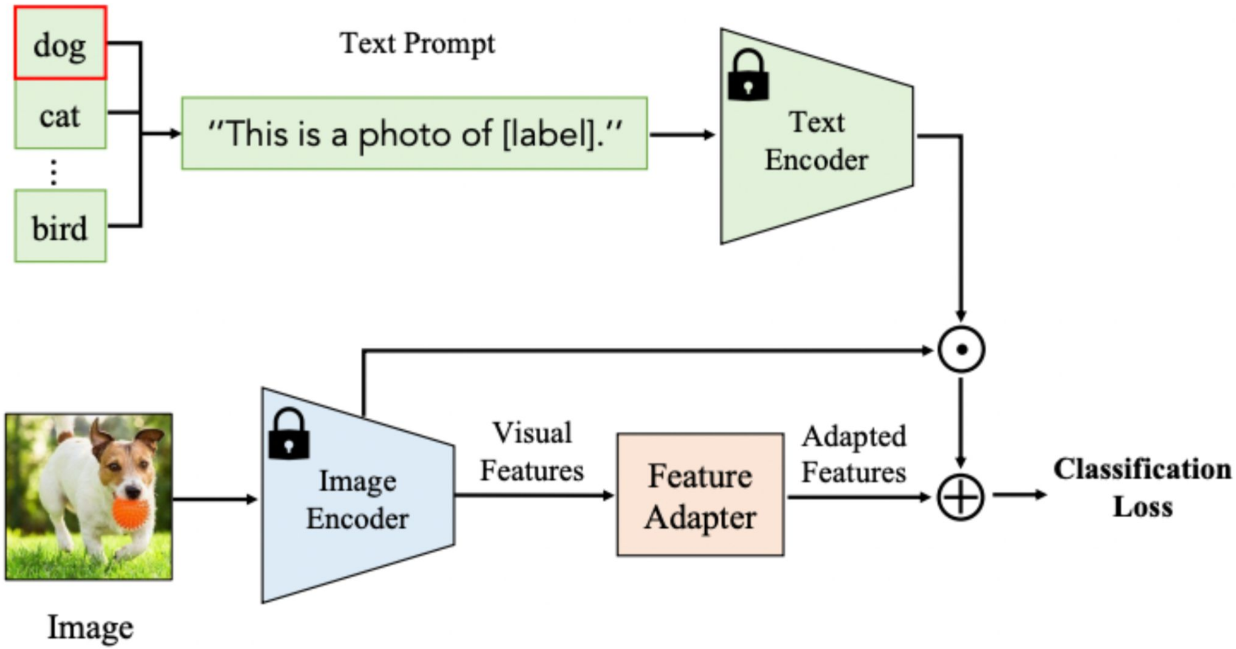
VLM 끄적끄적 7

Transfer Learning (2) Feature Adaptation + 기타

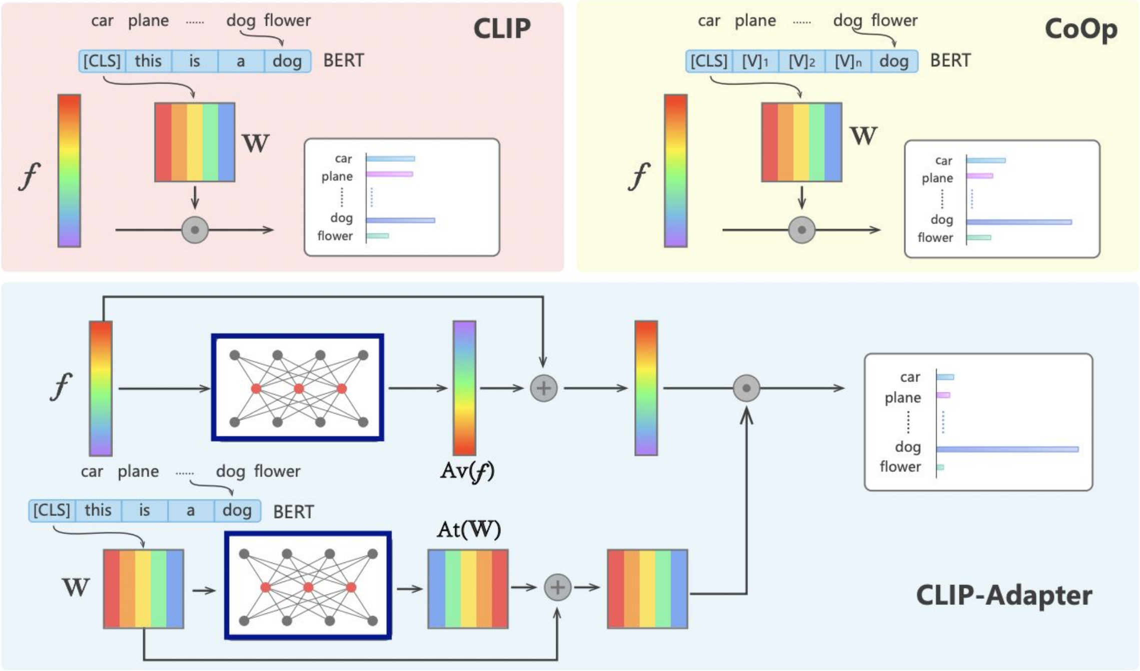
Feature Adaptation의 개요

VLM을 additional (light-weight) feature adapter를 사용하여 fine-tune 하자!



1. CLIP-Adapter (arxiv 2021)

