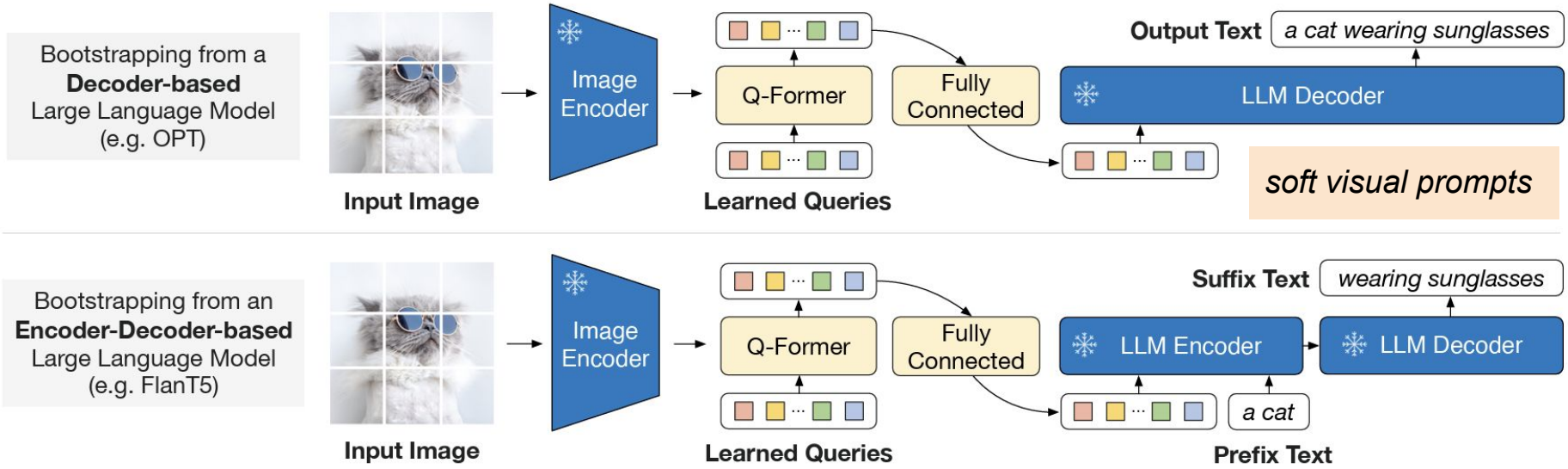
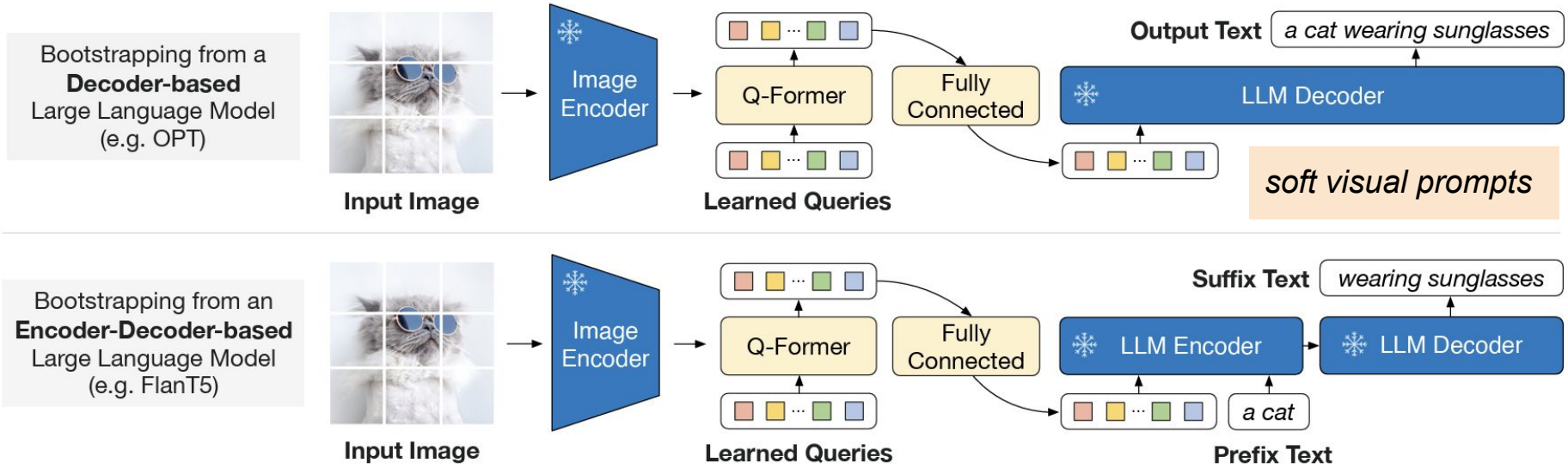
Bootstrap Vision-to-Language Generative Learning from a Frozen LLM



