

VLM 끄적 끄적 6

Transfer Learning (1) Prompt Tuning

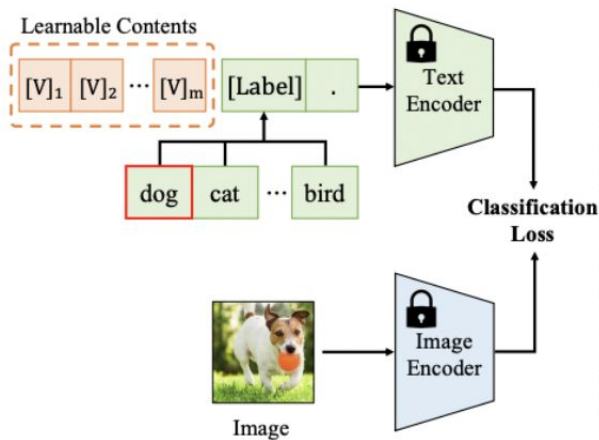
Prompt Tuning (PT)의 개요

(1) 필요성: 전체 VLM을 fine-tuning할 필요 없다.

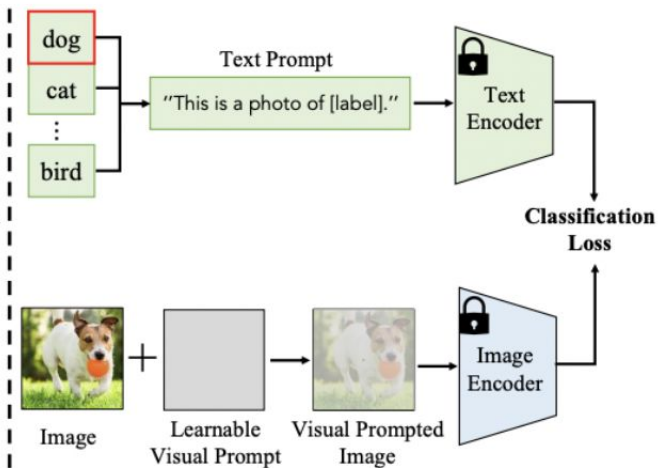
(2) 기존 연구: Prompt를 manual하게 design한다

(3) 세 갈래

- Text PT
- Visual PT
- Text-visual PT



Text PT (a)



(b) Visual PT

Prompt Tuning (PT)의 세 갈래

1. Text PT

- CoOp (IJCV 2022), CoCoOp (CVPR 2022), SubPT (arxiv 2023), LASP (CVPR 2023), VPT (ICLRW 2023), KgCoOp (CVPR 2023), SoftCPT (arxiv 2022), PLOT (ICLR 2023), DualCoOp (NeurIPS 2022), Tal-DP (CVPR 2023), DenseCLIP (CVPR 2022), ProTeCt (CVPR 2024), UPL (arxiv 2022), TPT (NeurIPS 2022)

2. Visual PT

- VP (arxiv 2022), RePrompt (arxiv 2024)

3. Text-visual PT

- UPT (arxiv 2022), MVLPT (WACV 2024), MaPLe (CVPR 2023), CAVPT (arxiv 2023)

Prompt Tuning (PT)의 세 갈래

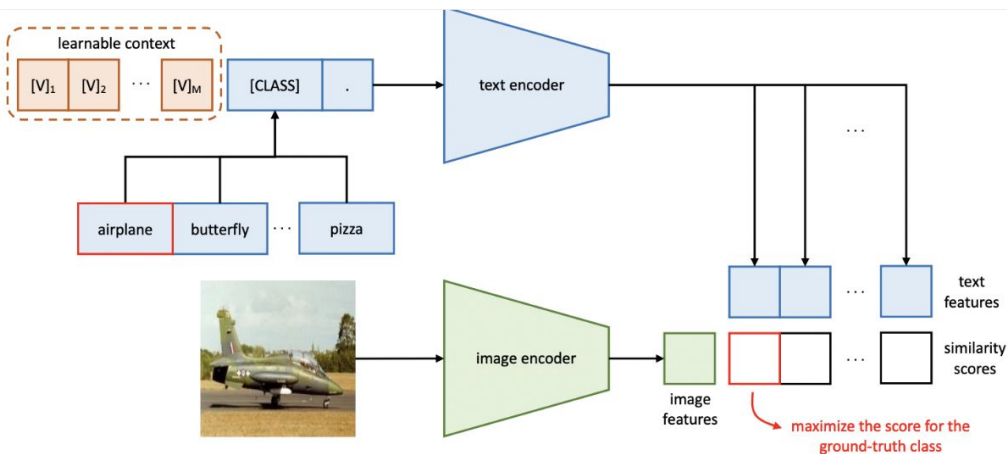
TABLE 4: Summary of VLM transfer learning methods. TPT: text-prompt tuning; VPT: visual-prompt tuning; FA: feature adapter; CA: cross-attention; FT: fine-tuning; AM: architecture modification; LLM: large-language model. [code] directs to code websites.