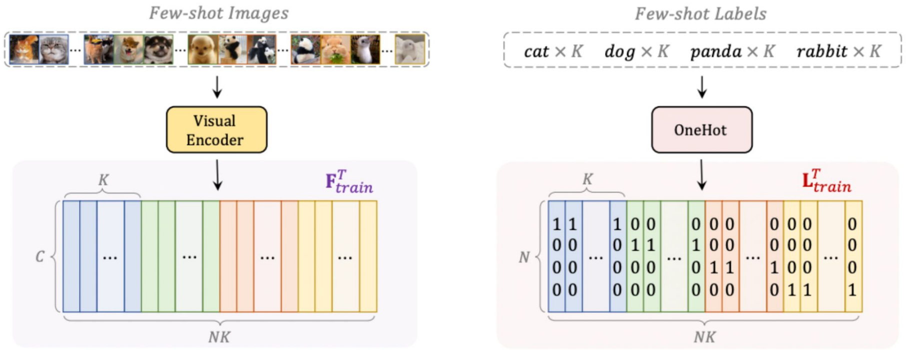
핵심: Language encoder & Image encoder 뒤에 linear layer 붙이자



2. Tip-Adapter (ECCV 2022)

Tip-Adapter = “Training-Free” CLIP-Adapter

- Few-shot 데이터에 대한 embedding을 곧 adapter로써 취급(사용)하자!
- Why ‘Training free’?
  - 별도의 adapter 학습(존재) X
  - pretrained encoder만을 사용 C



