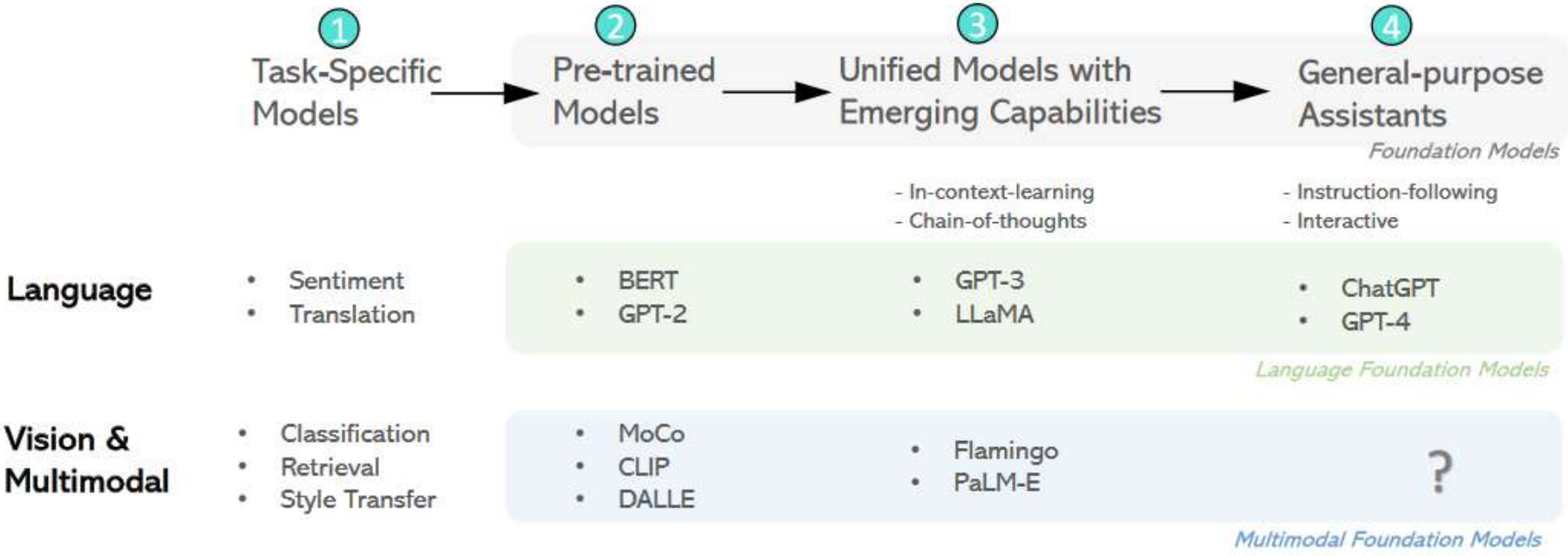
Multimodal Models

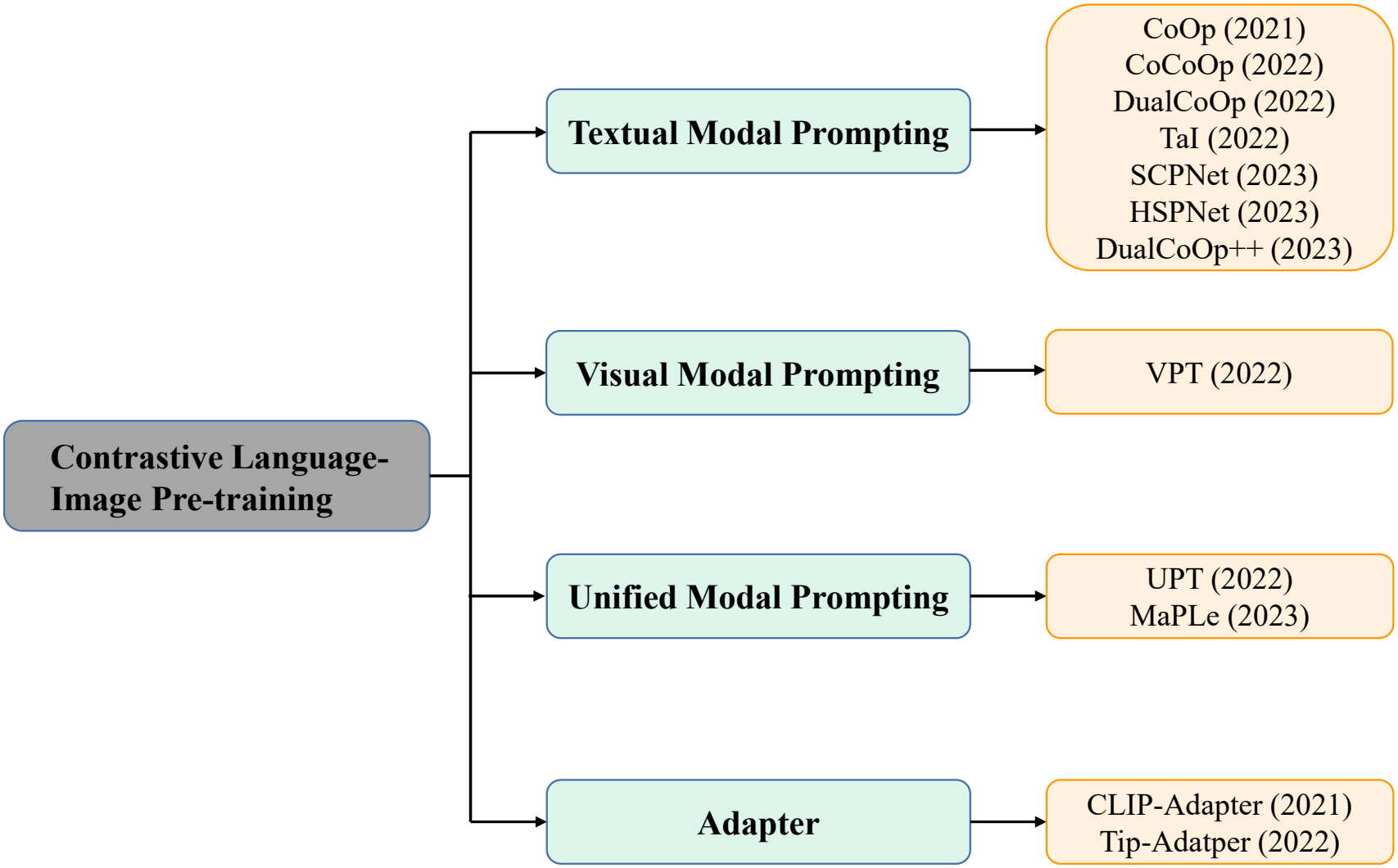
1. Contrastive Language-Image Pre-training