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.***

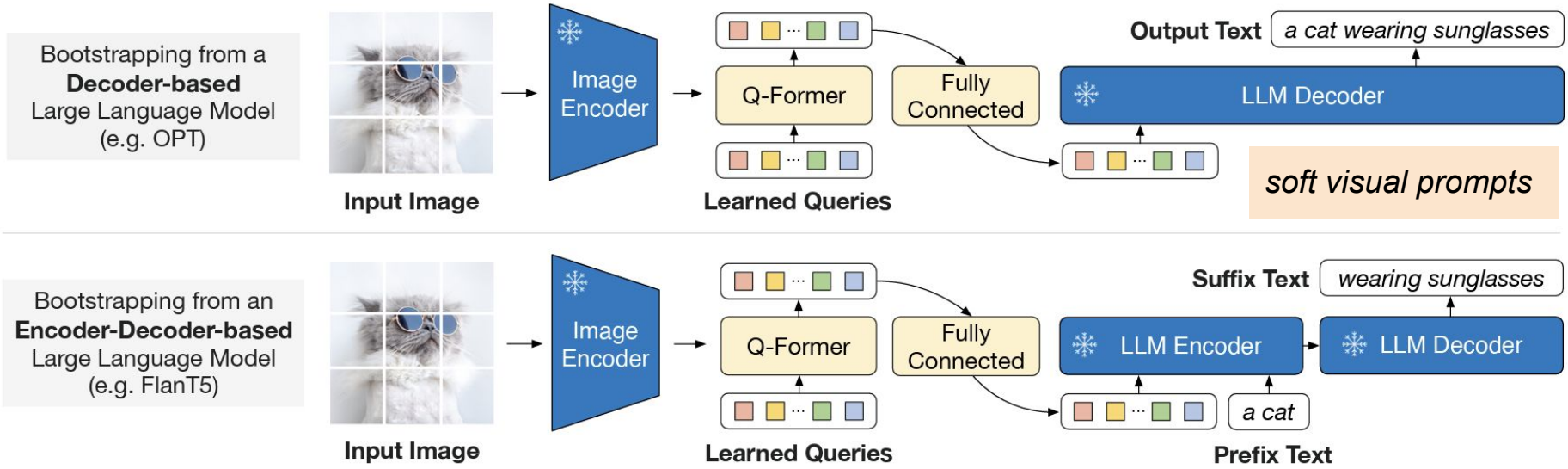
Bootstrap Vision-to-Language Generative Learning from a Frozen LLM



The projected query embeddings are then prepended to the input text embeddings. They function as ***soft visual prompts*** that condition the LLM on visual representation extracted by the Q-Former.

