PiTL: Cross-modal Retrieval with Weakly-supervised Vision-language Pre-training via Prompting

Zixin Guo  
Aalto University  
Espoo, Finland  
zixin.guo@aalto.fi

Tzu-Jui Julius Wang  
Aalto University  
Espoo, Finland  
tzu-jui.wang@aalto.fi

Selen Pehlivan  
Aalto University  
Espoo, Finland  
selen.pehlivantort@aalto.fi

Abduljalil Radman  
Aalto University  
Espoo, Finland  
abduljalil.saif@aalto.fi

Jorma Laaksonen  
Aalto University  
Espoo, Finland  
jorma.laaksonen@aalto.fi

ABSTRACT
Vision-language (VL) Pre-training (VLP) has shown to well generalize VL models over a wide range of VL downstream tasks, especially for cross-modal retrieval. However, it hinges on a huge amount of image-text pairs, which requires tedious and costly curation. On the contrary, weakly-supervised VLP (W-VLP) [33] explores means with object tags generated by a pre-trained object detector (OD) from images. Yet, they still require paired information, i.e. images and object-level annotations, as supervision to train an OD.

To further reduce the amount of supervision, we propose Prompts-in-The-Loop (PiTL) that prompts knowledge from large language models (LLMs) to describe images. Concretely, given a category label of an image, e.g., "refinery," the knowledge, e.g. "a refinery could be seen with large storage tanks, pipework, and ...", extracted by LLMs is used as the language counterpart. The knowledge supplements, e.g. the common relations among entities most likely appearing in a scene. We create IN14K, a new VL dataset of 9M images and 1M descriptions of 14K categories from ImageNet21K [8] with PiTL. Empirically, the VL models pre-trained with PiTL-generated pairs are strongly favored over other W-VLP works on image-to-text (I2T) and text-to-image (T2I) retrieval tasks, with less supervision. The results reveal the effectiveness of PiTL-generated pairs for VLP.

CCS CONCEPTS
• Information systems → Retrieval models and ranking.

KEYWORDS
Vision-language Retrieval; Pre-training; Knowledge Prompting

ACM Reference Format:

1 INTRODUCTION
Vision-language (VL) models have been advancing rapidly with the introduction of various Vision-language Pre-training (VLP) methods. The models for VLP can adapt to various downstream tasks involving VL modalities, such as cross-modal retrieval [7, 13, 14, 24, 28], visual question answering [4, 11, 14, 22, 38], visual captioning [12, 16, 17, 20, 21], etc. The secret recipe of a VLP model
VLP without an OD. PiTL capitalizes on image-level, i.e. a category label per image, instead of object-level annotations, e.g. the models excel better in downstream tasks with more pre-training data [1, 13, 15–18, 21, 28, 30, 34, 35, 37].

1.1 Weakly-supervised VLP
While the success of VLP methods relying on huge amounts of image-text annotations has been proven, a less visited research path is emerging to pursue more data-efficient VLP. The data-efficiency of a VL model can be viewed from the amount of supervision. That is, would a VL model pre-trained with less image-text data remain as performant in downstream tasks? This question leads to works [6, 19, 32, 33, 39] in weakly-supervised VLP (W-VLP) that aims at not relying on image-text pairs from, e.g. SBU Captions [26] and Conceptual Captions (CC) [5, 29].

Without the aligned images and texts, these works instead resort to a pre-trained object detector (OD) that generates object tags, i.e. the visual entities detected in the given image. The paired images and the object tags offer weaker supervision than those from the image-sentence pairs, but are still effective as the cross-domain bridge. However, training an OD relies on object-level annotations, which is still a form of supervision [33]. This seems to deviate these VLP works from the fully unsupervised path, which aims to remove any kind of cross-modal supervision.