Method	Category	Setup	Contribution
CoOp [31] [code]	TPT	Few-shot Sup.	Introduce context optimization with learnable text prompts for VLM transfer learning.
CoCoOp [32] [code]	TPT	Few-shot Sup.	Propose conditional text prompting to mitigate overfitting in VLM transfer learning.
SubPT [132] [code]	TPT	Few-shot Sup.	Propose subspace text prompt tuning to mitigate overfitting in VLM transfer learning.
LASP [133]	TPT	Few-shot Sup.	Propose to regularize the learnable text prompts with the hand-engineered prompts.
ProDA [134]	TPT	Few-shot Sup.	Propose prompt distribution learning that captures the distribution of diverse text prompts.
VPT [135]	TPT	Few-shot Sup.	Propose to model the text prompt learning with instance-specific distribution.
ProGrad [136] [code]	TPT	Few-shot Sup.	Present a prompt-aligned gradient technique for preventing knowledge forgetting.
CPL [137] [code]	TPT	Few-shot Sup.	Employ counterfactual generation and contrastive learning for text prompt tuning.
PLOT [138] [code]	TPT	Few-shot Sup.	Introduce optimal transport to learn multiple comprehensive text prompts.
DualCoOp [139] [code]	TPT	Few-shot Sup.	Introduce positive and negative text prompt learning for multi-label classification.
TaL-DPT [140] [code]	TPT	Few-shot Sup.	Introduce a double-grained prompt tuning technique for multi-label classification
SoftCPT [141] [code]	TPT	Few-shot Sup.	Propose to fine-tune VLMs on multiple downstream tasks simultaneously.
DenseClip [142] [code]	TPT	Supervised	Propose a language-guided fine-tuning technique for dense visual recognition tasks.
UPL [143] [code]	TPT	Unsupervised	Propose unsupervised prompt learning with self-training for VLM transfer learning.
TPT [144] [code]	TPT	Unsupervised	Propose test-time prompt tuning that learns adaptive prompts on the fly.
KgCoOp [145] [code]	TPT	Few-shot Sup.	Introduce knowledge-guided prompt tuning to improve the generalization ability.
ProTeCt [146]	TPT	Few-shot Sup.	Propose a prompt tuning technique to improve consistency of model predictions.
VP [147] [code]	VPT	Supervised	Investigate the efficacy of visual prompt tuning for VLM transfer learning.
RePrompt [148]	VPT	Few-shot Sup.	Introduce retrieval mechanisms to leverage knowledge from downstream tasks.
UPT [149] [code]	TPT, VPT	Few-shot Sup.	Propose a unified prompt tuning that jointly optimizes text and image prompts.
MVLP [150] [code]	TPT, VPT	Few-shot Sup.	Incorporate multi-task knowledge into text and image prompt tuning.
MaPLE [151] [code]	TPT, VPT	Few-shot Sup.	Propose multi-modal prompt tuning with a mutual promotion strategy.
CAVPT [152] [code]	TPT, VPT	Few-shot Sup.	Introduce class-aware visual prompt for concentrating more on visual concepts.
Clip-Adapter [33] [code]	FA	Few-shot Sup.	Introduce an adapter with residual feature blending for efficient VLM transfer learning.
Tip-Adapter [34] [code]	FA	Few-shot Sup.	Propose to build a training-free adapter with the embeddings of few labelled images.
SVL-Adapter [153] [code]	FA	Few-shot Sup.	Introduce a self-supervised adapter by performing self-supervised learning on images.
SuS-X [154] [code]	FA	Unsupervised	Propose a training-free name-only transfer learning paradigm with curated support sets.
CLIPPR [155] [code]	FA	Unsupervised	Leverage the label distribution priors for adapting pre-trained VLMs.
SgVA-CLIP [156]	TPT, FA	Few-shot Sup.	Propose a semantic-guided visual adapter to generate discriminative adapted features.
VT-Clip [157]	CA	Few-shot Sup.	Introduce visual-guided attention that semantically aligns text and image features.
CALIP [158] [code]	CA	Unsupervised	Propose parameter-free attention for the communication between visual and textual features.
TaskRes [159] [code]	CA	Few-shot Sup.	Propose a technique for better learning old VLM knowledge and new task knowledge.
CuPL [160]	LLM	Unsupervised	Employ large language models to generate customized prompts for VLMs.
VCD [161]	LLM	Unsupervised	Employ large language models to generate captions for VLMs.
Wise-FT [162] [code]	FT	Supervised	Propose ensemble-based fine-tuning by combining the fine-tuned and original VLMs.
MaskClip [163] [code]	AM	Unsupervised	Propose to extract dense features by modifying the image encoder architecture.
MUST [164] [code]	Self-training	Unsupervised	Propose masked unsupervised self-training for unsupervised VLM transfer learning.

