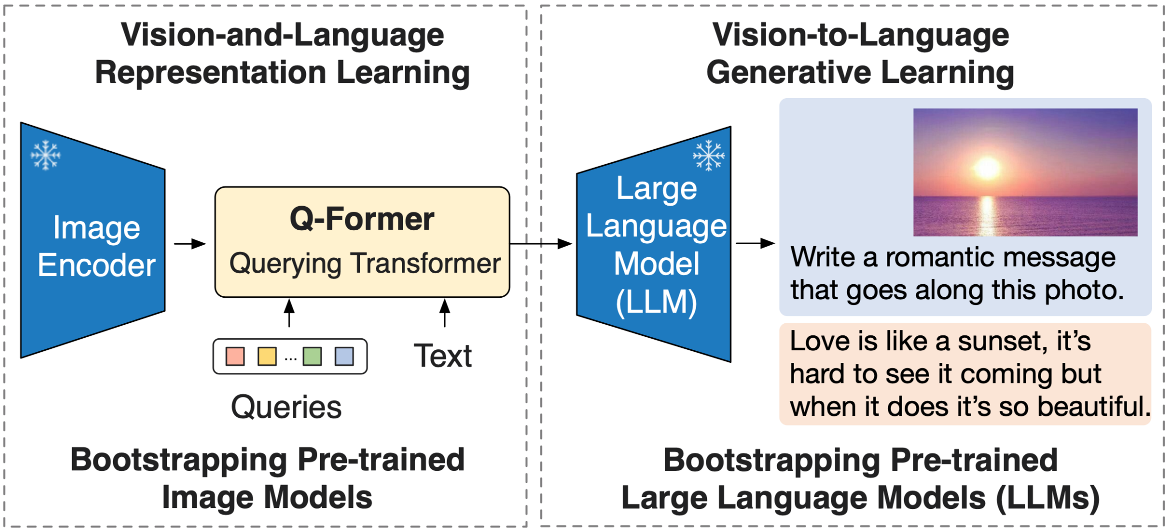
MyVLM: Personalizing VLMs for User-Specific Queries

Preliminaries

BILP-2 （Q-Former）

3 components:

1. a pretrained ViT-L/14 vision encoder
2. a pretrained language model
3. a trainable Querying Transformer (Q-Former) model tasked with bridging the vision-language modality gap.



Vision-and-Language Representation Learning
Image Encoder
Q-Former
Querying Transformer
Queries
Text
Bootstrapping Pre-trained Image Models
Vision-to-Language Generative Learning
Large Language Model (LLM)
Write a romantic message that goes along this photo.
Love is like a sunset, it's hard to see it coming but when it does it's so beautiful.
Bootstrapping Pre-trained Large Language Models (LLMs)

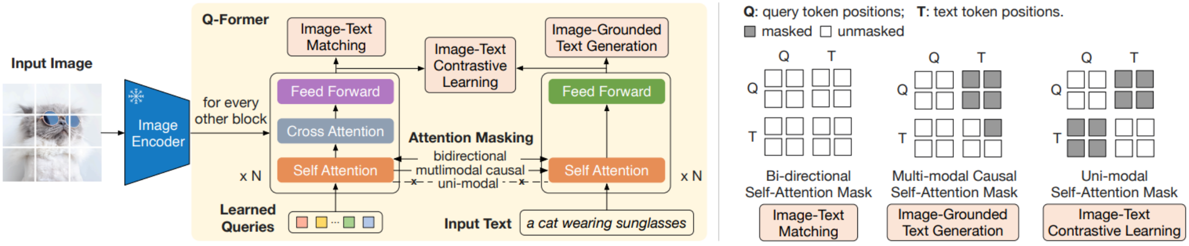


Figure 2. (**Left**) Model architecture of Q-Former and BLIP-2's first-stage vision-language representation learning objectives. We jointly optimize three objectives which enforce the queries (a set of learnable embeddings) to extract visual representation most relevant to the text. (**Right**) The self-attention masking strategy for each objective to control query-text interaction.

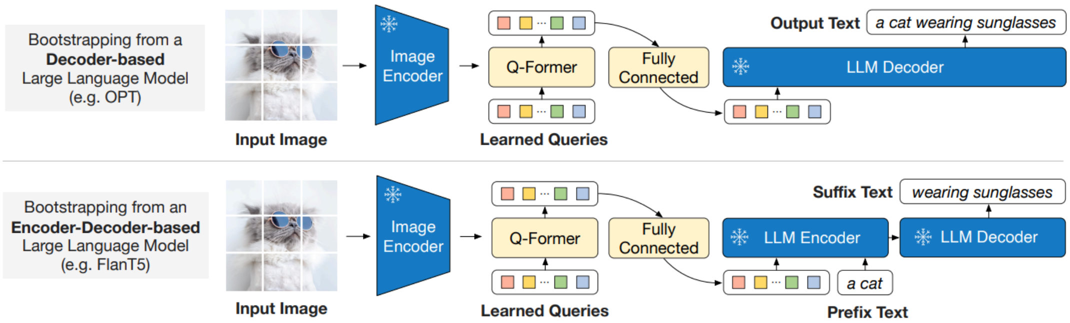


Figure 3. BLIP-2's second-stage vision-to-language generative pre-training, which bootstraps from frozen large language models (LLMs). (**Top**) Bootstrapping a decoder-based LLM (e.g. OPT). (**Bottom**) Bootstrapping an encoder-decoder-based LLM (e.g. FlanT5). The fully-connected layer adapts from the output dimension of the Q-Former to the input dimension of the chosen LLM.

