Knowledge Transfer from Interaction Learning

Supplementary Material

8. Vision-Language Interaction Extraction

In this section, we provide a detailed description of the Vision-Language Interaction Extraction process, as outlined in Section 4.2.

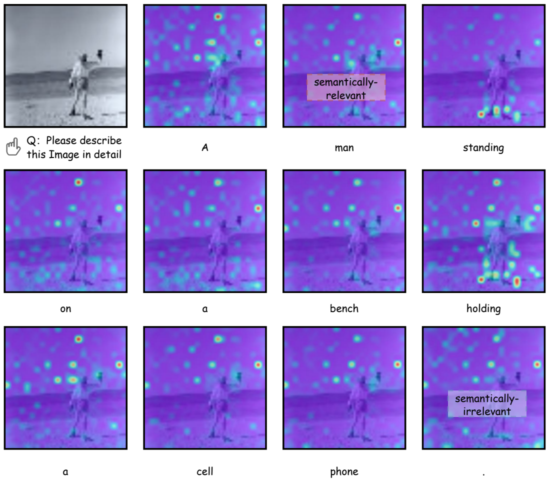


Figure 7. **Visual-Semantic Correlation in VLM Responses.** The outputs from Vision-Language Models (VLMs) are strongly associated with visual semantics. However, certain visual concepts generate high responses to semantically-irrelevant texts, such as punctuation marks, making it challenging to extract visual concepts directly from complex answers.

Given our use of Visual Question Answering (VQA) for interaction extraction, we first confirmed that visual concepts can effectively focus on relevant text during the VQA process, as demonstrated in Figure 7. While VLMs exhibit strong visual-semantic correlations, we observed that visual concepts also respond to semantically-irrelevant texts, such as punctuation marks, complicating the direct extraction of visual concepts from complex answers. Therefore, developing a robust pipeline for interaction extraction via VQA is essential. We compared the concepts extracted using various prompt settings:

1. **Prompt:** "If you are doing an image classification task, what is the foreground and what is the background? The answer format is as follows: {'foreground':{}, 'background':{}}. Please choose the foreground word from the list below:". We supplied a class vocabulary for LLaVA to select from, based on the categories in Tiny-ImageNet. We subtracted the background interaction C_{back} from the foreground interaction C_{fore} to align the concepts with human cognitive processes, and set the minimum value to 0 and normalized the interaction scores to a range of 0 to 1, as described in Section 3.2.
2. **Prompt:** "If you are doing an image classification task, What is the object in the picture? Answer the question using a single word or phrase." We directly take the interaction of the answer as our final interaction C_{VLM}.
3. **Prompt:** "Please describe the object in the picture in detail." Since there are several works in the sentence, we take the average of concepts as our final interaction C_{VLM}.