Background

Model development trajectory for language and vision/multimodality

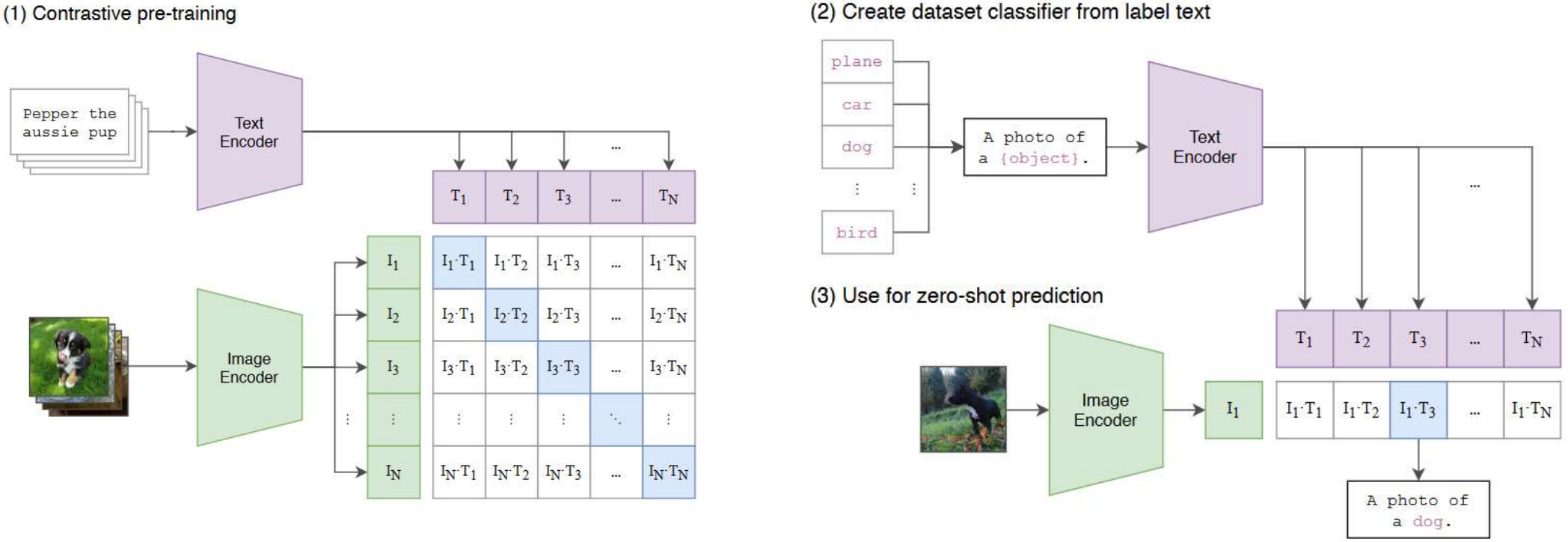


Taxonomy


Contrastive Language-Image Pre-training
Textual Modal Prompting
CoOp (2021)
CoCoOp (2022)
DualCoOp (2022)
TaI (2022)
SCPNet (2023)
HSPNet (2023)
DualCoOp++ (2023)
Visual Modal Prompting
VPT (2022)
Unified Modal Prompting
UPT (2022)
MaPLe (2023)
Adapter
CLIP-Adapter (2021)
Tip-Adatper (2022)

Background

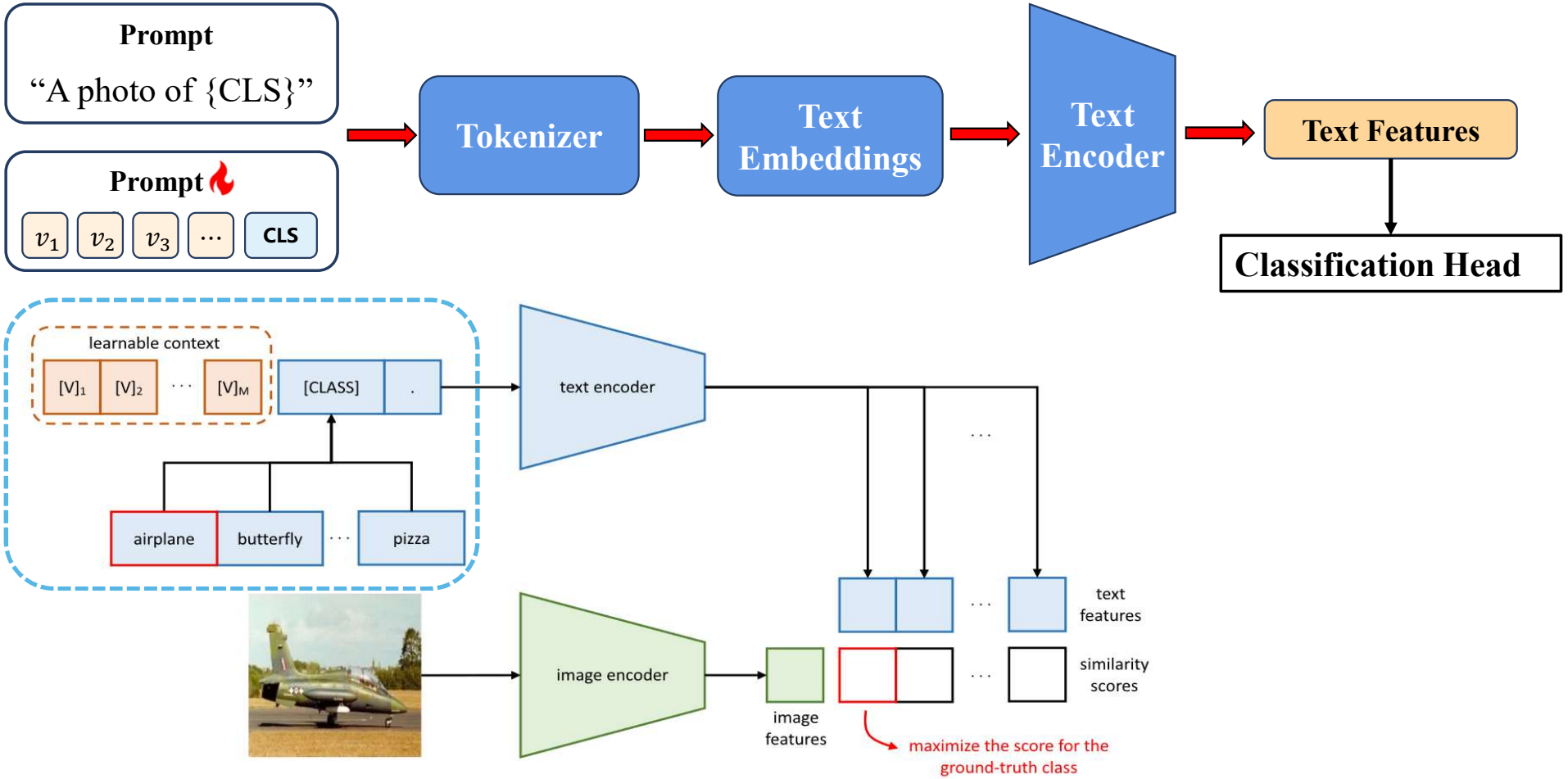
CLIP (2021)



Methods-Textual Modal Prompting

CoOp (2021)

Motivation: Learnable prompts are better than **hand-crafted** ones.



Methods-Textual Modal Prompting

CoCoOp (2022)

Motivation: The learned context is not generalizable to wider **unseen** classes within the same dataset, suggesting that CoOp **overfits base classes** observed during training.