1. Text PT

[1] CoOp (IJCV 2022)

- Context Optimization = CoOp

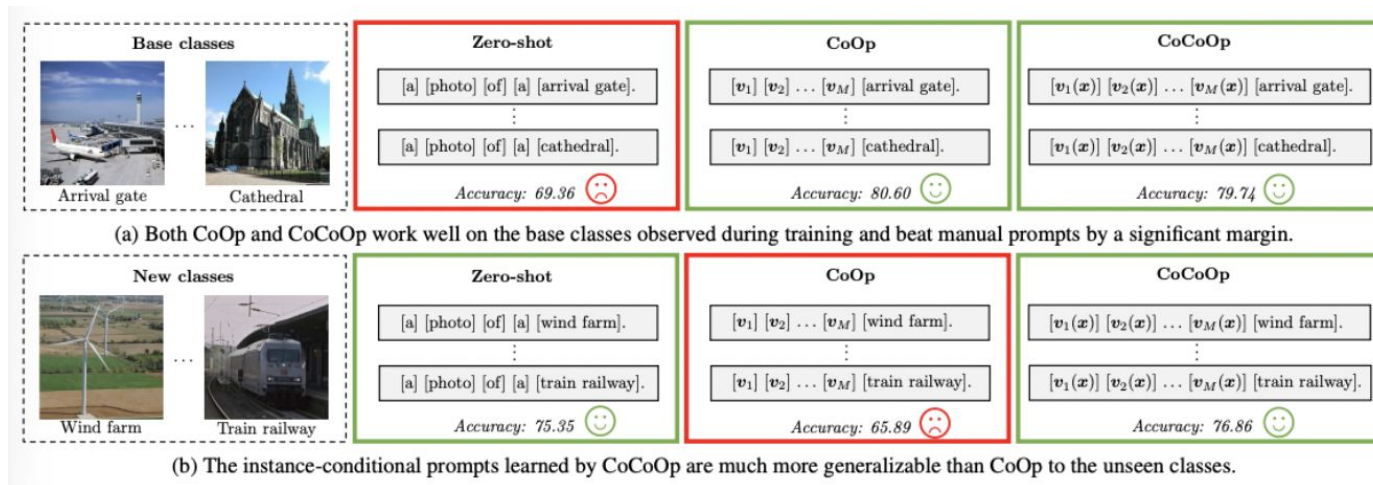


Caltech101	Prompt	Accuracy
	a [CLASS].	82.68
	a photo of [CLASS].	80.81
	a photo of a [CLASS].	86.29
	$[V]_1 [V]_2 \dots [V]_M [CLASS]$.	91.83

1. Text PT

[2] CoCoOp (CVPR 2022)

- CoCoOp = Conditional CoOp
- Generates a specific prompt “for each image”



1. Text PT

[3] SubPT (arxiv 2022)

- SubPT = Subspace Prompt Tuning

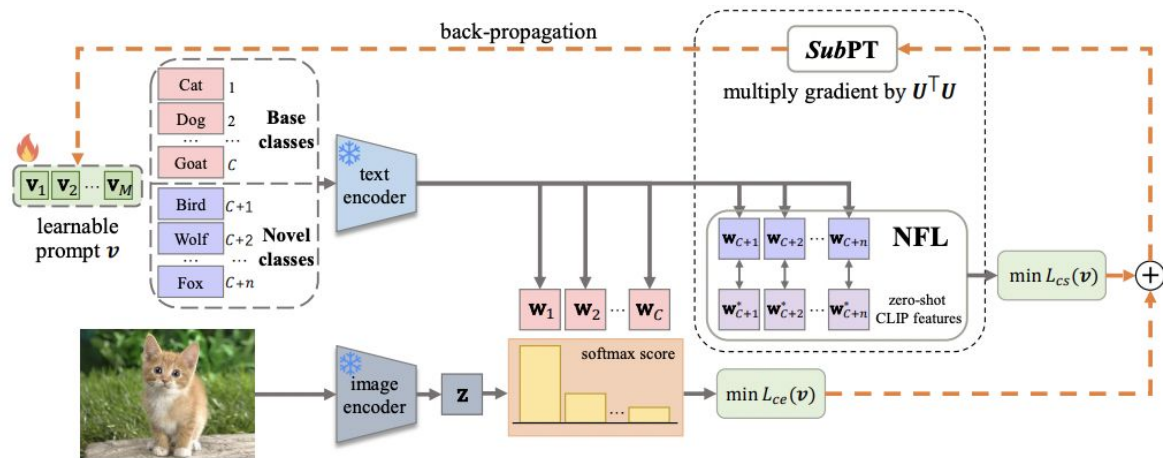
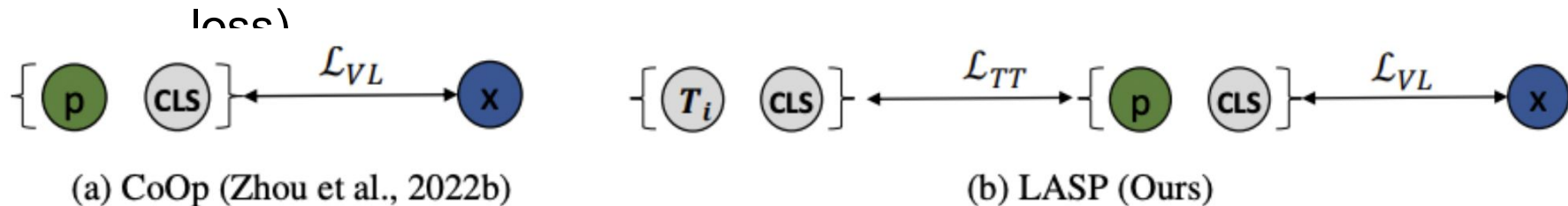


