## Overview of bootstrapping VL pre-training with the trained P-Former. The alignment loss introduced by P-Former is agnostic to input modalities, encoders, and X-to-language connection modules.

## Background

VLMs can be trained end-to-end in an image-conditioned language generation task.

**Limitation:** End-to-end optimization of such a model is challenging [1].

## Training of Prompt-Transformer

The P-Former training resembles an autoencoder, with the bidirectional P-Former as the encoder and a causal LLM (frozen) as the decoder.

The objective is to reconstruct input text auto-regressively. [CLS] representation serves as sentence embeddings, which are projected back to the length of prompts.

This training process is purely based on text, allowing the P-Former to benefit from text outside the image-text pair dataset.

## Experimental Results

Comparison with different methods on zero-shot VQA.

Comparison with different methods on NoCaps and COCO. All methods optimize the CE loss during fine-tuning.

\[
\text{VATEX English video caption. Baseline is a sequential model, training end-to-end with ITG.}
\]

Ablations on \(w_1\) and \(w_2\) (using OPT 2.7B).

Ablations on sentence datasets used to train P-Former (using OPT 2.7B).

## Main Contributions

- Introducing the Prompt-Transformer (P-Former), a model that predicts these ideal prompts, which is trained exclusively on linguistic data, bypassing the need for image-text pairings.
- Our experiments reveal that our framework significantly enhances the performance of BLIP-2, and effectively narrows the performance gap between models trained with either 4M or 129M image-text pairs.
- Our framework is modality-agnostic and flexible in terms of architectural design, as validated by its successful application in a video learning task using varied base modules.

## Background

- The P-Former training resembles an autoencoder, with the bidirectional P-Former as the encoder and a causal LLM (frozen) as the decoder.
- The objective is to reconstruct input text auto-regressively. [CLS] representation serves as sentence embeddings, which are projected back to the length of prompts.
- This training process is purely based on text, allowing the P-Former to benefit from text outside the image-text pair dataset.

## Experimental Results

Comparison with different methods on zero-shot VQA.

Comparison with different methods on NoCaps and COCO. All methods optimize the CE loss during fine-tuning.

\[
\text{C: CEDEr, S: SPICE, B: BLEU}
\]

Experiments using Flan-T5 XL.

Ablations on \(w_1\) and \(w_2\) (using OPT 2.7B).

Ablations on sentence datasets used to train P-Former (using OPT 2.7B).