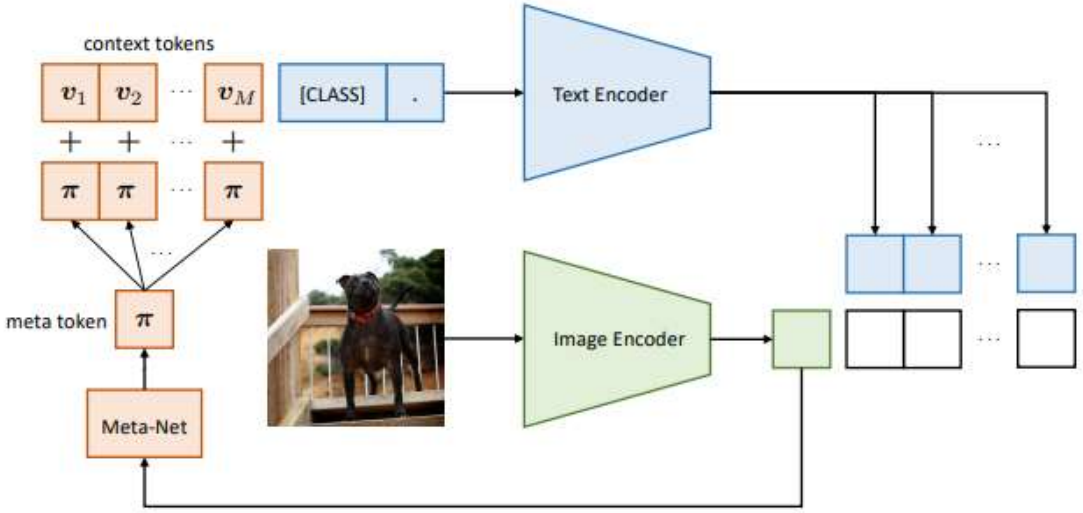


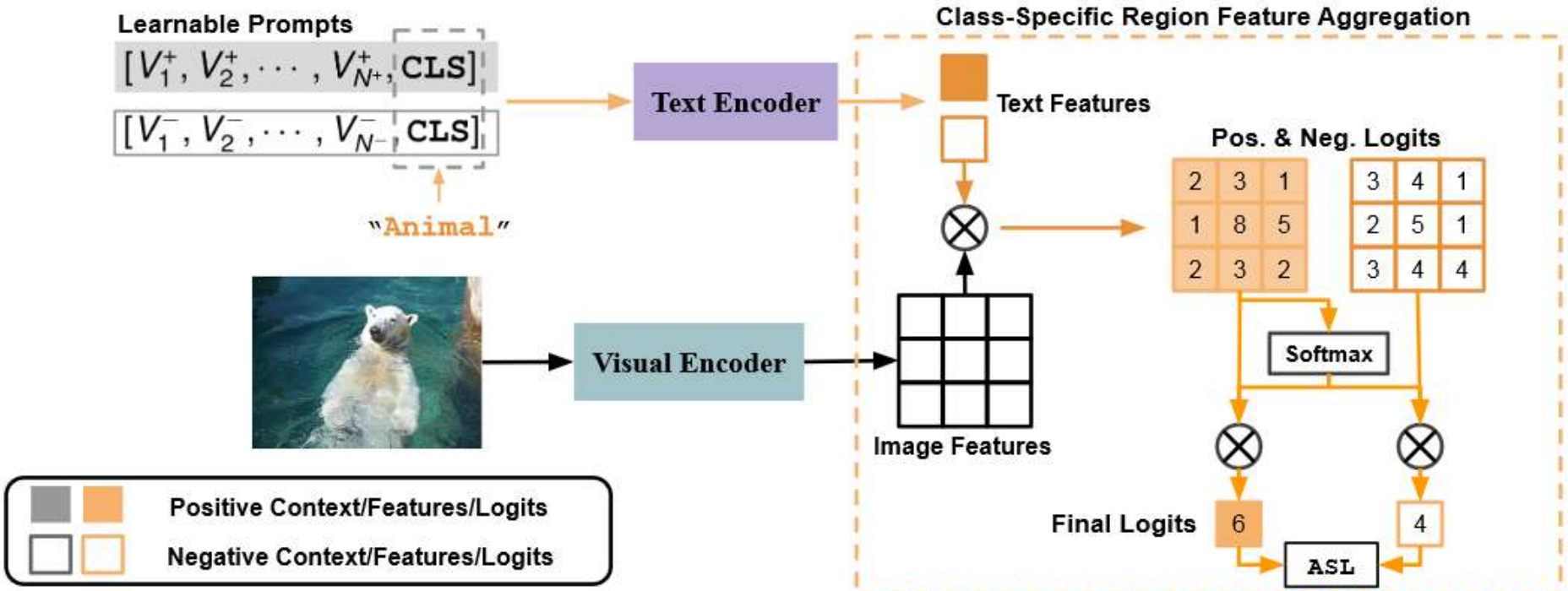
Figure 2. Our approach, Conditional Context Optimization (CoCoOp), consists of two learnable components: a set of context vectors and a lightweight neural network (Meta-Net) that generates for each image an input-conditional token.

- **Meta-Net.** Learn a lightweight neural network, called Meta-Net, to **generate** for each input a **conditional** token (vector), which is then combined with the context vectors.

Methods-Textual Modal Prompting

DualCoOp (2022)

Motivation: Encodes **positive** and **negative** contexts with **class names** as part of the linguistic input (i.e. prompts).