Fig. 3. **Overview of the proposed Subspace Prompt Tuning (SubPT) and Novel Feature Learner (NFL)**, surrounded by the black dotted box. To eliminate spurious components and mitigate overfitting, we project the gradient onto the low-rank **subspace** spanned by the dominant eigenvectors U of early-stage gradient flow during back-propagation. NFL learns text features towards zero-shot CLIP features on novel categories to enhance the generalization ability of the learned prompt embedding beyond the training set.

1. Text PT

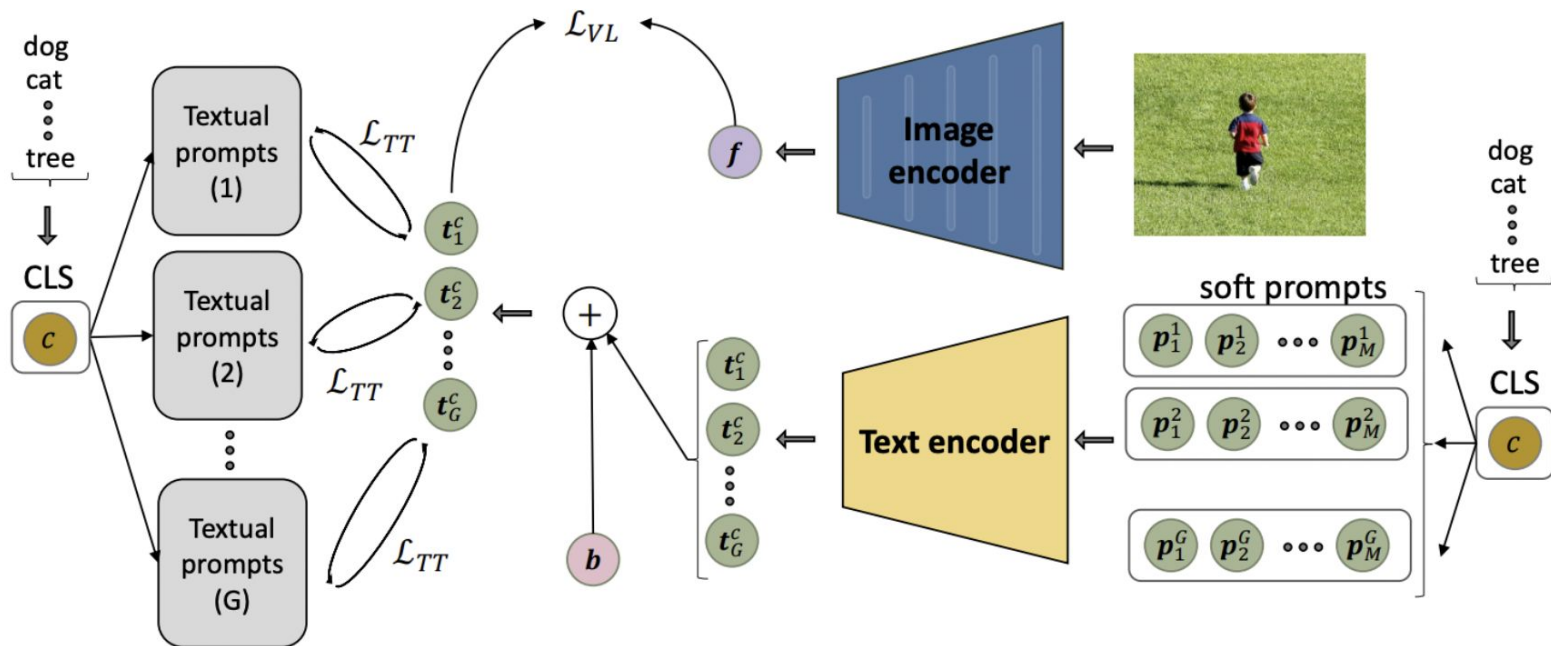
[4] LASP (CVPR 2023)

- 최근 트렌드: “SOFT” prompt learning (SP)
- 한계점: Overfitting
- 해결책: LASP = “Language-Aware” SP
 - Learned prompt & Hand-crafted prompt 사이의 규제를 통해! (CE



1. Text PT

[4] LASP (CVPR 2023)

