

LEARNING TO PROMPT FOR VISION-LANGUAGE MODELS



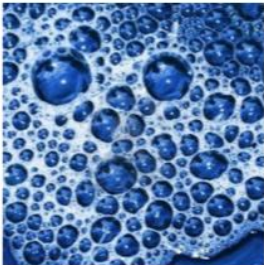

| | | |
|--|---|--------------|
|  Caltech101 | Prompt | Accuracy |
| | a [CLASS]. | 80.77 |
| | a photo of [CLASS]. | 78.99 |
| | a photo of a [CLASS]. | 84.42 |
| | $[V]_1 [V]_2 \dots [V]_M$ [CLASS]. | 92.00 |
| (a) | | |
|  Flowers102 | Prompt | Accuracy |
| | a photo of a [CLASS]. | 56.68 |
| | a flower photo of a [CLASS]. | 61.23 |
| | a photo of a [CLASS], a type of flower. | 62.32 |
| | $[V]_1 [V]_2 \dots [V]_M$ [CLASS]. | 93.22 |
| (b) | | |
|  Describable Textures (DTD) | Prompt | Accuracy |
| | a photo of a [CLASS]. | 38.24 |
| | a photo of a [CLASS] texture. | 37.71 |
| | [CLASS] texture. | 40.72 |
| | $[V]_1 [V]_2 \dots [V]_M$ [CLASS]. | 62.55 |
| (c) | | |
|  EuroSAT | Prompt | Accuracy |
| | a photo of a [CLASS]. | 22.30 |
| | a satellite photo of [CLASS]. | 31.12 |
| | a centered satellite photo of [CLASS]. | 31.53 |
| | $[V]_1 [V]_2 \dots [V]_M$ [CLASS]. | 81.60 |
| (d) | | |

Figure 1: **Prompt engineering vs. context optimization (CoOp)**. The latter uses only 16 shots for learning in these examples.

LEARNING TO PROMPT FOR VISION-LANGUAGE MODELS

$$\mathbf{t} = [\mathbf{V}]_1 [\mathbf{V}]_2 \dots [\mathbf{V}]_M [\mathbf{CLASS}],$$