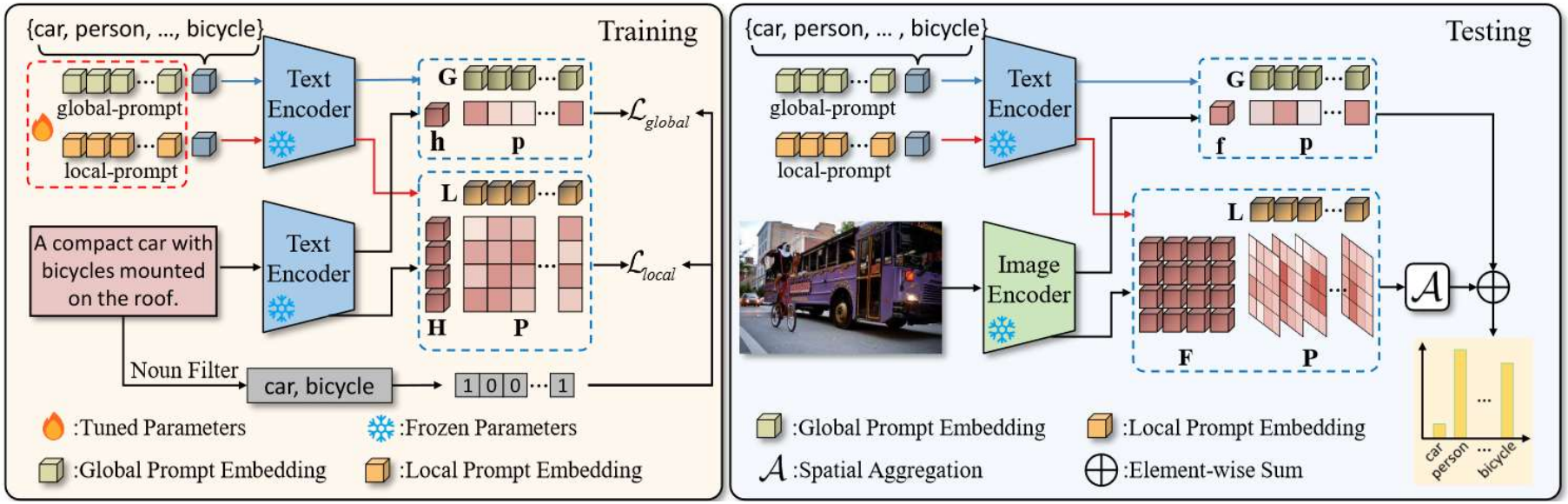


Methods-Textual Modal Prompting

TaI(2022) Dual-Prompt Design

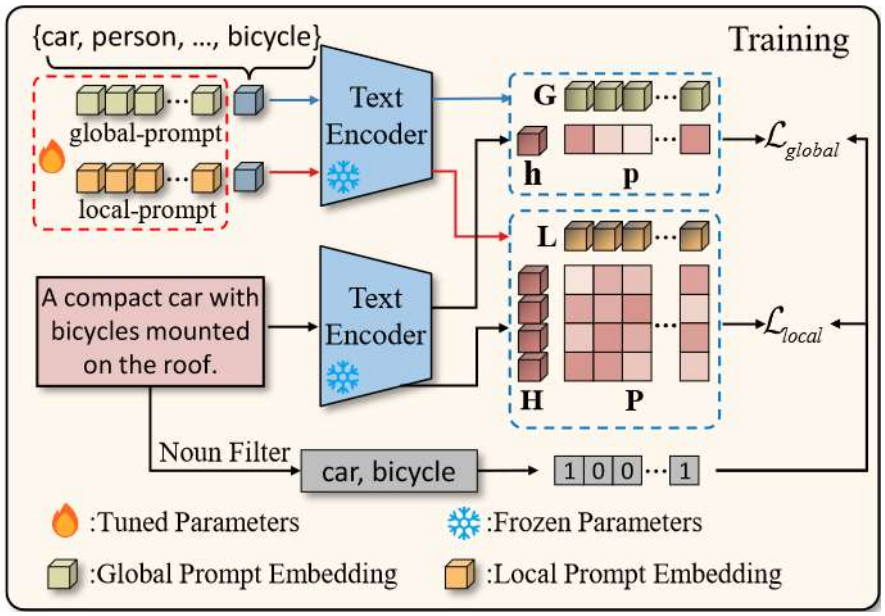
The **double-grained** prompt is defined as follows:

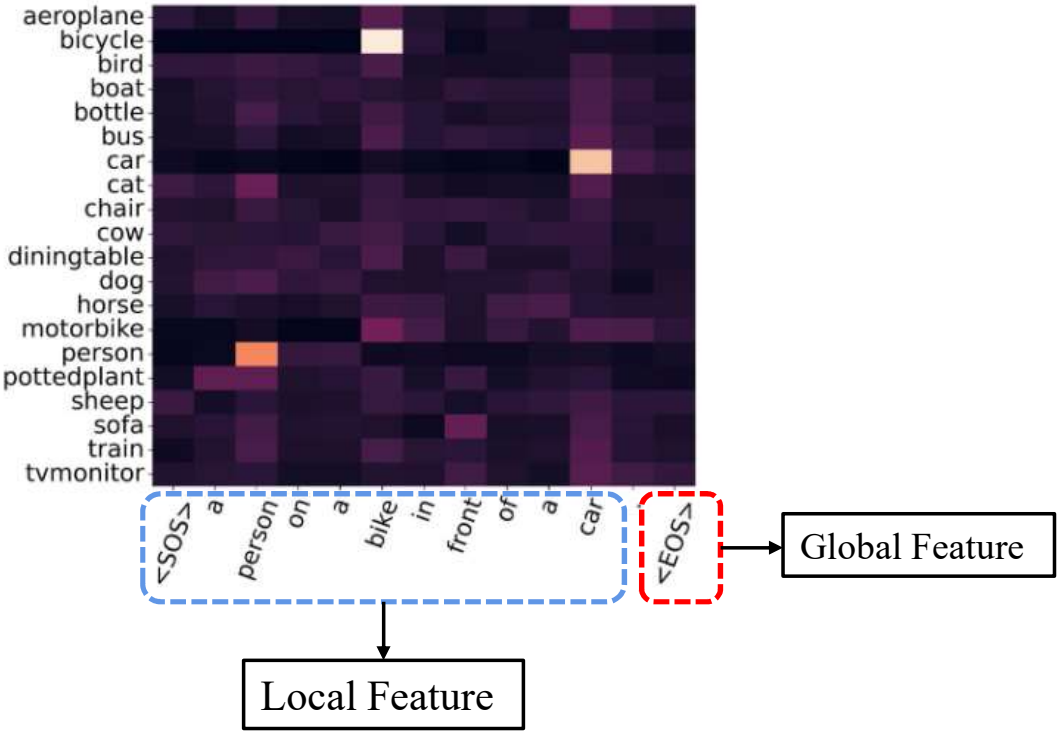
$$t_i^G = [v_1, v_2, v_3, ..., v_M, s_i],$$
$$t_i^L = [v'_1, v'_2, v'_3, ..., v'_M, s_i],$$



Methods-Textual Modal Prompting

TaI(2022) Training

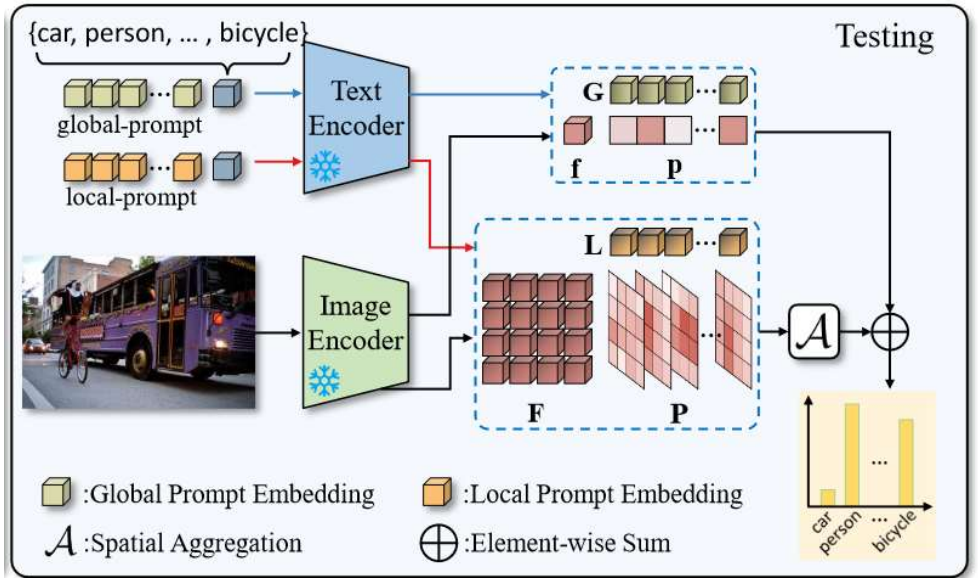


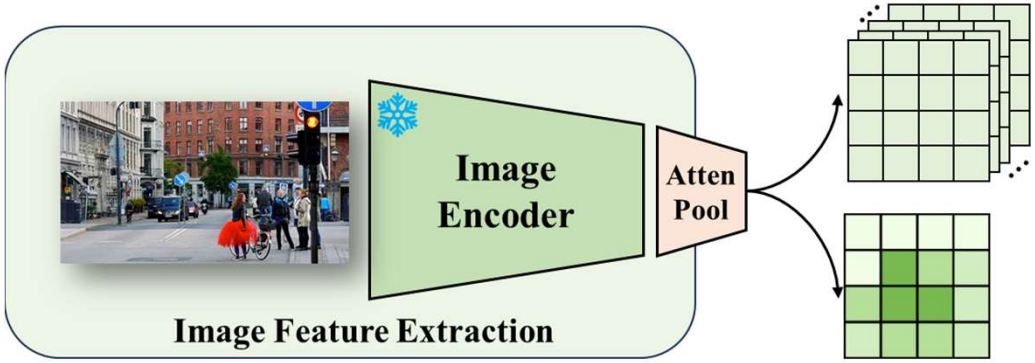


The sequential feature of **word tokens** from CLIP possesses **rich fine-grained information** which is very **similar to** the region feature of dense **image feature**.