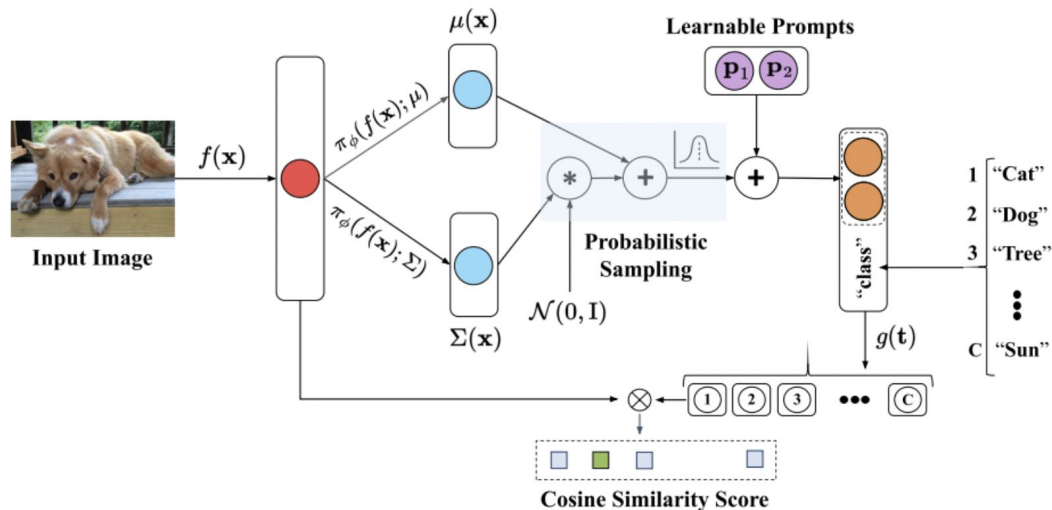


1. Text PT

[5] VPT (ICLRW 2023) $\mathbf{p}_\gamma(\mathbf{x}) = [\mathbf{p}_1 + \mathbf{r}_\gamma, \mathbf{p}_2 + \mathbf{r}_\gamma, \dots, \mathbf{p}_L + \mathbf{r}_\gamma], \mathbf{r}_\gamma \sim p_\gamma(\mathbf{x})$

$$\mathbf{r}(\mathbf{x}) \sim \mathcal{N}(\mu(\mathbf{x}), \Sigma(\mathbf{x}))$$

- VPT = VariationalPT
- Text prompt를 “instance-specific distribution”으로 모델링
- 총 L 개의 learnable prompt



1. Text PT

[6] KgCoOp (CVPR 2023)

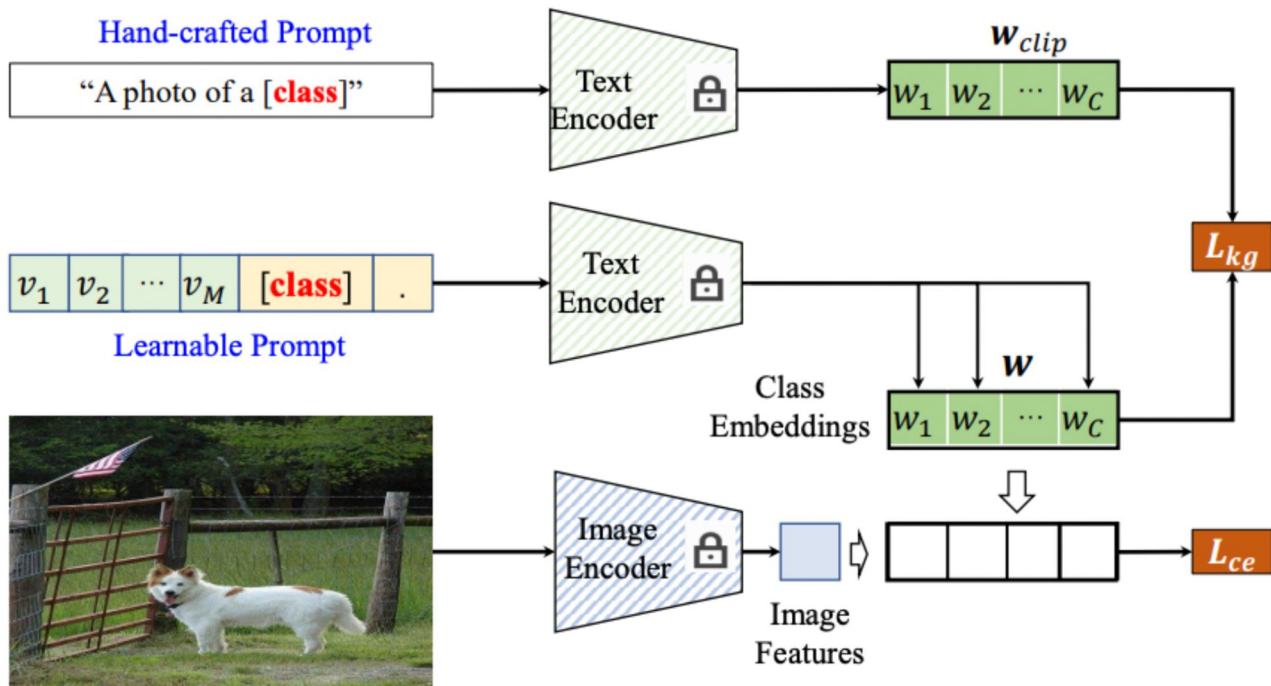
- Kg = Knowledge-guided
- Enhances the generalization of unseen class
 - By mitigating the forgetting of “textual knowledge”
- How? (a) Learnable prompt & (b) Hand-crafted prompt 규제
 - LASP (CVPR 2023)과 유사

$$\mathcal{L} = \mathcal{L}_{ce} + \lambda \mathcal{L}_{kg}.$$

$$\cdot \mathcal{L}_{kg} = \frac{1}{N_c} \sum_{i=1}^{N_c} \left\| \mathbf{w}_i - \mathbf{w}_i^{clip} \right\|_2^2$$