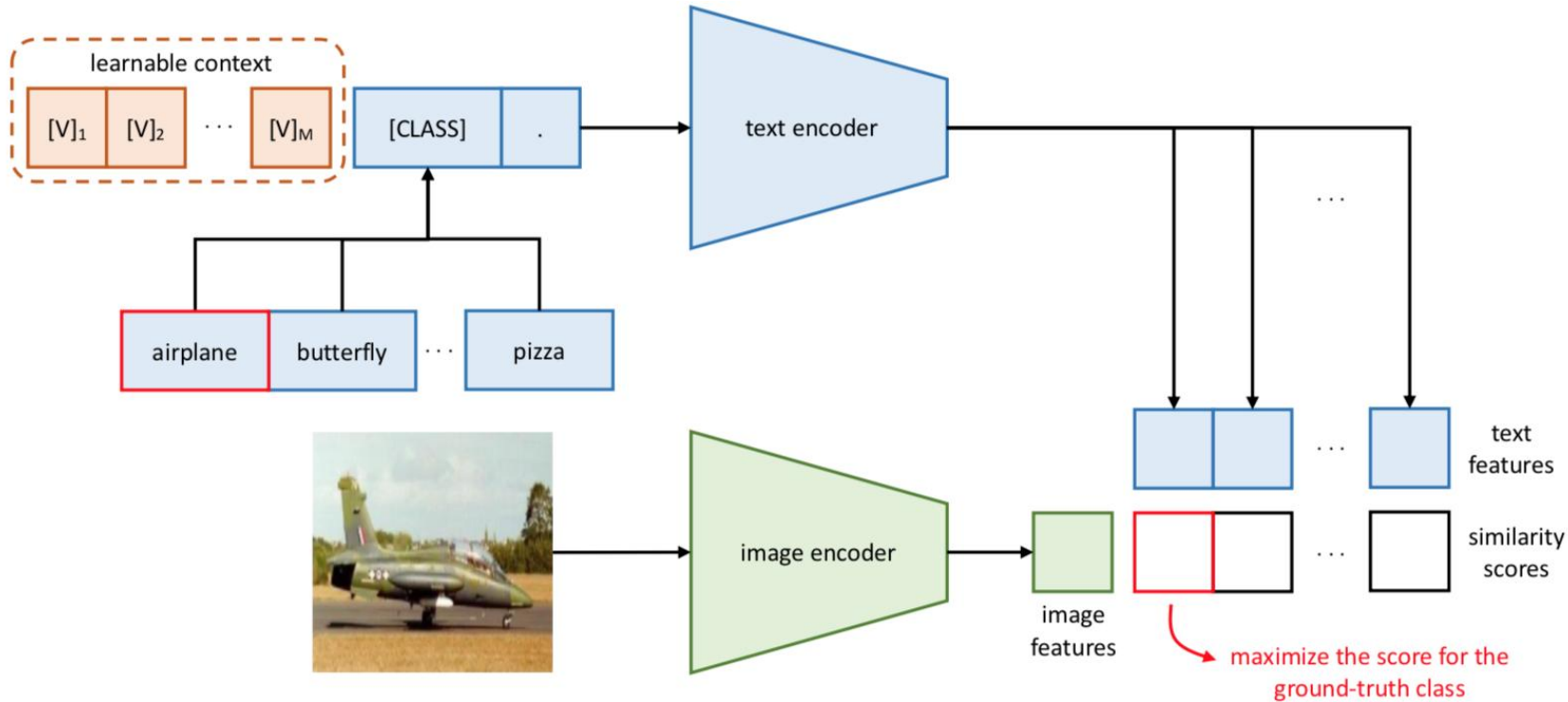
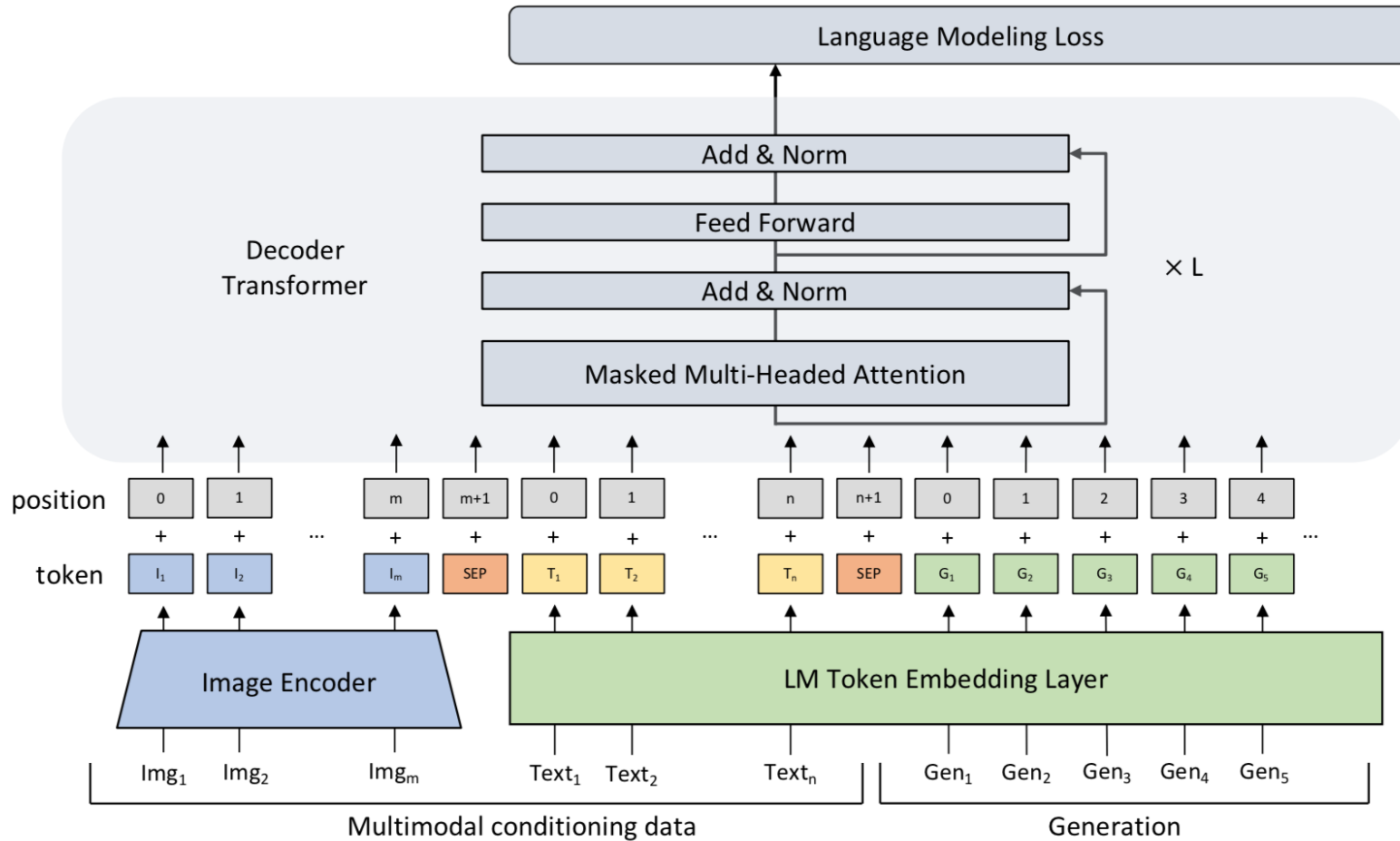