2. Tip-Adapter (ECCV 2022)

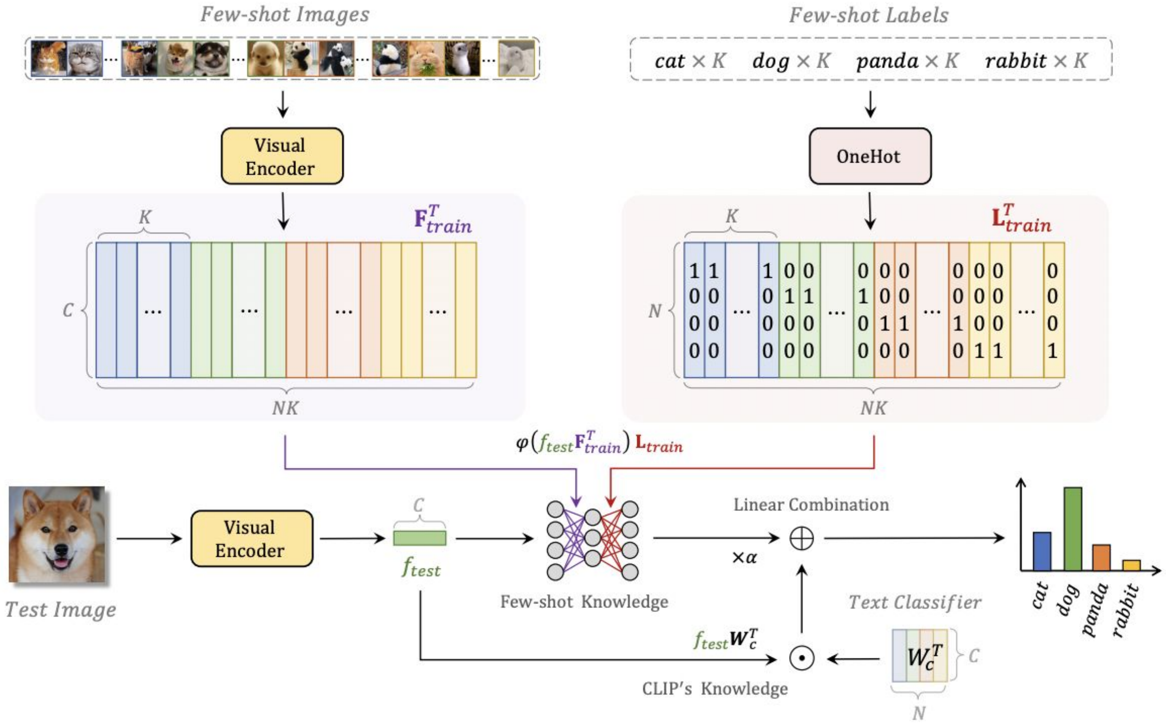
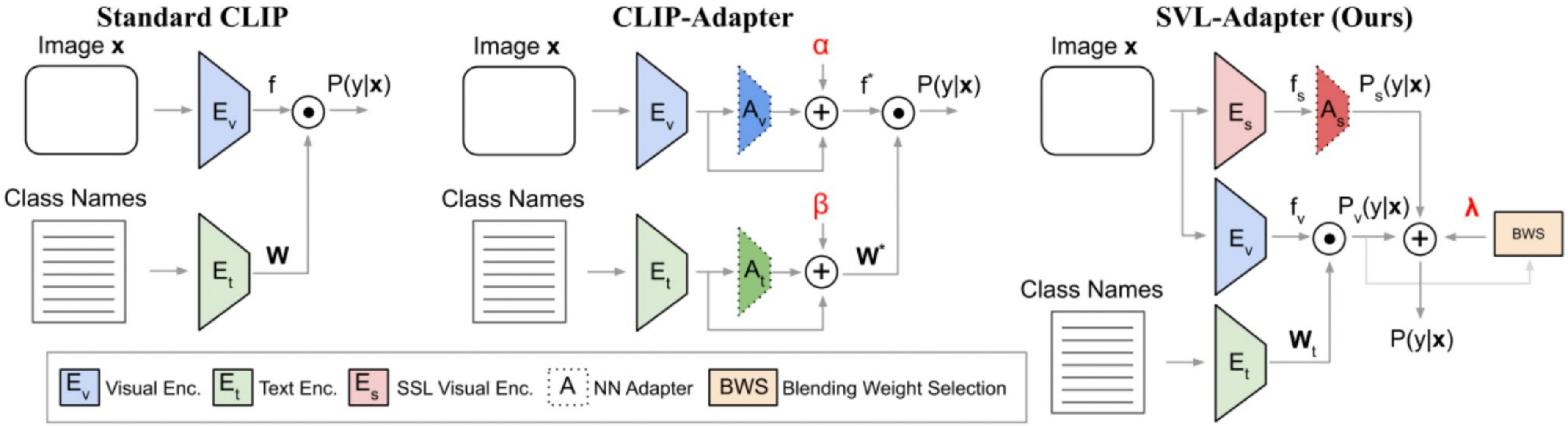


Figure 2. **The Pipeline of Tip-Adapter.** Given a K-shot N-class training set, we construct the weights of the two-layer adapter by creating a cache model from the few-shot training set. It contains few-shot visual features F_{train} encoded by CLIP's visual encoder and few-shot ground-truth labels L_{train}. F_{train} and L_{train} can be used as the weights for the first and second layers in the adapter.

3. SVL-Adapter (BMVC 2022)

SSL adapter 사용 (+ additional encoder)

- VL pretraining + SSL pretraining을 모두 활용하기 위해



4. 기타 방법론들

[1] WiSE-FT (CVPR 2022)

- 기존 FT 방법론들의 한계점: Distribution shift에 대해 robust X
- 해결책: WiSE-FT = Weight-Space Ensemble
  - Original VLM & Fine-tuned VLM의 weight를 combine한다.