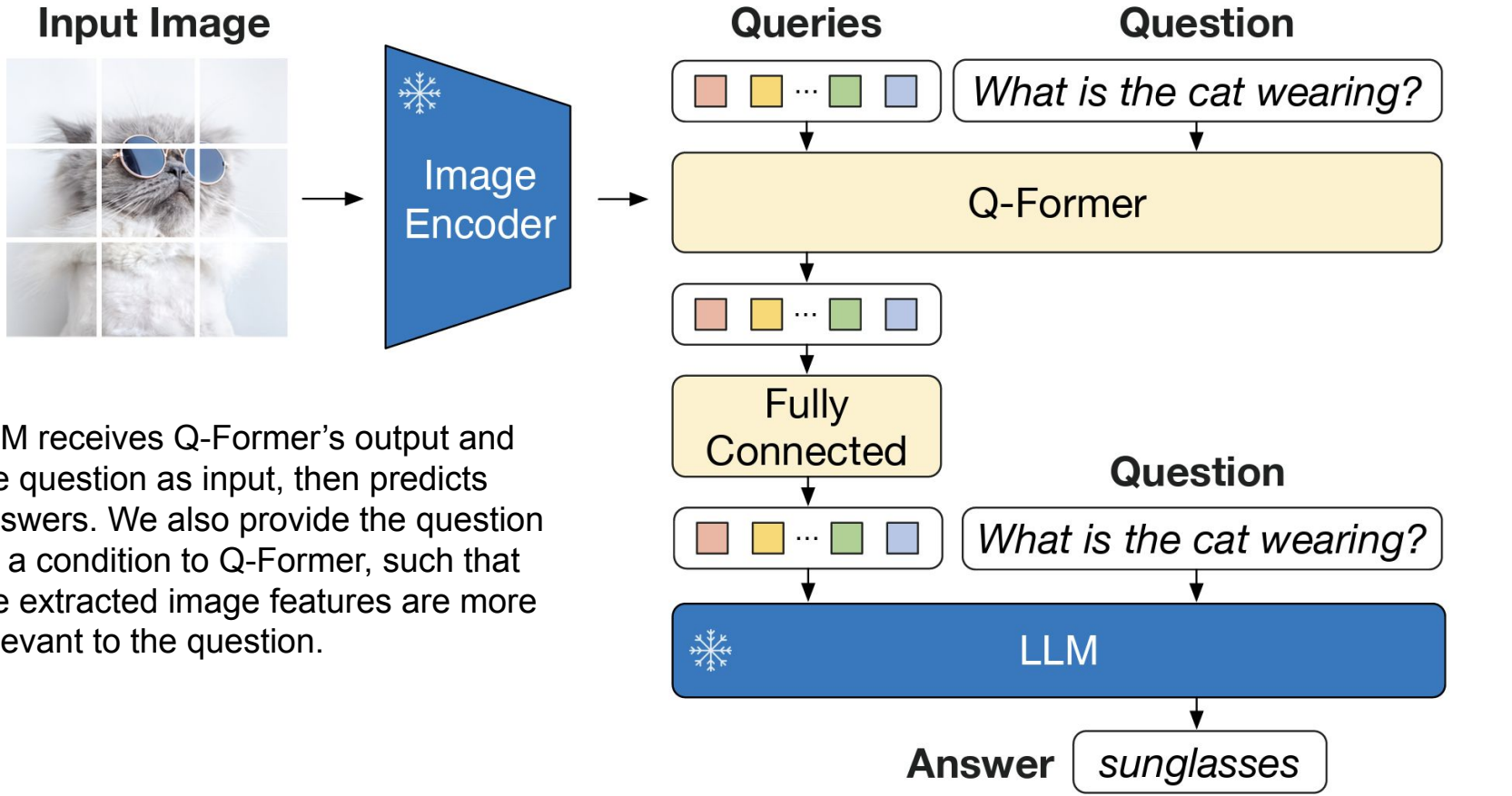
What is a soft prompt?

Bootstrap Vision-to-Language Generative Learning from a Frozen LLM



For decoder-based LLMs, we pre-train with the language modeling loss, where the frozen LLM is tasked to generate the text conditioned on the visual representation from Q-Former.
For **encoder-decoder-based LLMs**, we pre-train with the prefix language modeling loss, where we split a text into two parts. The prefix text is concatenated with the visual representation as input to the LLM's encoder. The suffix text is used as the generation target for the LLM's decoder.

VQA Finetuning



LLM receives Q-Former's output and the question as input, then predicts answers. We also provide the question as a condition to Q-Former, such that the extracted image features are more relevant to the question.

Example Outputs

Explain the advantages of this product.

The audi e-tron quattro concept is a plug-in hybrid electric sports car that has a range of 310 miles and can accelerate from 0 to 60 mph in just 4 seconds.

Tell me something about the history of this place.

The great wall of china was built in 221 bc by emperor qin shihuang to protect the emperor's capital from invaders from the north.

Write down the facts that you know about this flower.

It is a genus of flowering plants in the orchid family, orchidaceae. It is native to eastern asia and south east asia. It is a perennial flowering plant that grows up to a height of 1 m.