LLaVA: Large Language and Vision Assistant

Architecture

3 components:

1. a pre-trained CLIP visual encoder ViT-L/14
2. a simple linear layer W to connect image features into the word embedding space
3. a LLaMA

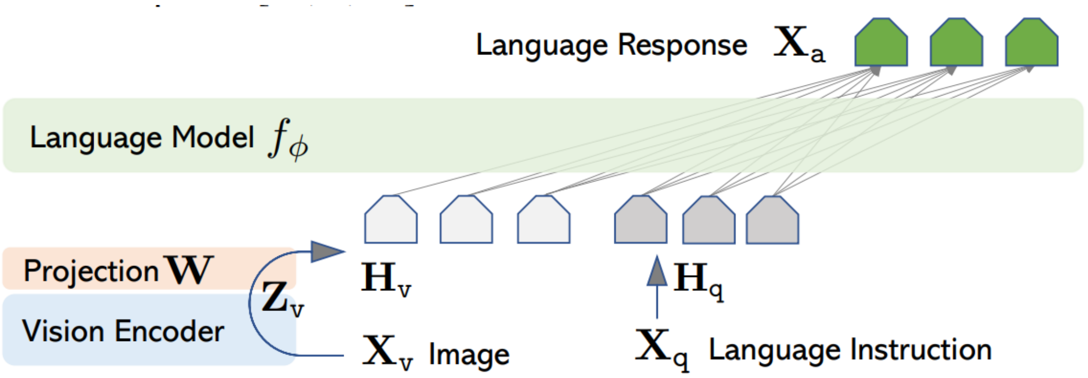


Figure 1: LLaVA network architecture.

GPT-assisted Visual Instruction Data Generation

[TODO]

Abstract

Recent large-scale vision-language models (VLMs) have demonstrated remarkable capabilities in understanding and generating textual descriptions for visual content. However, these models lack an understanding of user-specific concepts. In this work, we take a first step toward the personalization of VLMs, enabling them to learn and reason over user-provided concepts. For example, we explore whether these models can learn to recognize you in an image and communicate what you are doing, tailoring the model to reflect your personal experiences and relationships. To effectively recognize a variety of user-specific concepts, we augment the VLM with external concept heads that function as toggles for the model, enabling the VLM to identify the presence of specific target concepts in a given image. Having recognized the concept, we learn a new concept embedding in the intermediate feature space of the VLM. This embedding is tasked with guiding the language model to naturally integrate the target concept in its generated response. We apply our technique to BLIP-2 and LLaVA for personalized image captioning and further show its applicability for personalized visual question-answering. Our experiments demonstrate our ability to generalize to unseen images of learned concepts while preserving the model behavior on unrelated inputs.

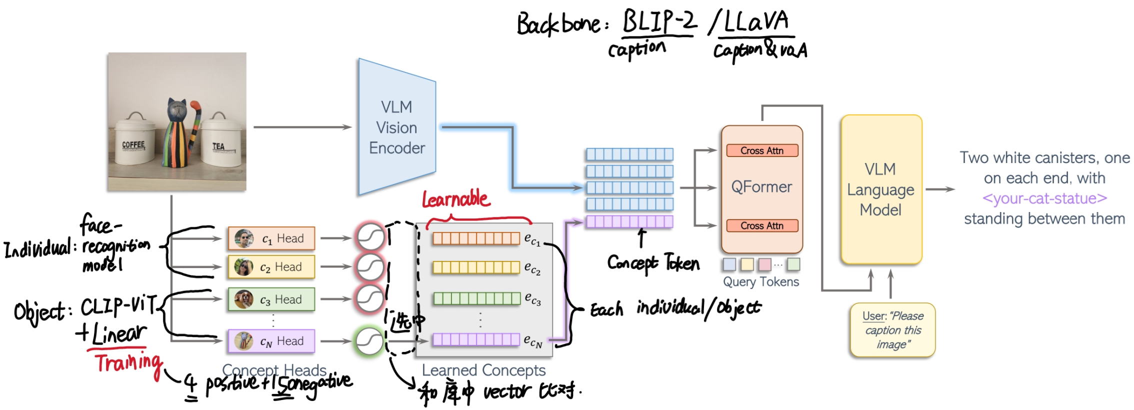


Figure 2. **MyVLM overview**, applied over BLIP-2. Given an input image, we pass it through the frozen vision encoder of the VLM. In parallel, we pass the image through a set of learned *concept heads*, each tasked with recognizing a single user-specific concept. We append the *concept embedding* of the identified concept to the extracted vision features. These features are then passed to the Q-Former via a set of cross-attention layers to extract relevant information from the image features and concept embedding. Given the Q-Former outputs and language instruction, the frozen LLM outputs a response incorporating the concept identifier while remaining aligned with the input.

Our technique is comprised of two key stages: first *recognizing* the concept within the given scene, and then *communicating* information about the concept to the language

trained CLIP model [29, 65]. To generate personaliz puts tailored to specific individuals, we utilize a pre face recognition network [24, 25] as an additional concept