Figure 2: Overview of context optimization (CoOp).

$$p(y = i | \mathbf{x}) = \frac{\exp(\langle g(\mathbf{t}_i), \mathbf{f} \rangle / \tau)}{\sum_{j=1}^K \exp(\langle g(\mathbf{t}_j), \mathbf{f} \rangle / \tau)},$$

Multimodal Conditionality for Natural Language Generation



Given a sequence of token vectors $x = (x_1, \dots, x_n)$, language models learn the probability $p(x)$,

$$p(x) = \prod_{i=1}^n p(x_i | x_1, \dots, x_{i-1})$$

Here, we adapt a pretrained language model into a multimodal conditional model that learns the conditional probability distribution $p(x|y)$, where $y = (y_1, \dots, y_n)$ consists of tokens of any modality.

$$p(x|y) = \prod_{i=1}^n p(x_i | y, x_1, \dots, x_{i-1})$$

The goal is to learn $p(x|y)$ given supervised dataset of x, y pairs. To achieve this, we frame the problem using

Figure 1. An overview of the MANTiS architecture. Conditioning images are passed as input through an image encoder and mapped to textual token space of language model. Input text is encoded using the language model's encoder and together with image tokens form the conditionality prefix. The language modeling loss is computed only for the text tokens. Here m and n represent the number of input images and text tokens respectively and L is the number of decoder transformer layers.