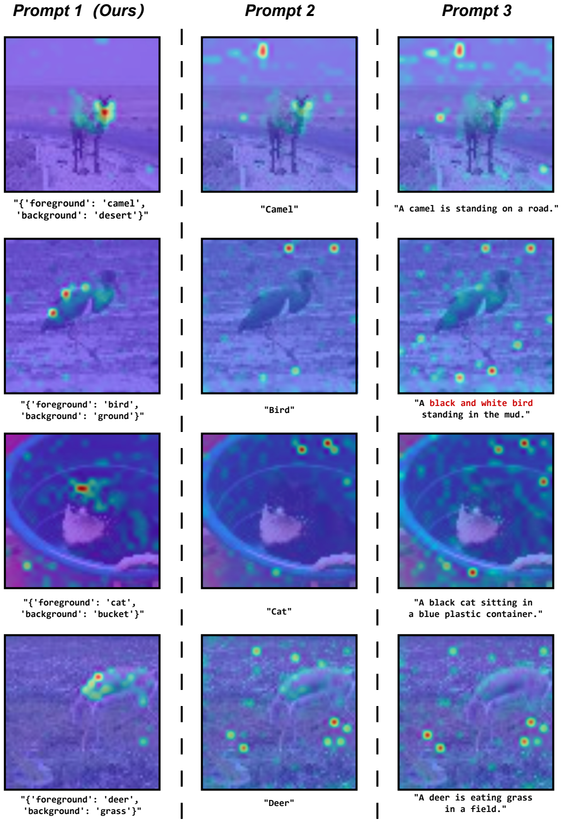


Figure 8. **Impact of Prompts on Concept Extraction.** Prompt 1 effectively focuses on the instance itself. However, Prompt 2, which instructs VLMs to output target content as single words, still leads to concepts emphasizing the background rather than the instance. Meanwhile, the detailed description approach of Prompt 3 results in concepts being distributed across the entire image and occasionally leads to incorrect responses *(highlighted in red)*.

The results, illustrated in Figure 8, demonstrate that different prompts elicit distinct concepts. Prompt 1 successfully emphasizes the object itself, while Prompt 2 exhibits the same issue as existing Visual Feature Models (VFMs), where concepts tend to focus on the background rather than the object. In contrast, Prompt 3's detailed description approach disperses concepts across the image, sometimes leading to inaccurate responses.

Table 8. Comparison of different Prompts

Model	Prompt	Top-1 Acc.(%)
ViT-S/16	/	79.94
I-ViT-S/16	Prompt 3	80.59
I-ViT-S/16	Prompt 2	81.26
I-ViT-S/16	Prompt 1	**81.52**

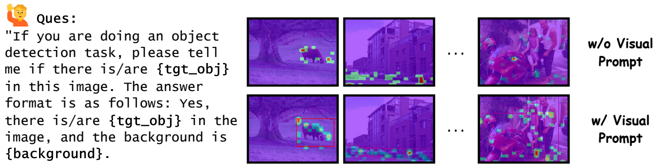
The impact of different prompts on VFM performance is compared in Table 8. Generally, incorporating concepts from VLMs enhances performance. However, the extent of improvement correlates with the prompts' ability to focus on the object. A stronger focus on the object indicates a more precise cognitive process, leading to superior outcomes.

For dense prediction tasks requiring multi-objective awareness, we synergistically combine **Language Prompts** and **Visual Prompts** to guide the VLM's focus. The language prompt is structured as:

> *"If you are doing an object detection task, please tell me if there is/are {tgt_obj} in this image. The answer format is as follows: Yes, there is/are {tgt_obj} in the image, and the background is {background}."*

where {tgt_obj} and {background} are dynamically replaced with target objects (e.g., *dog, car*) and contextual attributes (e.g., *grass, urban*) from task-specific annotations. This interrogative template forces the VLM to explicitly verify each object's presence and environmental context.

Concurrently, **Visual Prompts** are implemented by overlaying ground-truth bounding boxes on input images (Figure 9), spatially constraining the VLM's attention to instance regions. These boxes act as positional anchors during cross-modal interaction computation, reducing distraction from cluttered backgrounds.



Ques:
"If you are doing an object detection task, please tell me if there is/are {tgt_obj} in this image. The answer format is as follows: Yes, there is/are {tgt_obj} in the image, and the background is {background}.

Figure 9. **Interaction Extraction for Dense Prediction.** The process uses {**tgt_obj**} to aggregate concepts.

Table 9. Impact of Visual Prompts on COCO val2017

Configuration	**mAP**
Language Prompt Only	41.6
Language + Visual Prompts	43.6 (+2.0)

As evidenced by Table 9, integrating visual prompts yields gains: +2.0 mAP, establishing a closed-loop feedback between linguistic verification and visual grounding.

9. Human evaluation

To ensure the reliability and objectivity of our evaluation, we employed a double-blind assessment methodology. All participants were senior researchers with recognized expertise in the field. During the annotation process, participants were blinded to the annotations of their peers as well as to the results of C_{VLM}, C_{AGT}, and C_{VFM}. However, they were provided with the corresponding image labels to maintain objectivity. Participants were tasked with annotating two categories: (1) image tokens that directly determine the label with high confidence (1.0), and (2) tokens that indirectly influence the label with low confidence (0.5), such as background regions. 20 participants provided $\sim 1k$ annotated results, which were used to assess the similarity between different concepts and human annotations. Examples of the evaluation process are illustrated in Figure 10, and the findings align with those presented in Section 5.2.3.