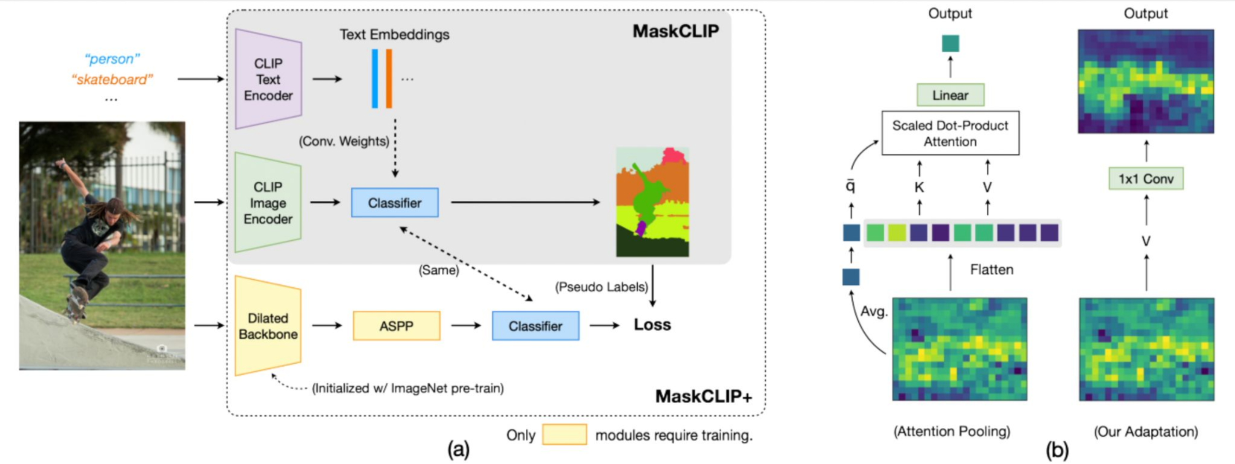
Step 2: Weight-space ensembling. For a *mixing coefficient* $\alpha \in [0, 1]$, we consider the *weight-space ensemble* between the zero-shot model with parameters θ_0 and the model obtained via standard fine-tuning with parameters θ_1. The predictions of the weight-space ensemble wse are given by

$$\mathsf{wse}(x, \alpha) = f(x, (1-\alpha)\cdot\theta_0 + \alpha\cdot\theta_1)\ , \tag{1}$$

4. 기타 방법론들

[2] MaskCLIP (ECCV 2022)

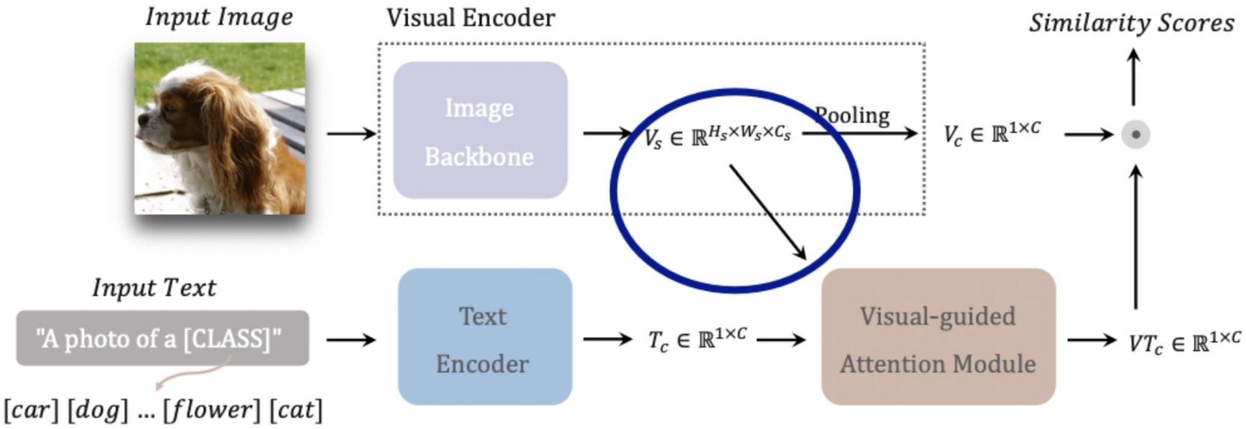
- 태스크: Dense prediction
- Goal: Dense prediction을 위해, CLIP image encoder 아키텍처 수정하자!



4. 기타 방법론들

[3] VT-CLIP (arxiv 2021)

- CLIP 기반 연구의 한계점: Dataset들 간의 “semantic gap”
- 해결책: Visual-guided text로써 보완! $VT_c = \text{Cross Attn}(V_s, V_s, T_c) + T_c$