SIMVLM: SIMPLE VISUAL LANGUAGE MODEL PRETRAINING WITH WEAK SUPERVISION

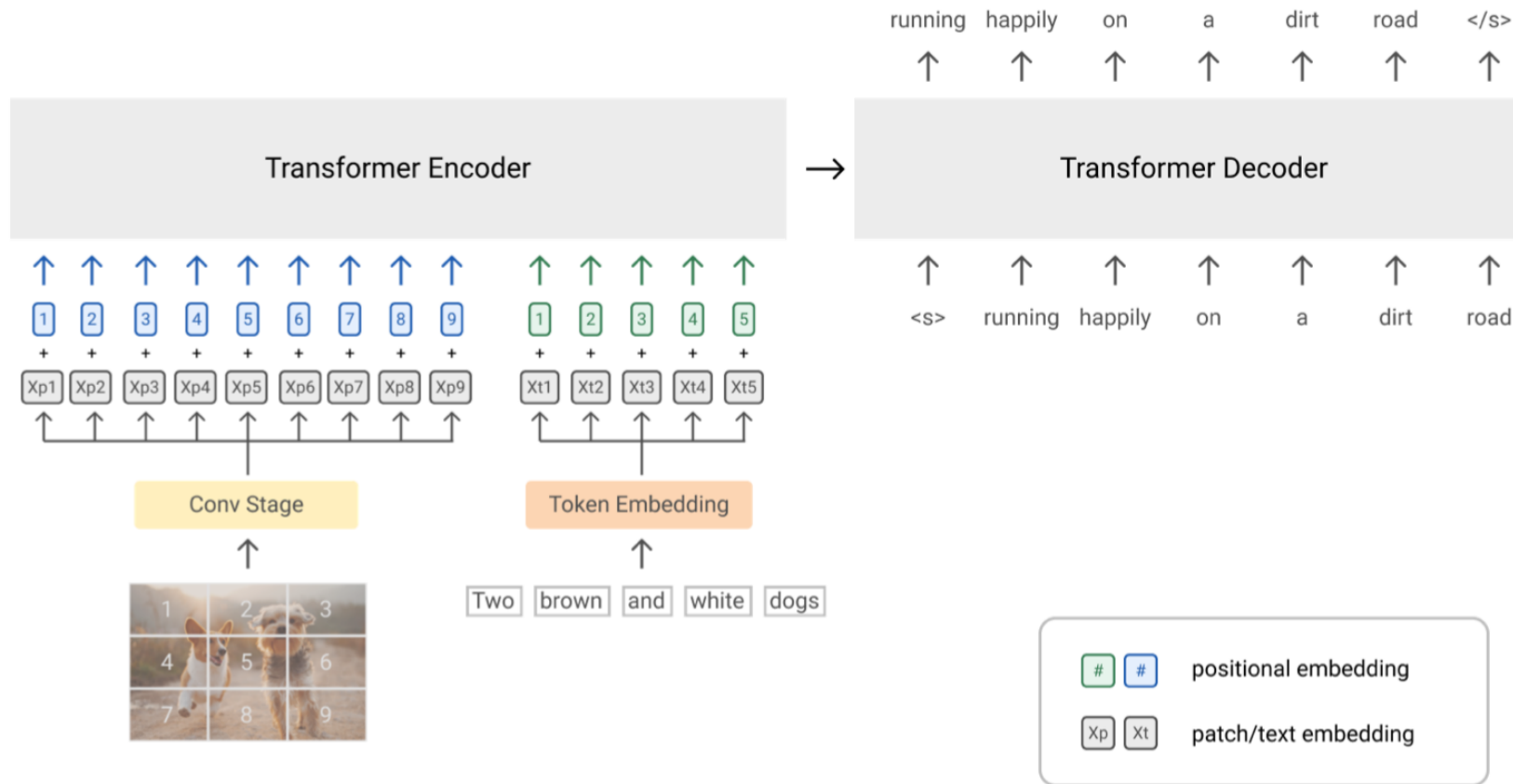


Figure 2: Illustration of the SimVLM model. This shows an example of training with PrefixLM of an image-text pair. For text-only corpora, it is straightforward to remove the image patches and utilize textual tokens only.