Methods-Textual Modal Prompting

TaI (2022) Testing





local image feature z

global image feature $\bar{z}$

- Suppose we use ResNet-50 as the image encoder, so we can denote image feature as $\mathbf{x}_4 \in \mathbb{R}^{H_4 W_4 \times C}$
- Different from the original ResNet , CLIP makes a small modification by adding an **attention pooling** layer.
- Specifically, CLIP first performs global average pooling to obtain a global feature $\bar{\mathbf{x}}_4 \in \mathbb{R}^{1 \times C}$

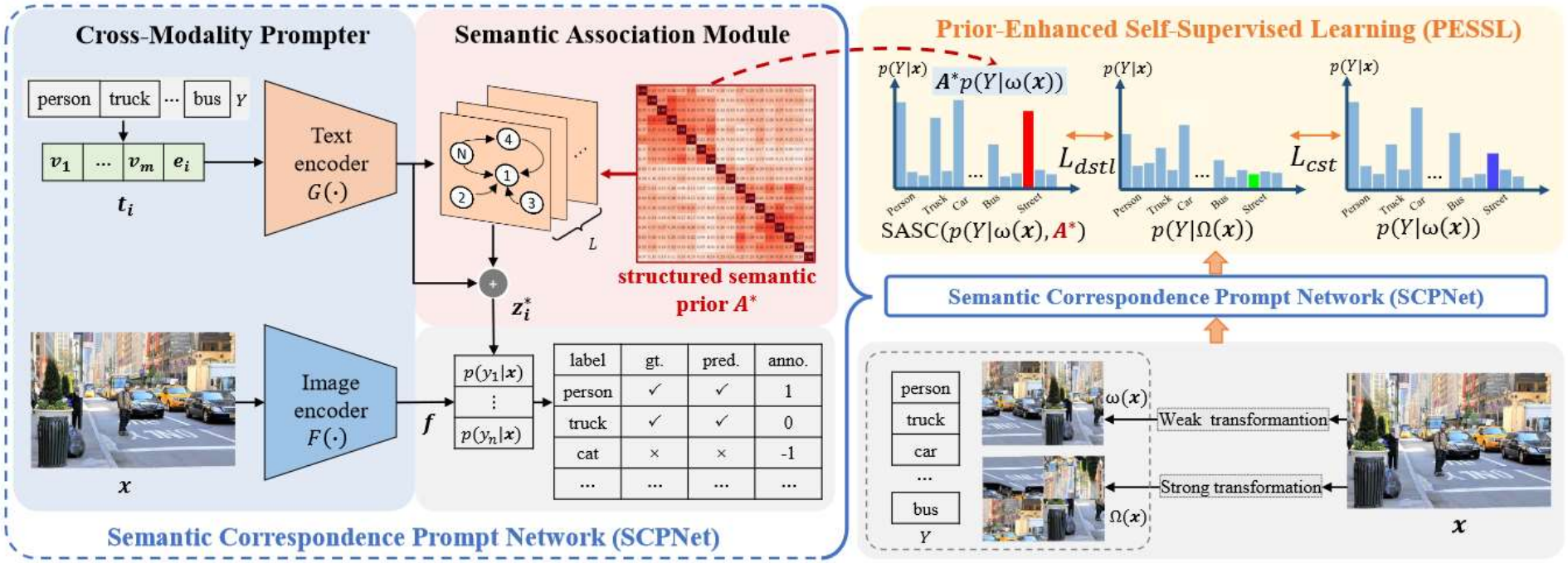
The concatenated features are then fed into an multi-head self-attention layer (MHSA):

$$[\bar{\mathbf{z}}, \mathbf{z}] = \mathrm{MHSA}([\bar{\mathbf{x}}_4, \mathbf{x}_4]).$$

Methods-Textual Modal Prompting

SCPNet (2023)

Motivation: **Correlation prior** among labels is helpful for multi-label learning.

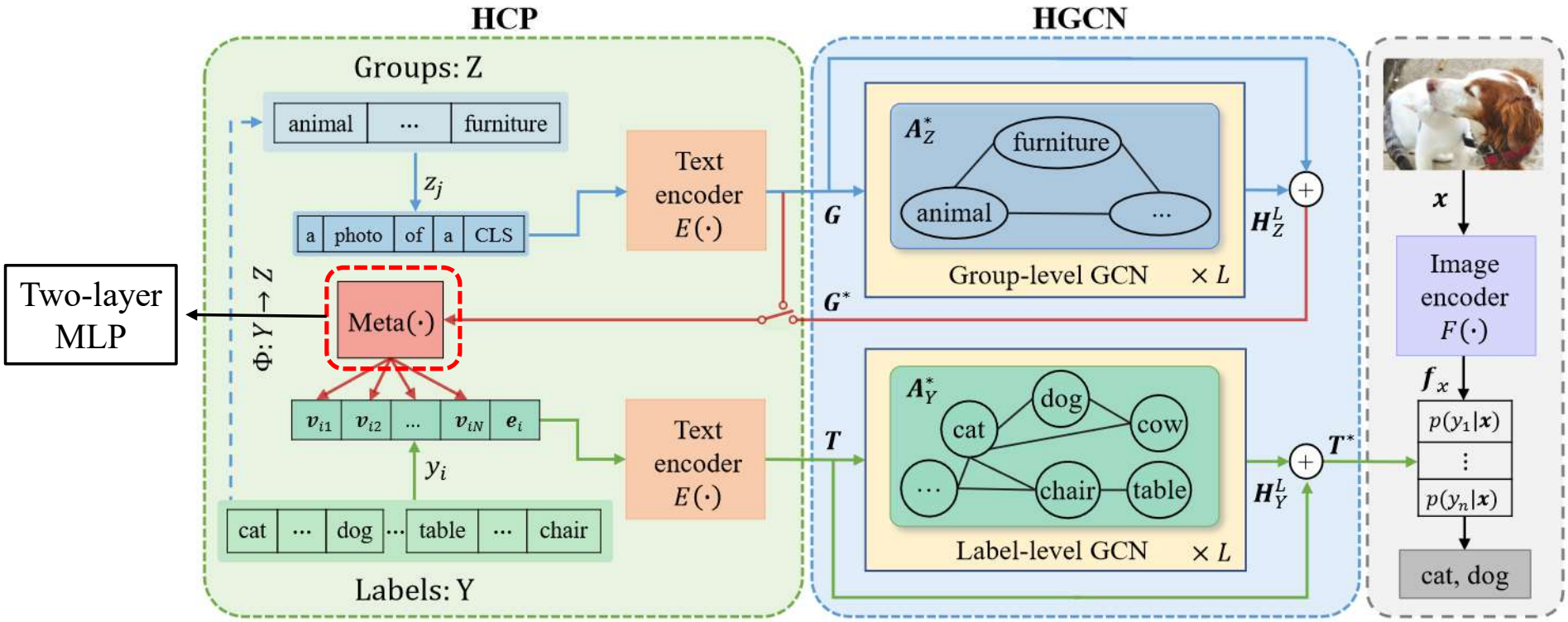


- **Semantic Association Module.** After extracting text feature of the same template prompt "**A photo of [CLS]**", **correlation** between different classes is estimated **as similarity?** And then use GCN to refine text feature.
- **Structure-Aware Semantic Calibration.** Calibrate prediction with sematic prior matrix.