1. Text PT

[6] KgCoOp (CVPR 2023)



1. Text PT

[7] SoftCPT (arxiv 2022)

- 핵심: Soft context sharing
 - 기본 아이디어: 많은 task들은 서로 correlated 되어 있다. sharing하자!
 - Proposal
 - (1) Fine-tune on “multiple tasks” jointly
 - (2) “Task-shared” meta network
 - 각 Task에 대한 prompt context
- = a) Task 명 + b) Learnable Task context

1. Text PT

[8] PLOT (ICLR 2023)

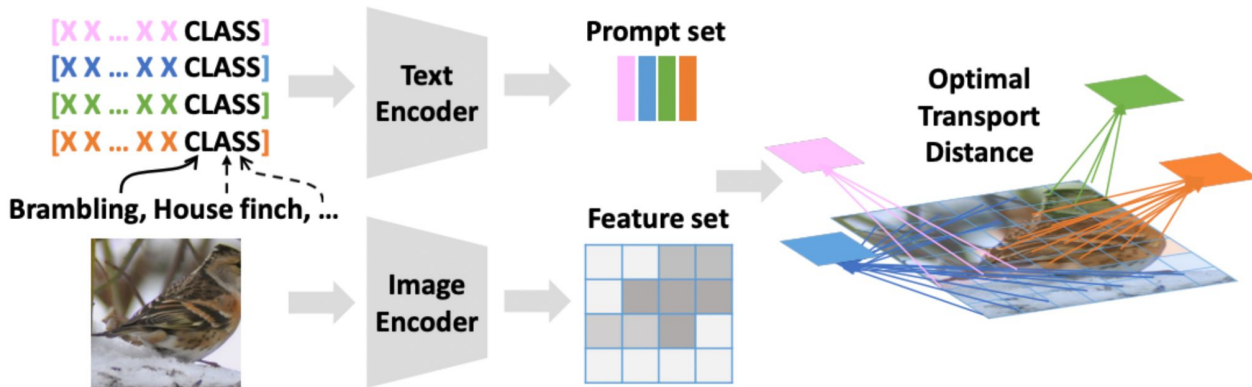
- 기존: 1 종류의 prompt

PLOT: N 종류의 prompt

- N개가 1개로 collapse되는 문제: Optimal transport로써 해결!



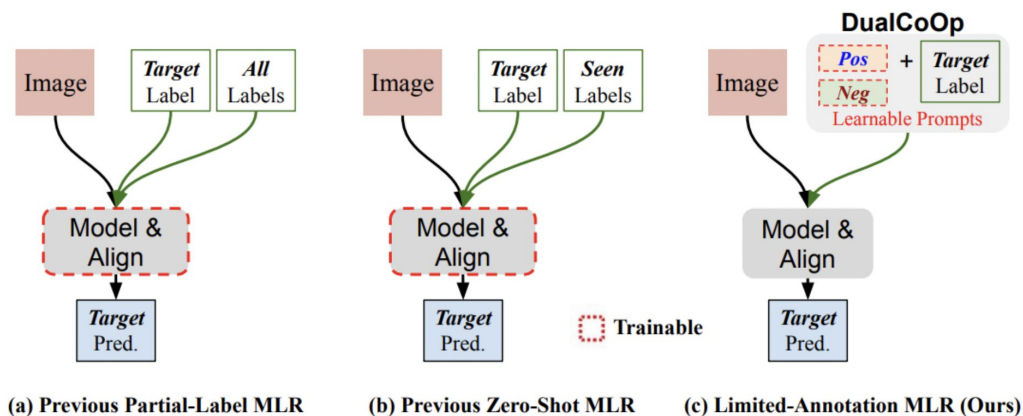
Figure 1: The motivation that one category can be complementarily described in different views (An example of “Brambling”).