Maria: A Visual Experience Powered Conversational Agent

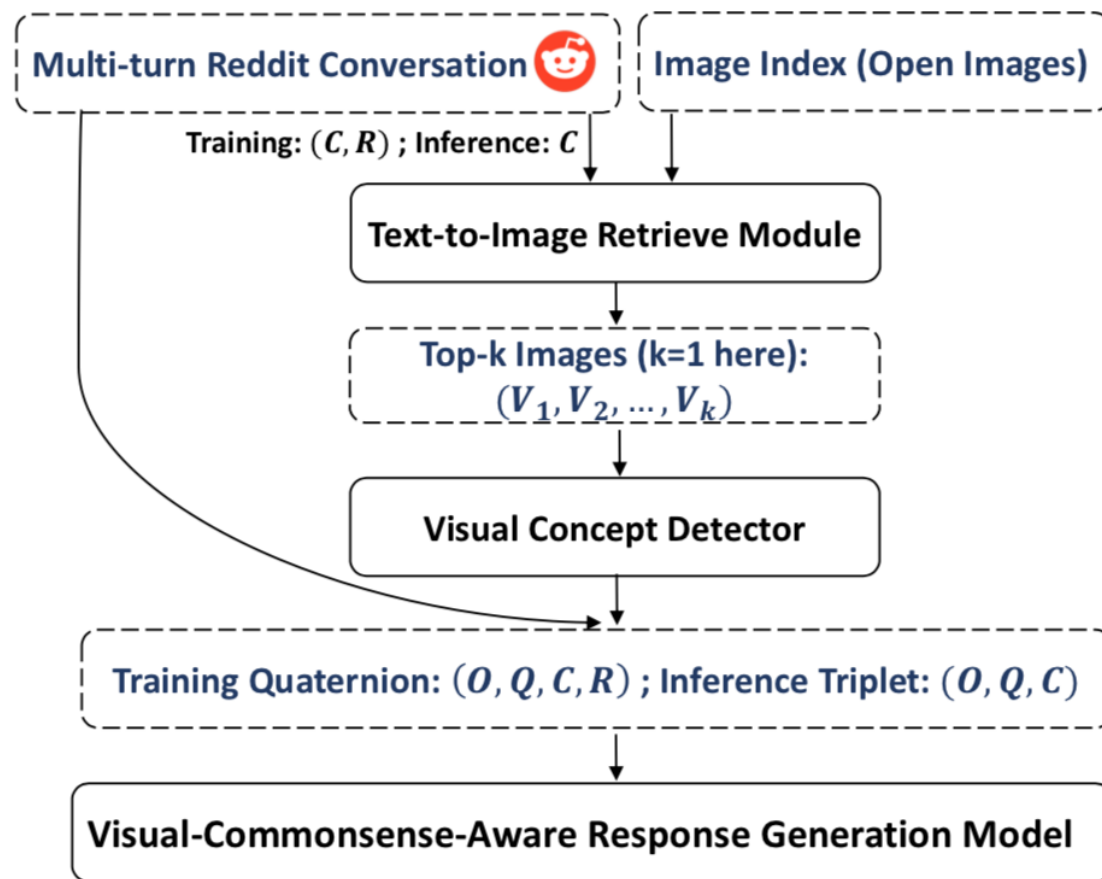


Figure 3: The flowchart of our framework. O, Q, C, R represents the image region features, extracted visual concepts, dialog context and response.

Maria: A Visual Experience Powered Conversational Agent

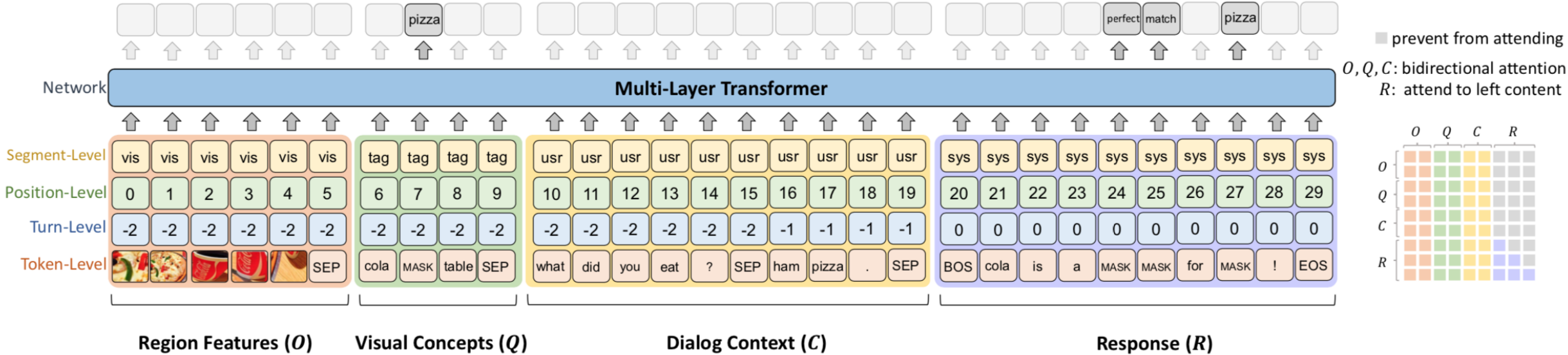


Figure 4: The overview of the response generation model. There are four kinds of inputs, *i.e.*, image region features O , extracted visual concepts Q , dialog context C and response R . The self-attention mask in R is unidirectional, *i.e.*, can only attend to the left context, while the self-attention mask in other segments is bidirectional.

Our Model