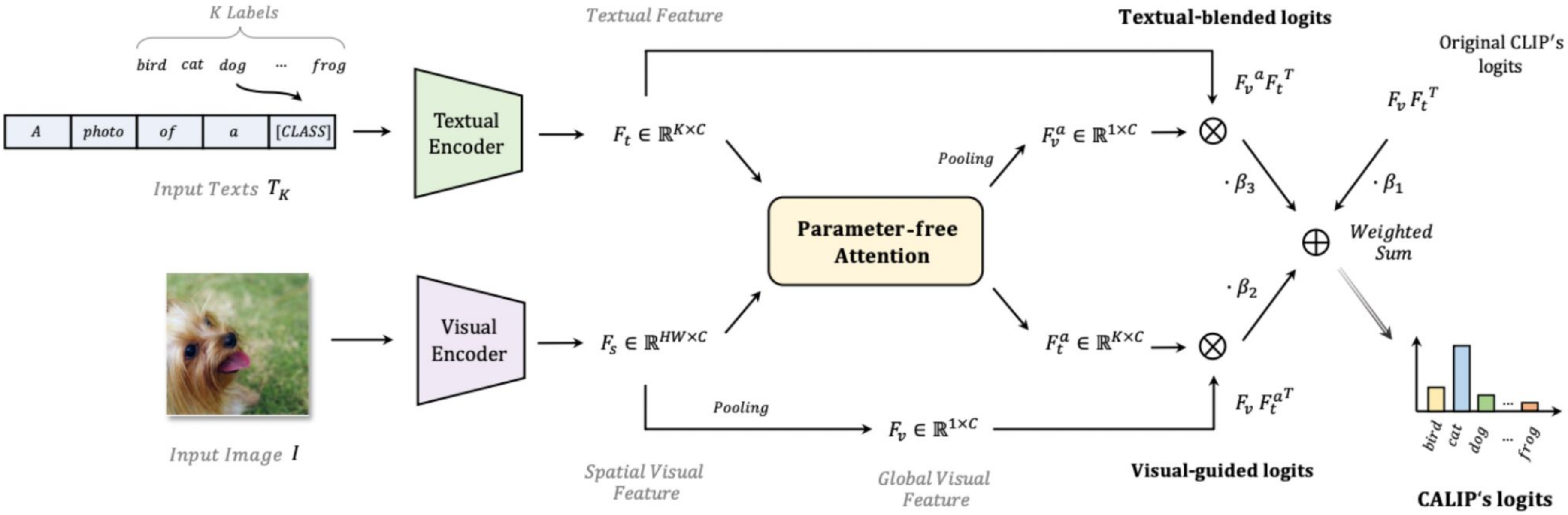
4. 기타 방법론들

[4] CALIP (AAAI 2023)

- 기존: CLIP + additional learnable module
- 제안: CALIP = CLIP + parameter-free attention module
    - 핵심: visual & textual rep가 서로 interact하도록!
- 두 가지 버전
    - parameter X: 파라미터 없이, 단지 내적으로 유사도 측정!
    - parameter O: 파라미터 있음 (간단)

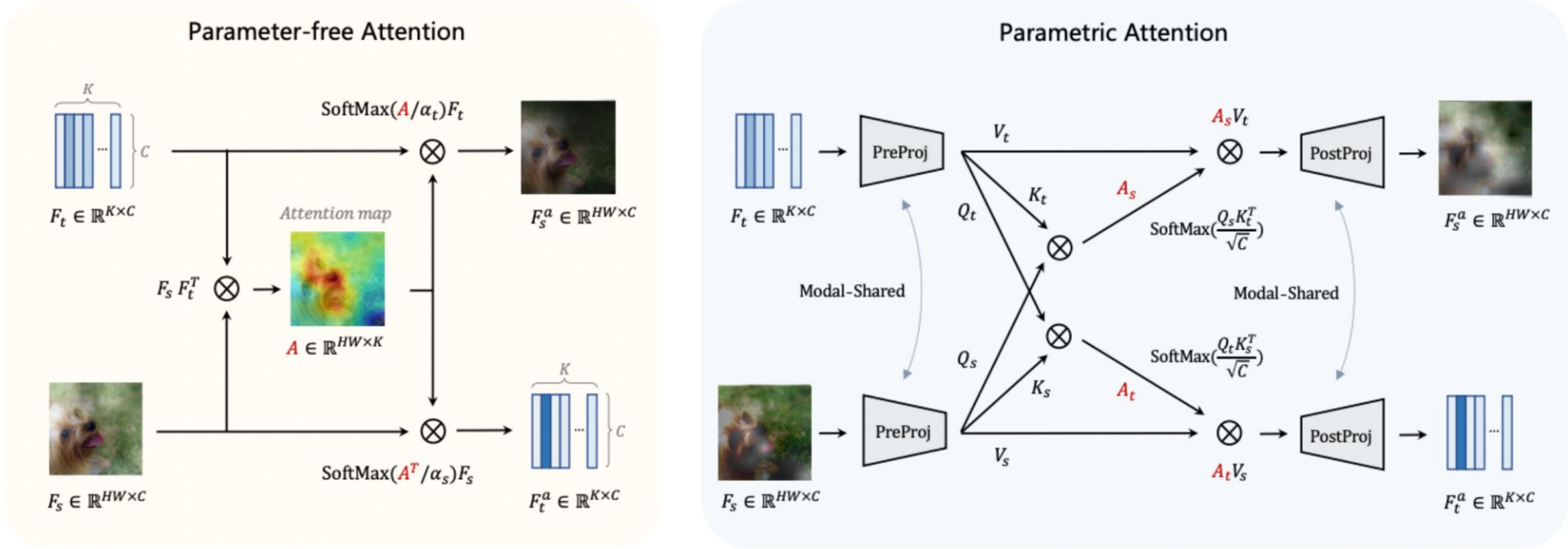
4. 기타 방법론들

[4] CALIP (AAAI 2023)