1.2 Towards Unsupervised VLP
The unsupervised VLP (U-VLP), which aims at learning a VL model without any supervision across modalities, remains a daunting challenge. As a step towards U-VLP, we introduce Prompts-in-The-Loop (PiTL) that generates highly effective image-text pairs for W-VLP without an OD. PiTL capitalizes on image-level, i.e. a category label per image, instead of object-level supervision from the object bounding boxes and the corresponding object categories. This leads to a much harder W-VLP setting since much underlying information about an entity could no longer be inferred, such as the common co-occurrence of the visual entities, e.g. a chair and a desk, in a scene, and the entity relations, e.g. a person usually sits on a bench.

Specifically, given images with category labels, e.g. a duck as shown in Fig. 1, we prompt large language models (LLMs), e.g. GPT-3 [3], to generate descriptions as the external knowledge of the category labels of the images. In fact, different prompts provide different focuses on each target category, e.g. one that emphasizes colors: "Describe the colors seen from a/an <category>？", and another that emphasizes relations with other entities: "Describe what could a/an <category> be seen with?". This encourages a VL model to associate all plausible visual traits, entities, actions, and scenes pertaining to the target categories.

The prompting paradigm is becoming trendy. Recent works [25, 27, 36] reveal that textual prompts generated by LLMs lead to significant improvement in zero-shot image classification with VLP models like CLIP [28]. For instance, prompting a VL model with the LLM-generated description, "Goldfish are small, orange fish with shiny scales.\(^1\) is more likely to find matches in the visual domain than the generic "A photo of a goldfish." Likewise, our work explores if the LLM-generated descriptions could be proved useful as well in the W-VLP setting.

Our contributions are summarized as follows. Firstly, we propose PiTL that generates an image-text dataset IN14K containing 9M images with 1M descriptions of 14K categories from the “Winter21” release of ImageNet-21K [8]. Second, trained with half of the samples in IN14K, our models are shown competitive to the state of the arts in image-to-text (I2T) and text-to-image (T2I) retrieval tasks. With full IN14K, our models significantly outperform them, e.g. on MSCOCO-5K [23] by 11% and 10%, respectively, on I2T and T2I. Moreover, our models are comparable with VLP models trained with 4M aligned image-text pairs from, e.g. CC3M and SBU Captions, etc. Lastly, PiTL does not only come with the least cross-modal supervision among W-VLP works, but also leads to a small gap between the W-VLP and VLP performances.

2 W-VLP WITH IMAGE-LEVEL SUPERVISION
VLP aims at learning VL alignments given a large number of image-sentence pairs. The methodology is concluded as (1) learning shared semantics around VL modalities, (2) learning cross-modal context, e.g. masked modeling [9], and (3) learning to explicitly match images and texts. With the same aim, the existing W-VLP methods leverage OD-generated tags to form VL pairs. What is usually neglected is the cost of pre-training such an OD, which usually requires 10+ object-level annotations to be effective. The proposed PiTL aims at relaxing the requirement of having an OD via prompting LLMs [3] to generate descriptions for the object categories.

2.1 Forming Image-text Pairs via Prompting
PiTL elicits knowledge about an object category from nine prompts of different perspectives with an LLM. Five descriptions are collected for each prompt. Table 1 summarizes the nine prompts and their focuses. Some of them are more visually-relevant (P1-6), some focus more on knowledge around the target category (P7-8), and some are more open-ended (P9). We study the effectiveness of the descriptions generated by each prompt later in Sec. 3.3.

Among PiTL-generated pairs, an image can be paired with different descriptions as long as they are of the same category. In pre-training, the positive pairs for the Image-Text Contrastive and Image-Text Matching losses (introduced later in Sec. 2.3) are drawn from the images and descriptions of the same categories, and the negative pairs from those of the different categories. In this setup, pre-training with PiTL-generated pairs encourages the VL models to learn cross-modal alignment at the category level, i.e. images of a target category aligned to a group of descriptions, instead of instance level, i.e. an image aligned with a description as in other VLP works. As such, a given image would be associated with the plausible categorical visual traits, entities, actions, and scenes through the VL models.