10. Concept weights on different layers

Although we incorporate additional C_{VLM} supervision into every layer of the Transformer in I-ViT, the model's preference for C_{VFM} and C_{VLM} evolves with increasing model depth, as shown in Figure 11. We posit that the primary reason for this phenomenon is the rapid acquisition of task-related concepts in the initial layers, followed by their refinement in the deeper layers through the integration of VLM concepts. This observation suggests the following:

1. The Gated Control Network (GCN) effectively modulates the weights between different concepts, preventing over-reliance on any single interaction and ensuring balanced consideration.
2. Different concepts are complementary rather than interchangeable.

By effectively leveraging concepts from various sources, the model achieves improved generalization and robustness.

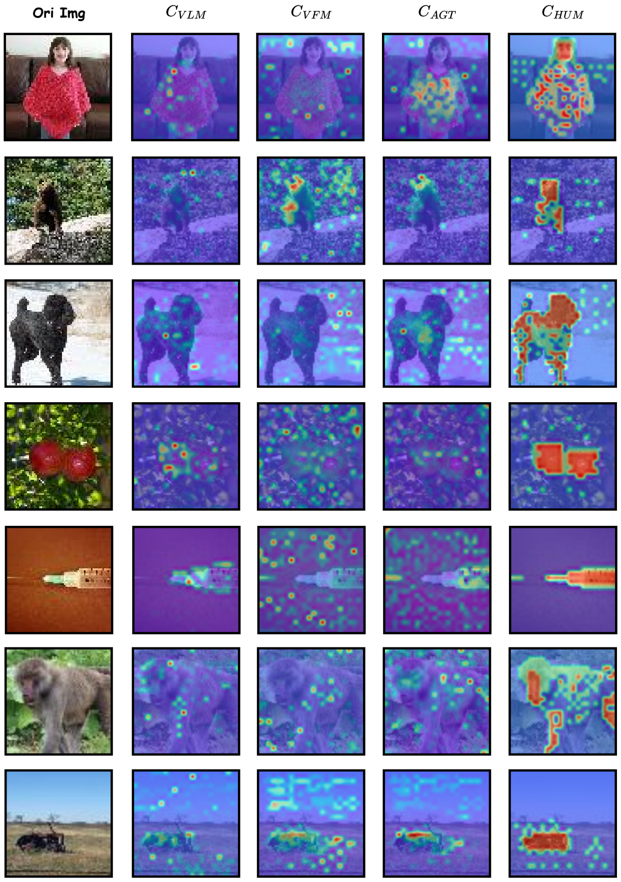


Figure 10. **Examples of Human Evaluation.** The figure illustrates the annotation process, where participants labeled image tokens based on their direct or indirect influence on the image label.

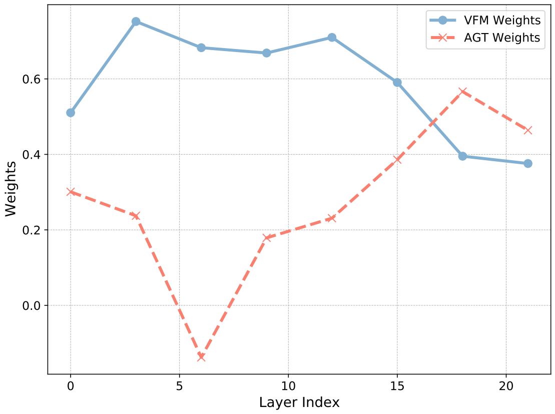


Figure 11. **Weights of Different Concepts Across Layers.** The model rapidly captures task-related concepts in early layers and further refines them in later layers using concepts from the VLM.