4. 기타 방법론들

[4] CALIP (AAAI 2023)



4. 기타 방법론들

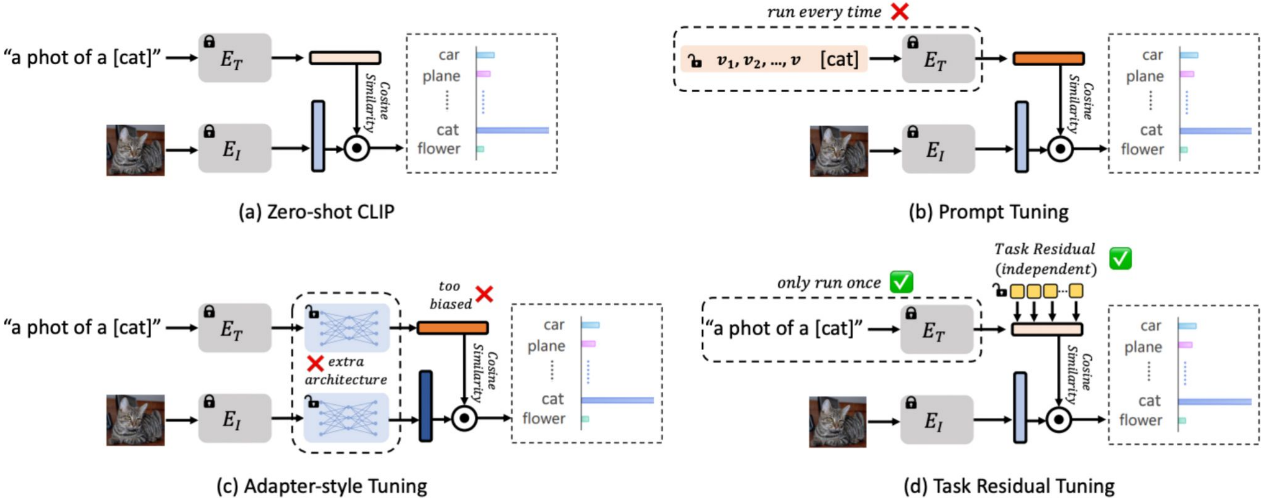
Set of tunable parameters $\mathbf{x} \in \mathbb{R}^{K \times D}$

[5] TaskRes (CVPR 2023)

New classifier $\mathbf{t}'$ for the target task:

$$\mathbf{t}' = \mathbf{t} + \alpha \mathbf{x}$$

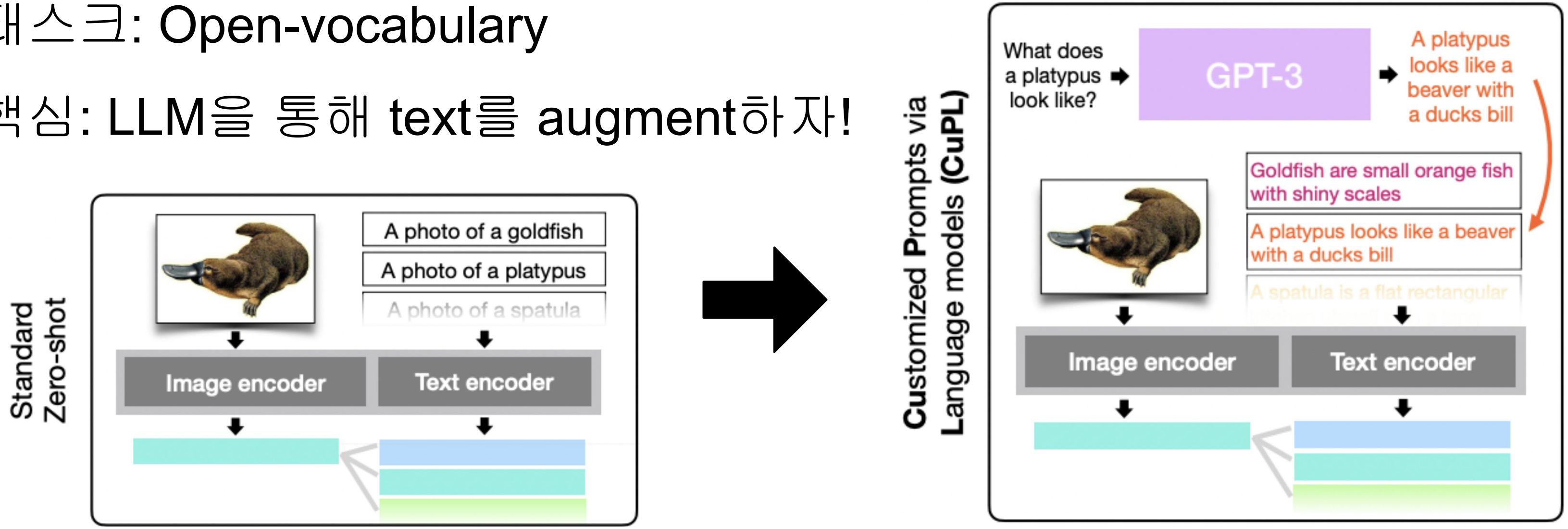
- TaskRes = Task Residual Tuning
- Prior knowledge & New knowledge를 완전히 decouple 시킴