2.2 VL Model Architecture
Our model architecture follows a state-of-the-art VL model, ALBEF [18], which has a multi-modal encoder fusing the representations generated by visual and textual encoders. Indeed, any VL model

---

\(^1\)Some of the gathered image-text pairs may not be highly relevant as they are not validated by a human.
with image-text inputs could also be used instead, as proposing a new VL architecture is not the focus of this work. Specifically, given an image $V$ and its paired text description $T$, the vision encoder $g_V(\cdot)$ follows ViT [10] consisting of a 12-layer Transformer that generates the image embedding as $g_V(V) = (x_{CLS}^V, x_1^V, \ldots, x_n^V)$, where $x_{CLS}^V$ measures the dot-product similarity of an image-text pair $(V, T)$. $p^T_m(V)$ and $p^T_m(T)$ are image-to-text and text-to-image similarities, respectively. $\tau$ is a learnable temperature parameter and $M$ is the size of the queues storing the image and textual class embeddings. The ITC loss is then defined as

$$L_{ITC} = \frac{1}{2} \sum_{m=1}^{M} \sum_{i=1}^{M} [H(y^T_m, p^T_m(V)) + H(y^T_m, p^T_m(T))],$$

where $y^T_m$ and $y^T_m$ indicate whether the pair is matched or not. It is worth noting that $L_{ITC}$ encourages the model to retain the uni-modal semantics provided by the pre-trained weights of the vision and textual encoders, complementing $L_{ITC}$, $L_{ITM}$, and $L_{LM}$, all of which promote multi-modal alignments.

### 3 EXPERIMENTS

#### 3.1 Settings

Our vision encoder is instantiated by ViT [10] and initialized with DeiT [31] or BeiT-B/16 [2] pre-trained weights. The textual encoder is initialized with BERT-Base [9]. The proposed PiTL W-VLP model is pre-trained on three subsets IN1K, IN6K, and IN14K created from ImageNet21K. Note that IN6K and IN14K are created in this work. Specifically, IN14K contains IN6K which also contains IN1K samples. Each prompt, out of the nine shown in Table 1, generates five responses for a category. Individual prompts are created for multiple synonyms, e.g. snorkeling and snorkel_diving under the same category. Statistics of IN1K, IN6K, IN14K along with other datasets, e.g. CC3M, BookCorpus (BC) [40], and VL-Full [32] used by other VLP methods, are shown in Table 2 under the Pre-training Corpus column. We assess the pre-training quality on I2T and T2I with MSCOCO-5K and Flickr30K [26]. The retrieval models are evaluated on recall at rank K (R@K).

#### 3.2 Quantitative Results

Table 2 shows the main results of PiTL and the comparisons against the state-of-the-art VLP and W-VLP on the I2T and T2I tasks. **Effects of Initialization and Dataset Sizes.** We initialize the image encoder with weights pre-trained with no supervision (i.e. the self-supervised BeiT\(^\dagger\)) and with image-level supervision (i.e. ViT, BeiT\(^\dagger\), and BeiT\(*\)). The best performances are obtained with BeiT\(*\), whose weights contain the strongest visual semantics, on IN1K to IN14K. PiTL's results steadily improve with more images and descriptions. Compared to models without pre-training, BeiT\(^\dagger\) pre-trained on IN1K has larger improvements in R@1 than the other initializations, i.e. 8.6% for I2T and 10.8% for T2I R@1 on MSCOCO-5K.

**PiTL with BeiT-B/16\(^\dagger\) vs. W-VLP Works.** To purely assess the generated image-text pairs, models initialized with self-supervised BeiT-B/16\(^\dagger\) weights are mainly benchmarked. VLMixer starts out as a better model than PiTL when both are not pre-trained on any image-text pairs. However, once pre-trained, PiTL appears to be strongly competitive. For instance, PiTL pre-trained on IN1K outperforms VLMixer pre-trained on CC3M, across all the I2T metrics, in the same sentence. 15% masking probability is set. The MLM loss [9] is denoted as $L_{LM}$.