1. Text PT

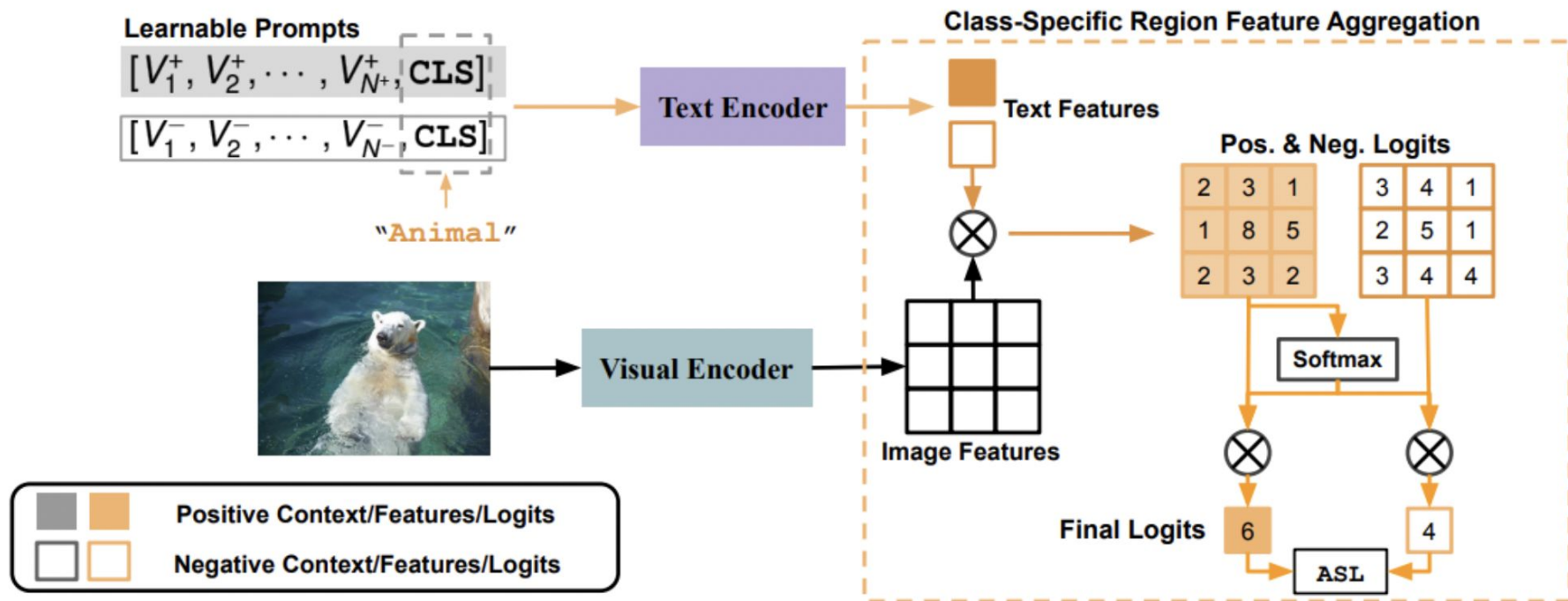
[9] DualCoOp (NeurIPS 2022)

- 태스크: Multi-label Recognition (MLR)
 - 하나의 이미지에 여러 개의 (미정의) **label**이 있을 수 있음!
- 특정 **prompt**에 대해, 긍정 & 부정의 개념이 각각, 총 2개씩 존재!



1. Text PT

[9] DualCoOp (NeurIPS 2022)



1. Text PT

[10] Tai-DP (CVPR 2023)

- 태스크: Multi-label Recognition (MLR)
- 핵심: “Text as Image” (TAI)
- Double-grained prompt tuning (coarse & fine-grained embeddings)
- Details:
 - Training 시, Text Encoder만을 사용하여 prompt를 학습한다.
 - 근거: Text descriptions are easy to collect!
 - Inference 시, text description을 image로 대체!

1. Text PT

[10] Tai-DP (CVPR 2023)