Methods-Textual Modal Prompting

HSPNet (2023)

Motivation: Labels can be clustered into different groups.

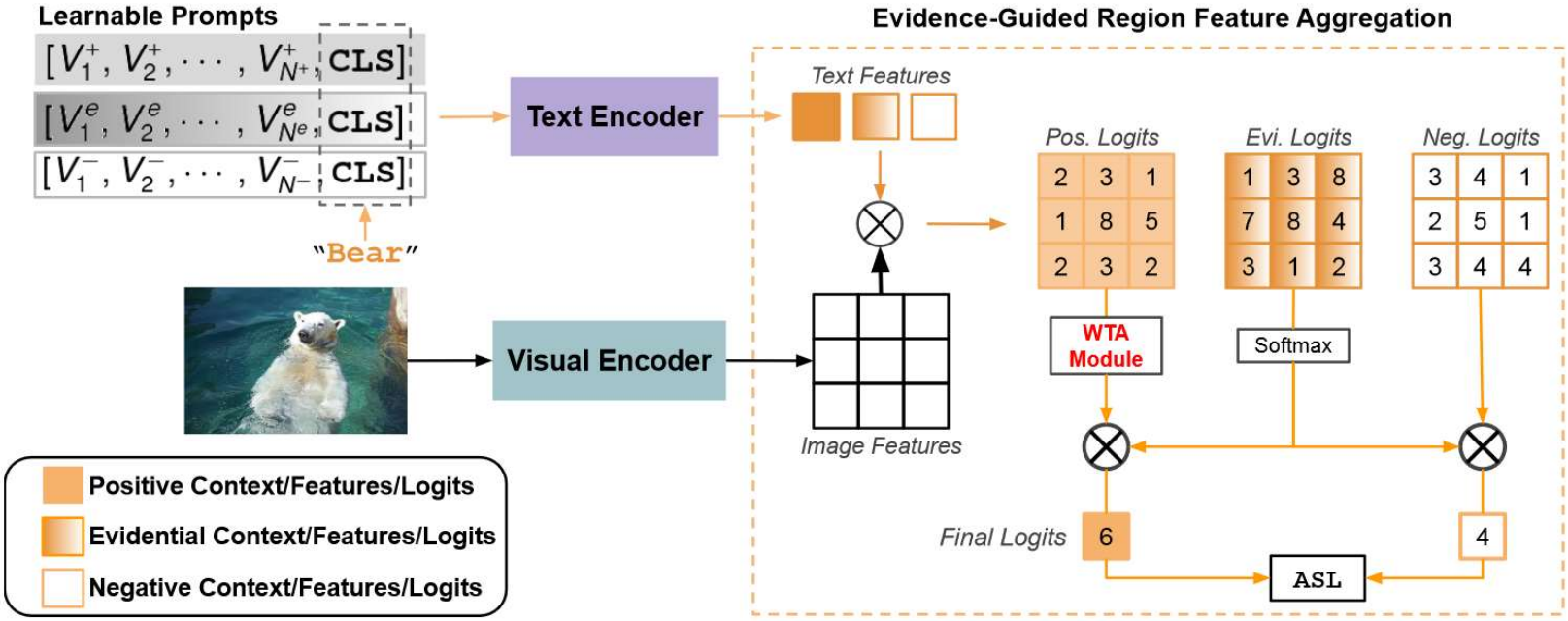


- **Hierarchical Conditional Prompt.** Categories that belong to the **same group** have the **same learnable prompts**. Each of them is then concatenate with their class names to obtain the final prompts.
- **Hierarchical Graph Convolutional Network.** Use two GCNs to capture the high-order inter-label and inter-group relationships.

Methods-Textual Modal Prompting

DualCoOp++ (2023)

Motivation: Introduce a new evidential prompt based on DualCoOp. The evidential context aims to discover all the related visual content for the target class.



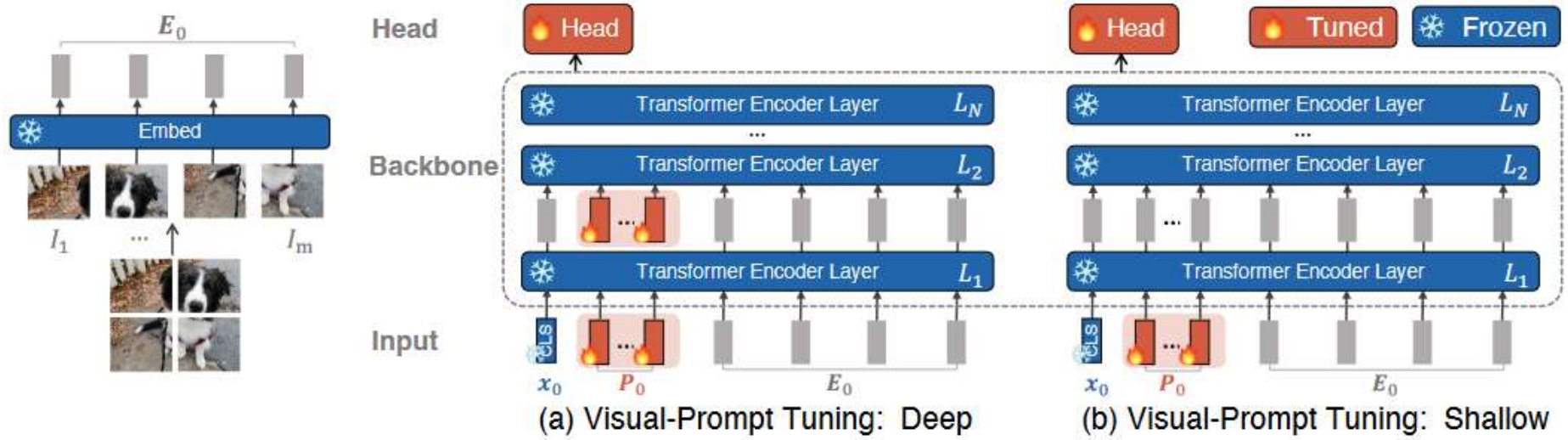
- **Evidence-Guided Region Feature Aggregation.** Use the logits generated by **evidential prompt** to reweight positive and negative logits.
- **Winner-Take-All Regularization.** Each **spatial region** to only positively respond to at most **one class**.