4. 기타 방법론들

[6] CuPL (ICCV 2023)

- CuPL = “Customized” prompts via “Language” model
- 태스크: Open-vocabulary
- 핵심: LLM을 통해 text를 augment하자!



4. 기타 방법론들

[6] CuPL (ICCV 2023)

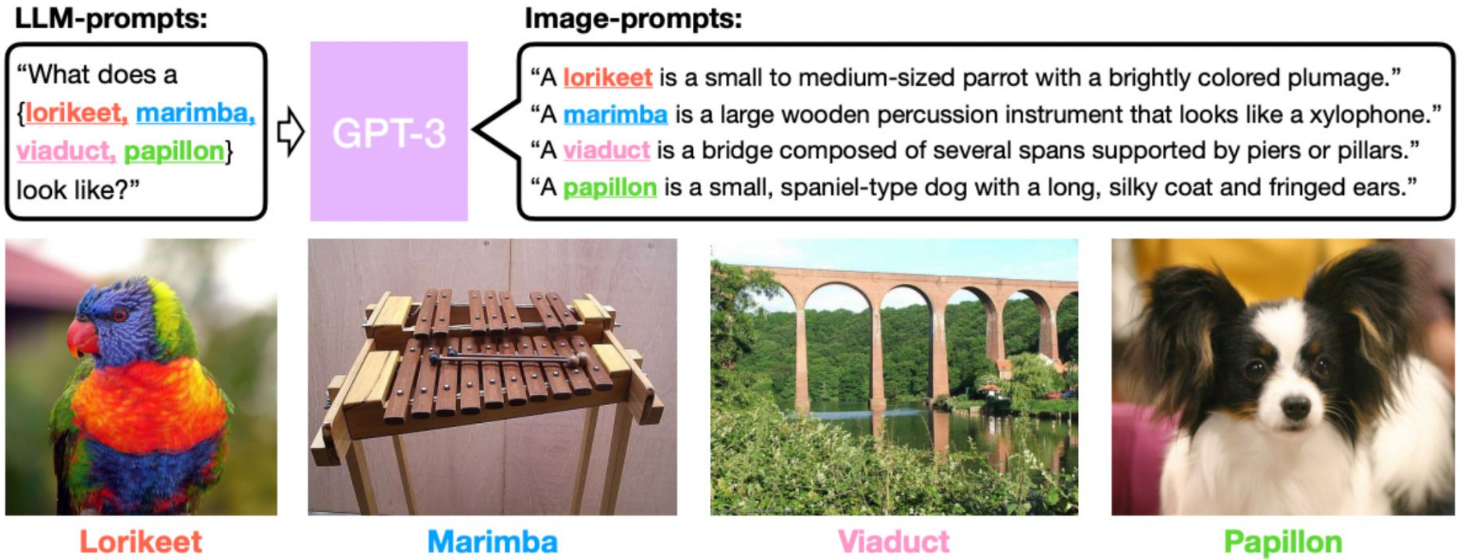


Figure 2. **Example CuPL LLM-prompts and Image-prompts.** LLM-prompts are filled in with a class name and then used as input to GPT-3, which then outputs image-prompts. Example LLM generated image-prompts and associated images from ImageNet are shown. Only image-prompts are used for the downstream image classification.