**Intra-Modal Contrastive (IMC)** aims to differentiate the semantics between the positive and negative pairs within the same modality [35], i.e. image-image and text-text pairs with similarities:

$$p^T_m(V) = \frac{e^{s(V,Y_m)/\tau}}{\sum_{i=1}^{M} e^{s(V,Y_i)/\tau}},$$

$$p^T_m(T) = \frac{e^{s(T,Y_m)/\tau}}{\sum_{i=1}^{M} e^{s(T,Y_i)/\tau}}.$$  

(5)

The IMC objective is defined as

$$L_{IMC} = \frac{1}{2} \sum_{V,T \in D} [H(y^H_{IT}(\cdot), y^H_{IT}(\cdot)) + H(y^T_{IT}(\cdot), y^T_{IT}(\cdot))],$$

(6)

where $y^H_{IT}$ and $y^T_{IT}$ indicate whether the pair is matched or not.

### 2.3 Pre-training Losses

Our PiTL VL-models are pre-trained with four losses [18, 35] that all contribute equally to the total loss $L$:

$$L = L_{ITC} + L_{ITM} + L_{LM} + L_{IMC},$$

(1)

where each objective is described as follows.

**Image-Text Contrastive (ITC)** aims to retain high and low similarities between the positive and negative image-text pairs, respectively. To obtain the ITC loss, one first calculates

$$p^T_m(V) = \frac{e^{s(V,Y_m)/\tau}}{\sum_{i=1}^{M} e^{s(V,Y_i)/\tau}}.$$  

(2)

where $s(V,T) = (x_{CLS}^V, x_{CLS}^T)$ measures the dot-product similarity of an image-text pair $(V, T)$. $p^T_m(V)$ and $p^T_m(T)$ are image-to-text and text-to-image similarities, respectively. $\tau$ is a learnable temperature parameter and $M$ is the size of the queues storing the image and textual class embeddings. The ITC loss is then defined as

$$L_{ITC} = \frac{1}{2} \sum_{m=1}^{M} \sum_{i=1}^{M} [H(y^T_m, p^T_m(V)) + H(y^T_m, p^T_m(T))].$$

(3)

**Image-Text Matching (ITM)** aims to predict whether an image-text pair is matched. The token embedding $x_{CLS}^V$ of the fusion encoder predicts the binary classification probability $p^T_{ITM}$. The ITM loss is defined as

$$L_{ITM} = \sum_{m=1}^{M} \sum_{i=1}^{M} [H(y^T_m, p^T_{ITM}(V, T))].$$

(4)

where $y^T_{ITM}$ is a binary vector indicating the matching pairs.

**Masked Language Modeling (MLM)** predicts the masked tokens in a sentence given an image and the unmasked textual tokens.

#### Table 1: LLM prompts and focuses for <category>, e.g. duck.

<table>
<thead>
<tr>
<th>Type</th>
<th>Prompts</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>Describe colors of a &lt;category&gt;</td>
</tr>
<tr>
<td>P2</td>
<td>Describe shapes of a &lt;category&gt;</td>
</tr>
<tr>
<td>P3</td>
<td>Describe textures of a &lt;category&gt;</td>
</tr>
<tr>
<td>P4</td>
<td>Describe visual appearances of a &lt;category&gt;</td>
</tr>
<tr>
<td>P5</td>
<td>Describe a &lt;category&gt; in a scene</td>
</tr>
<tr>
<td>P6</td>
<td>Describe what a &lt;category&gt; could be seen with</td>
</tr>
<tr>
<td>P7</td>
<td>Describe the places a &lt;category&gt; has been seen</td>
</tr>
<tr>
<td>P8</td>
<td>Describe the main activities of a &lt;category&gt;</td>
</tr>
<tr>
<td>P9</td>
<td>Describe what it is like to be a &lt;category&gt;</td>
</tr>
</tbody>
</table>

- Colors
- Shapes
- Textures
- Visual appearances
- Scene
- Associated things
- Places
- Activities
- Feelings
Table 2: Results of image-text retrieval tasks on MSCOCO-5K and Flickr30K test splits. * denotes the results reported by [33]. † denotes BEiT [2] self-supervised pre-trained on the images of ImageNet21K. ‡ denotes BEiT intermediate fine-tuned on the images and labels of ImageNet21K. * denotes BEiT† intermediate fine-tuned on the images and labels of ImageNet21K.