$$w_i = \text{softmax}(\gamma \cdot S_i^+ \cdot \max_m(S_i^+)) \qquad (S_i^+)' = w_i \odot S_i^+$$

Methods-Visual Modal Prompting

Visual Prompt Tuning (2022) **Transformer Architecture Only**

Motivation: Introduce prompts into visual encoder. **Like P-tuning v2**

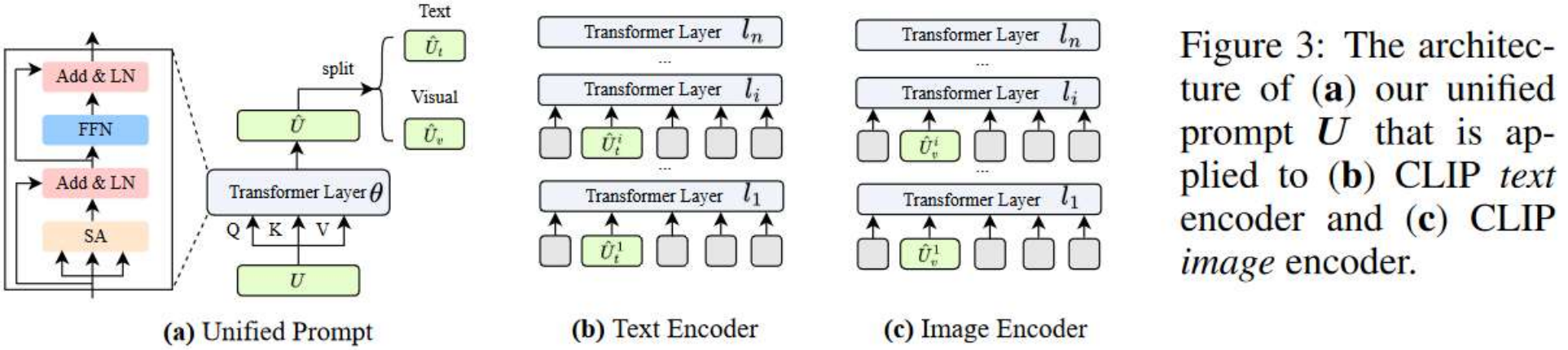


- **VPT-Shallow.** Prompts are inserted into the first Transformer layer L1 only.
- **VPT-Deep.** Prompts are introduced at every Transformer layer's input space.

Methods-Unified Modal Prompting

Unified Prompt Tuning (2022) Transformer Architecture Only

Motivation: None of the **unimodal** prompt tuning methods **performs** consistently **well**: text prompt tuning fails on data with high intra-class visual variances while visual prompt tuning cannot handle low inter-class variances.



(a) Unified Prompt **(b)** Text Encoder **(c)** Image Encoder

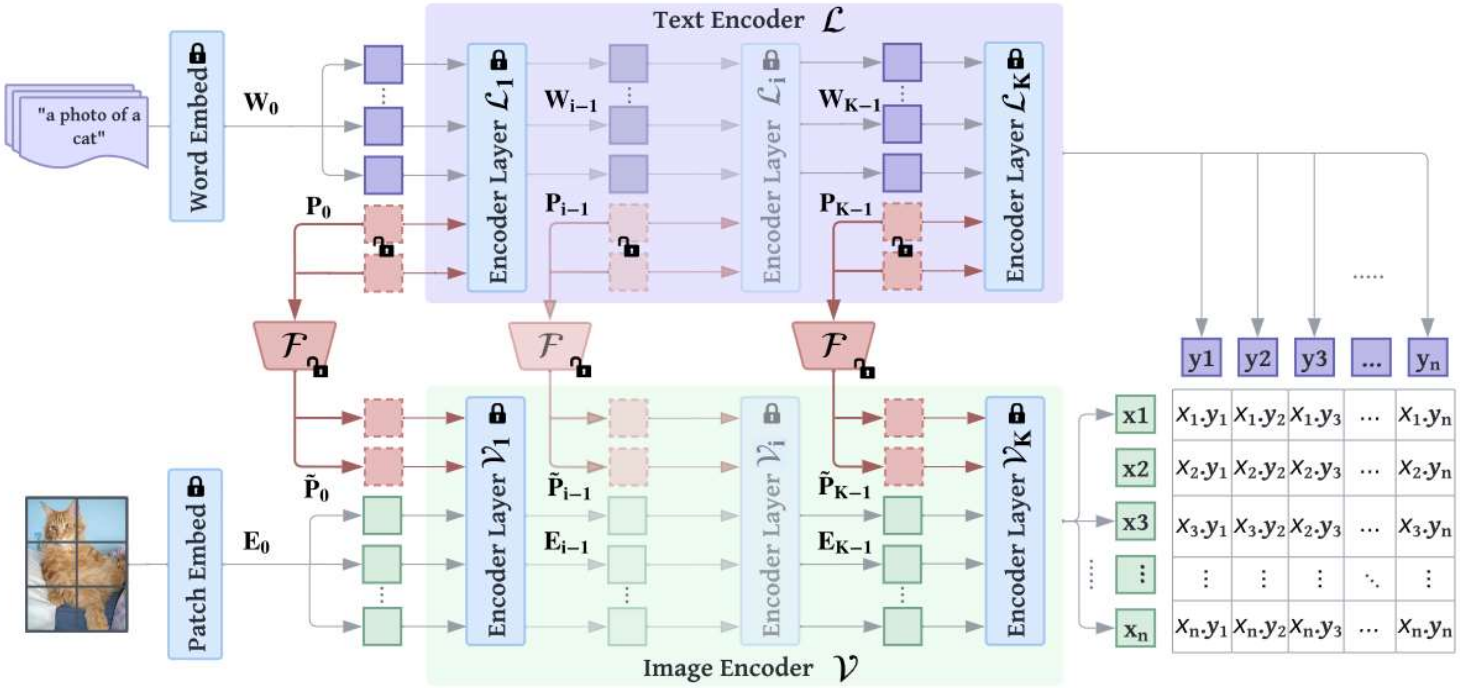
Figure 3: The architecture of (**a**) our unified prompt U that is applied to (**b**) CLIP *text* encoder and (**c**) CLIP *image* encoder.

- **Unified Prompt Tuning (UPT).** Employ a lightweight Transformer layer to extract prompt feature for both visual and textual modalities.

Methods-Unified Modal Prompting

MaPLe (2023) **Transformer Architecture Only**

Motivation: Using prompting to adapt representations in a **single** branch of **CLIP** (language or vision) is **sub-optimal** since it does **not allow the flexibility** to dynamically adjust both representation spaces on a downstream task.

