



The diagram you provided outlines two fundamental VTL components:

1. **Vision Representation Learning:** This encompasses a collection of techniques that empower computers to process and encode visual data extracted from images.
2. **Generative Learning:** In this stage, the computer leverages the encoded visual information to produce textual descriptions.

The lower portion of the image depicts a practical example of this concept. Here, a large language model (LLM) is used to craft a romantic message inspired by an image of a sunset.

Here's a more technical breakdown of the process:

1. **Bootstrapping Pre-trained Image Models:** A pre-trained image recognition model furnishes the groundwork by encoding the sunset image into a machine-interpretable format.
2. **Q-Former Encoder:** This component functions as a query formulator. It accepts a question as input (in this instance, the question pertains to composing a romantic message) and generates a refined query that can be used to search the encoded image data.
3. **Text Generation:** Finally, the LLM takes both the refined query and the encoded image data as inputs and generates a textual description, which manifests as the romantic message “Love is like a sunset, it’s hard to see it coming but when it does it’s so beautiful.”