<table>
<thead>
<tr>
<th>VLP</th>
<th>Supervision</th>
<th>Pre-training Corpus</th>
<th># of Images</th>
<th># of Texts</th>
<th>MSCOCO-5K</th>
<th>Flickr30K</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>I2T</td>
<td>I2T</td>
</tr>
<tr>
<td>VIT-B/16</td>
<td>[14]</td>
<td>image-level annotations + aligned VL pairs</td>
<td>VIT-B/16</td>
<td>4M</td>
<td>4M</td>
<td>61.5</td>
</tr>
<tr>
<td>ALBEF [16]</td>
<td>image-level annotations</td>
<td>VIT-B/16 (DeiT)</td>
<td>4M</td>
<td>4M</td>
<td>73.1</td>
<td>91.4</td>
</tr>
<tr>
<td>TCL [33]</td>
<td>image-level annotations</td>
<td>VIT-B/16 (DeiT)</td>
<td>4M</td>
<td>4M</td>
<td>73.6</td>
<td>92.8</td>
</tr>
<tr>
<td>W-VLP</td>
<td>image-level annotations</td>
<td>BEiT-B/16</td>
<td>4M</td>
<td>4M</td>
<td>62.2</td>
<td>86.3</td>
</tr>
<tr>
<td>U/VisualBERT [19]</td>
<td>image-level + object-level annotations</td>
<td>BERT-Base</td>
<td>CC: 3M</td>
<td>CC: 3M</td>
<td>62.8</td>
<td>86.3</td>
</tr>
<tr>
<td>WPH [33]</td>
<td>image-level + object-level annotations</td>
<td>BERT-Base</td>
<td>CC: 3M</td>
<td>CC: 3M</td>
<td>57.4</td>
<td>84.0</td>
</tr>
<tr>
<td>E2E-UVL [6]</td>
<td>image-level + object-level annotations</td>
<td>BERT-Base</td>
<td>CC: 3M</td>
<td>CC: 3M</td>
<td>62.7</td>
<td>84.1</td>
</tr>
<tr>
<td>ALBEF [16]</td>
<td>image-level annotations</td>
<td>BERT-Base</td>
<td>CC: 3M</td>
<td>CC: 3M</td>
<td>62.2</td>
<td>86.3</td>
</tr>
<tr>
<td>PiTL</td>
<td>image-level annotations</td>
<td>BERT-Base</td>
<td>CC: 3M</td>
<td>CC: 3M</td>
<td>62.8</td>
<td>86.3</td>
</tr>
</tbody>
</table>

Table 3: Effects of each type of prompt on Table 1 on retrieval tasks. VA is short for visual appearances.