4. 기타 방법론들

[7] VCD (ICLR 2023)

- VCD = “V”isual “C”lassification via “D”escription from LLMs
- CLIP의 한계점
  - (상세 설명이 아닌) 단순히 class만을 맞힌다
  - 왜 해당 class가 선택되었는지에 대한 설명이 없다.
- 해결책: Classification by de

- $s(c, x) = \frac{1}{|D(c)|} \sum_{d \in D(c)} \phi(d, x).$
  - $D(c)$: Set of descriptors for the category c
  - $\phi(d, x)$: Log probability that descriptor d pertains to the image x
    - Represent d via a natural language sentence
- Classification: $\arg\max s(c, x)$

4. 기타 방법론들

[7] VCD (ICLR 2023)

- $s(c, x) = \frac{1}{|D(c)|} \sum_{d \in D(c)} \phi(d, x).$
  - $D(c)$: Set of descriptors for the category c
  - $\phi(d, x)$: Log probability that descriptor d pertains to the image x.
    - Represent d via a natural language sentence
- Classification: $\arg\max_{c \in C} s(c, x)$

4. 기타 방법론들

[7] VCD (ICLR 2023)

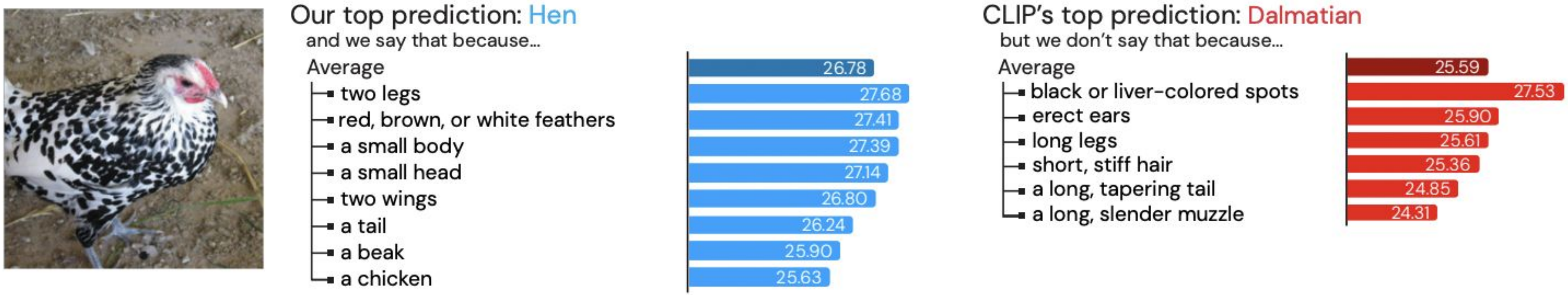


Figure 1: On the left, we show an example decision by our model in addition to its justification (blue bars). On the right, we show how CLIP classifies this image. Our model does not make the same mistake because it cannot produce a compatible justification with the image (red bars).

Visual classification via description from large language model

4. 기타 방법론들

[7] VCD (ICLR 2023)

```
Q: What are useful visual features for distinguishing a lemur in a photo?
A: There are several useful visual features to tell there is a lemur in a
    photo:
- four-limbed primate
- black, grey, white, brown, or red-brown
- wet and hairless nose with curved nostrils
- long tail
- large eyes
- furry bodies
- clawed hands and feet
```

School bus
- large, yellow vehicle
- the words "school bus" written on the side
- a stop sign that deploys from the side of the bus
- flashing lights on the top of the bus
- large windows

Shoe store
- a building with a sign that says "shoe store"
- a large selection of shoes in the window
- shoes on display racks inside the store
- a cash register
- a salesperson or customer

Volcano
- a large, cone-shaped mountain
- a crater at the top of the mountain
- lava or ash flowing from the crater
- a plume of smoke or ash rising from the crater

Barber shop
- a building with a large, open storefront
- a barber pole or sign outside the shop
- barber chairs inside the shop
- mirrors on the walls
- shelves or cabinets for storing supplies
- a cash register
- a waiting area for customers

Cheeseburger
- a burger patty
- cheese
- a bun
- lettuce
- tomato
- onion
- pickles
- ketchup
- mustard

Violin
- a stringed instrument
- typically has four strings
- a wooden body
- a neck and fingerboard
- tuning pegs
- a bridge
- a soundpost
- f-holes
- a bow

Pirate ship
- a large, sailing vessel
- a flag with a skull and crossbones
- cannons on the deck
- a wooden hull
- portholes
- rigging
- a crow's nest