, T., Zhang, T., Zhu, M., Shen, H., Lee, K., Lu, X., Yin, J.: An explainable toolbox for evaluating pre-trained vision-language models. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (2022)
63. Zhou, M., Yu, L., Singh, A., Wang, M., Yu, Z., Zhang, N.: Unsupervised vision-and-language pre-training via retrieval-based multi-granular alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)