<table>
<thead>
<tr>
<th>Types</th>
<th>R@1</th>
<th>R@5</th>
<th>R@10</th>
<th>AvgR</th>
<th>R@1</th>
<th>R@5</th>
<th>R@10</th>
<th>AvgR</th>
<th>Overall</th>
<th>AvgR</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1: captions</td>
<td>83.4</td>
<td>97.7</td>
<td>98.9</td>
<td>94.0</td>
<td>71.9</td>
<td>91.6</td>
<td>94.9</td>
<td>86.0</td>
<td>90.0</td>
<td></td>
</tr>
<tr>
<td>P2: shapes</td>
<td>85.0</td>
<td>97.7</td>
<td>98.9</td>
<td>94.0</td>
<td>71.9</td>
<td>91.6</td>
<td>94.9</td>
<td>86.0</td>
<td>90.0</td>
<td></td>
</tr>
<tr>
<td>P3: textures</td>
<td>85.0</td>
<td>97.7</td>
<td>98.9</td>
<td>94.0</td>
<td>71.9</td>
<td>91.6</td>
<td>94.9</td>
<td>86.0</td>
<td>90.0</td>
<td></td>
</tr>
<tr>
<td>P4: summarized VA</td>
<td>85.0</td>
<td>97.7</td>
<td>98.9</td>
<td>94.0</td>
<td>71.9</td>
<td>91.6</td>
<td>94.9</td>
<td>86.0</td>
<td>90.0</td>
<td></td>
</tr>
<tr>
<td>P5: scenes</td>
<td>85.0</td>
<td>97.7</td>
<td>98.9</td>
<td>94.0</td>
<td>71.9</td>
<td>91.6</td>
<td>94.9</td>
<td>86.0</td>
<td>90.0</td>
<td></td>
</tr>
<tr>
<td>P6: relations</td>
<td>85.0</td>
<td>97.7</td>
<td>98.9</td>
<td>94.0</td>
<td>71.9</td>
<td>91.6</td>
<td>94.9</td>
<td>86.0</td>
<td>90.0</td>
<td></td>
</tr>
<tr>
<td>P7: places</td>
<td>85.0</td>
<td>97.7</td>
<td>98.9</td>
<td>94.0</td>
<td>71.9</td>
<td>91.6</td>
<td>94.9</td>
<td>86.0</td>
<td>90.0</td>
<td></td>
</tr>
<tr>
<td>P8: activities</td>
<td>85.0</td>
<td>97.7</td>
<td>98.9</td>
<td>94.0</td>
<td>71.9</td>
<td>91.6</td>
<td>94.9</td>
<td>86.0</td>
<td>90.0</td>
<td></td>
</tr>
<tr>
<td>P9: first-person</td>
<td>85.0</td>
<td>97.7</td>
<td>98.9</td>
<td>94.0</td>
<td>71.9</td>
<td>91.6</td>
<td>94.9</td>
<td>86.0</td>
<td>90.0</td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>85.0</td>
<td>97.7</td>
<td>98.9</td>
<td>94.0</td>
<td>71.9</td>
<td>91.6</td>
<td>94.9</td>
<td>86.0</td>
<td>90.0</td>
<td></td>
</tr>
</tbody>
</table>

3.3 Ablation Studies
We study the retrieval performances on Flickr30K on the types of prompts (P1–P9 in Table 1) and the effect of the paired images and descriptions generated by PiTL.

On Effects of Prompts P1–P9. We dissect the effect contributed by prompts P1–9 to the retrieval performance. We pre-train models initialized with BEiT-B/16★ on descriptions from each prompt. Table 3 shows that, with all the descriptions available, the models achieve the best R@1 scores for both retrieval tasks. On I2T, P9 results in the best average recall over R@1,5,10. We reckon that P9 is a more open-ended question than the others, which focus specifically on appearances, relations, etc. The open-endedness could result in more diverse descriptions and hence better I2T result. On T2I, P1–4 usually yield better T2I results than the other prompts. This seems to be expected since P1–4 are more visually relevant.

On the Effect of Increased Uni-modal Shuffled Sources. To study whether the improvements actually come from the proposed prompt-based weak supervision, rather than from the pure increase in the number of images and texts, we pre-train the models (initialized with BEiT-B/16★ on IN1K and IN6K with shuffled images and descriptions. Indeed, on Flickr30K, the models degrade in R@1 from 60.7 to 56.1 for I2T, and from 48.2 to 42.7 for T2I, when pre-trained on shuffled IN1K and IN6K, respectively. This concludes the effectiveness of PiTL-generated pairs for VLP.

4 CONCLUSION
In this work, we proposed Prompts-in-The-Loop (PiTL), a weakly-supervised method to pre-train VL-models for cross-modal retrieval tasks. Without object-level supervision based on OD, our PiTL comes with image-level supervision from images and their categories described by large language models. Retrieval results on the cross-modal datasets demonstrated the effectiveness of PiTL.

ACKNOWLEDGMENTS
This work is supported by the Academy of Finland in project 345791. We acknowledge the LUMI supercomputer, owned by the EuroHPC Joint Undertaking, hosted by CSC and the LUMI consortium.
REFERENCES