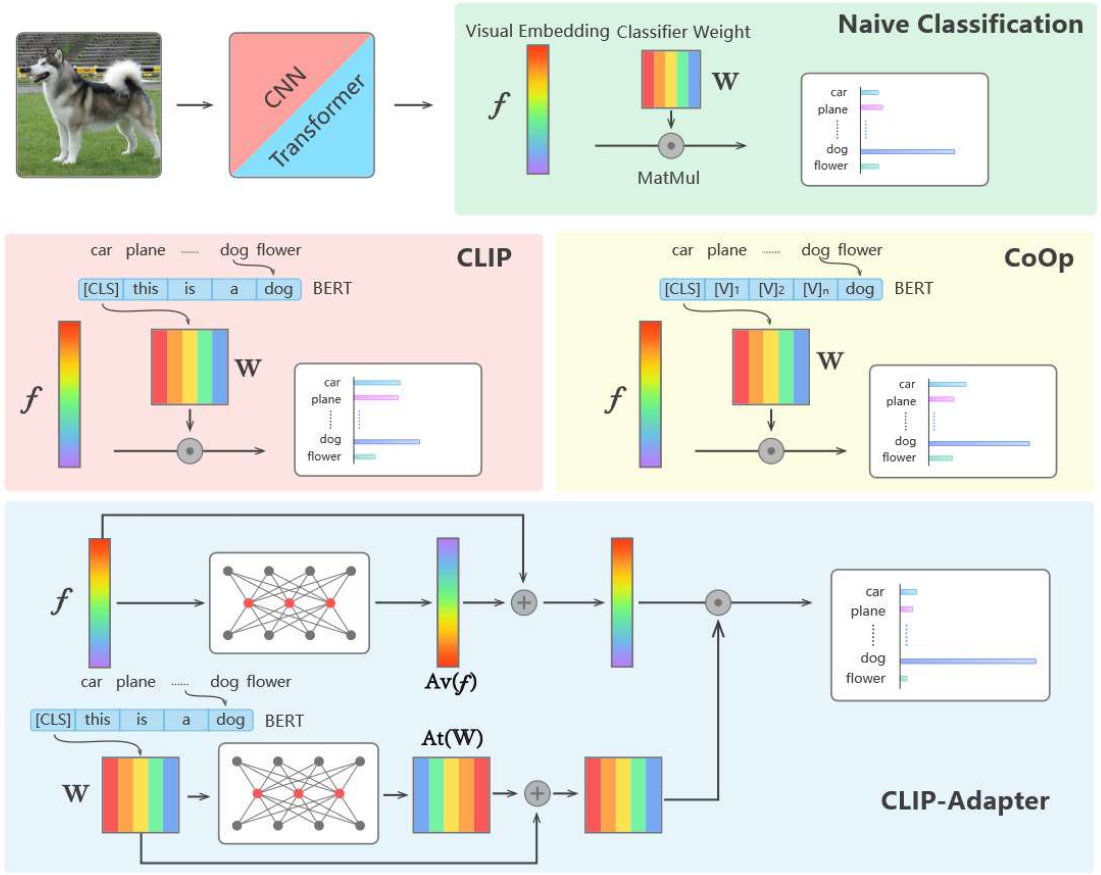


- **Vision Language Prompt Coupling.** Project language prompts P via vision-to-language projection which we refer to as V-L coupling function $F(\cdot)$. The coupling function is implemented as a linear layer which maps d_l dimensional inputs to d_v.

Methods-Adapter

CLIP-Adapter (2021)

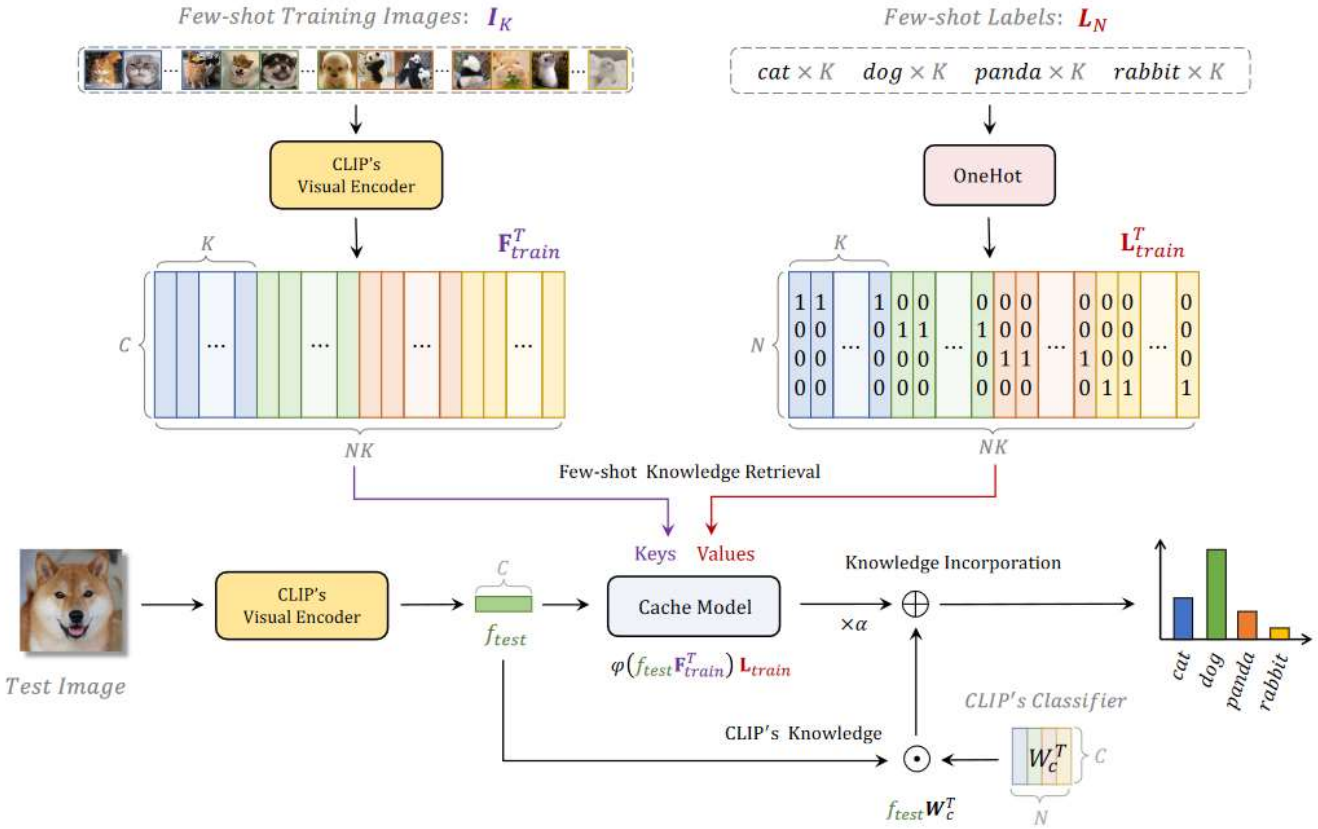
Motivation: Conduct fine-tuning with feature adapters on both visual and language branches.



Methods-Adapter

Tip-Adapter (2022)

Motivation: Constructs the adapter via a key-value cache model from the few-shot training set, and updates the prior knowledge encoded in CLIP by feature retrieval.



Cache Model Construction. There are K annotated images in each of the N categories, denoted as I_K with their labels I_N. Use feature of test image as query to retrieve relevant knowledge from **Cache Model**.

Tip-Adapter. Use text feature as classification head to generate original logit, and it is merged with the query result from Cache Model.

$$\begin{aligned}\text{logits} &= \alpha A\mathbf{L}_{\text{train}} + f_{\text{test}}W_c^T \\ &= \alpha\varphi(f_{\text{test}}\mathbf{F}_{\text{train}}^T)\mathbf{L}_{\text{train}} + f_{\text{test}}W_c^T,\end{aligned}$$

Thanks