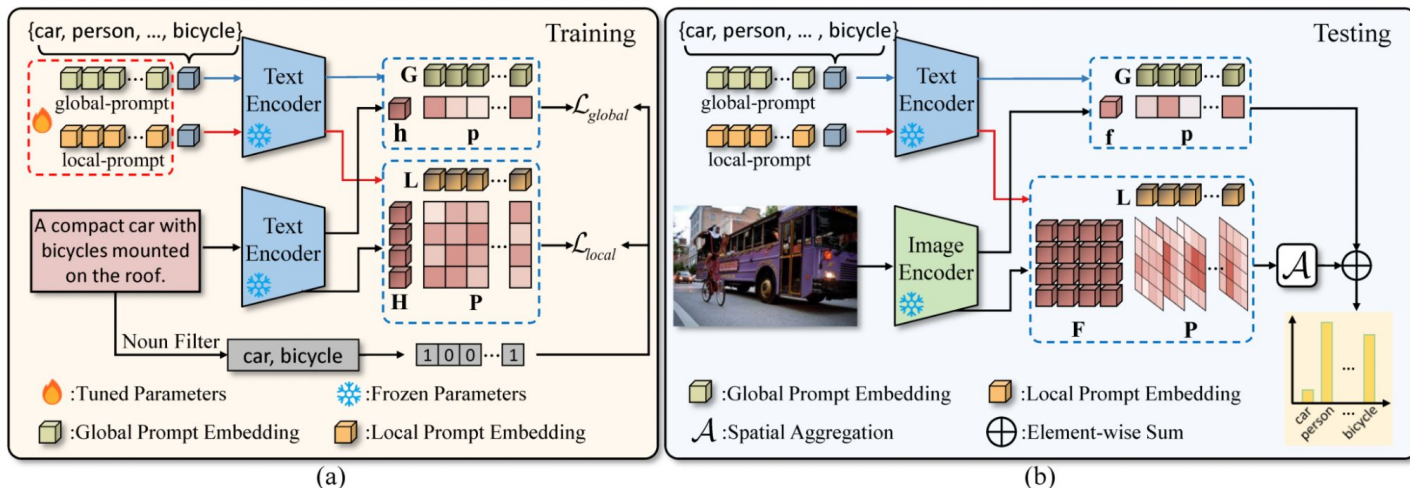
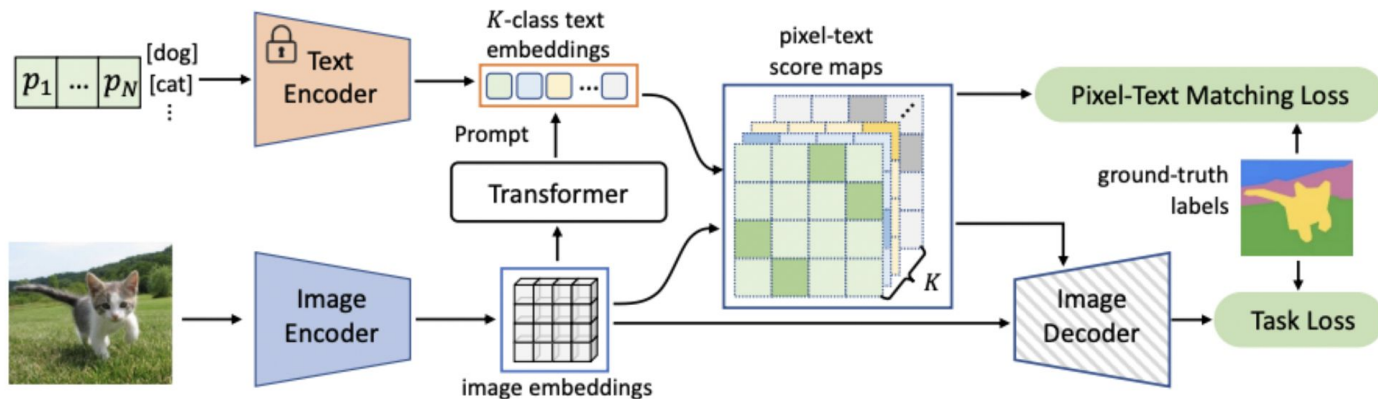


Figure 2. Training and testing pipeline of our proposed Text-as-Image (TaI) prompting, where we use text descriptions instead of labeled images to train the prompts. (a) During training, we use two identical text encoders from pre-trained CLIP to extract the global & local class embeddings (G&L) and overall & sequential text embeddings (h&H) respectively from the prompts and text description. The corresponding cosine similarity (p&P) between the embeddings are guided by the derived pseudo labels with ranking loss. (b) During testing, we replace the input from text descriptions to images. The global and local class embeddings can discriminate target classes from global & local image features (f&F). The final classification results are obtained by merging the scores of the two branches.

1. Text PT

[11] DenseCLIP (CVPR 2022)

- 태스크: Dense prediction
- 핵심: Language guided dense prediction!
- 기존: Image-text matching <-> 제안: Pixel-text matching



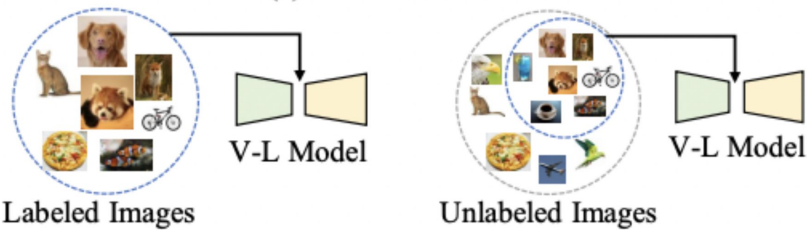
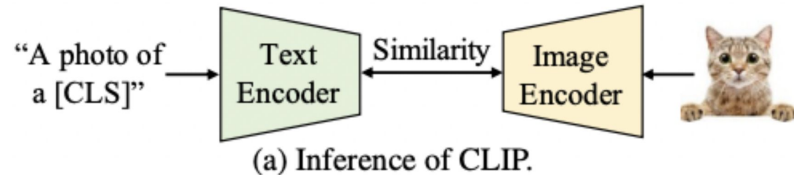
1. Text PT

[12] UPL (arxiv 2022)

- UPL = Unsupervised PL
- 기존 연구 한계점: “Labeled” data 필요
- 핵심: Self-training on pseudo-label

$$p_c = \frac{\exp(\langle f_c^{\text{text}}, f^{\text{image}} \rangle / \tau)}{\sum_{j=1}^C \exp(\langle f_j^{\text{text}}, f^{\text{image}} \rangle / \tau)}$$

$$\hat{y} = \underset{c}{\operatorname{argmax}} p_c.$$



(b) Existing methods adapt V-L model on labeled images.

(c) Ours. Adapt V-L model on selected unlabeled images.

1. Text PT

[12] UPL (arxiv 2022)

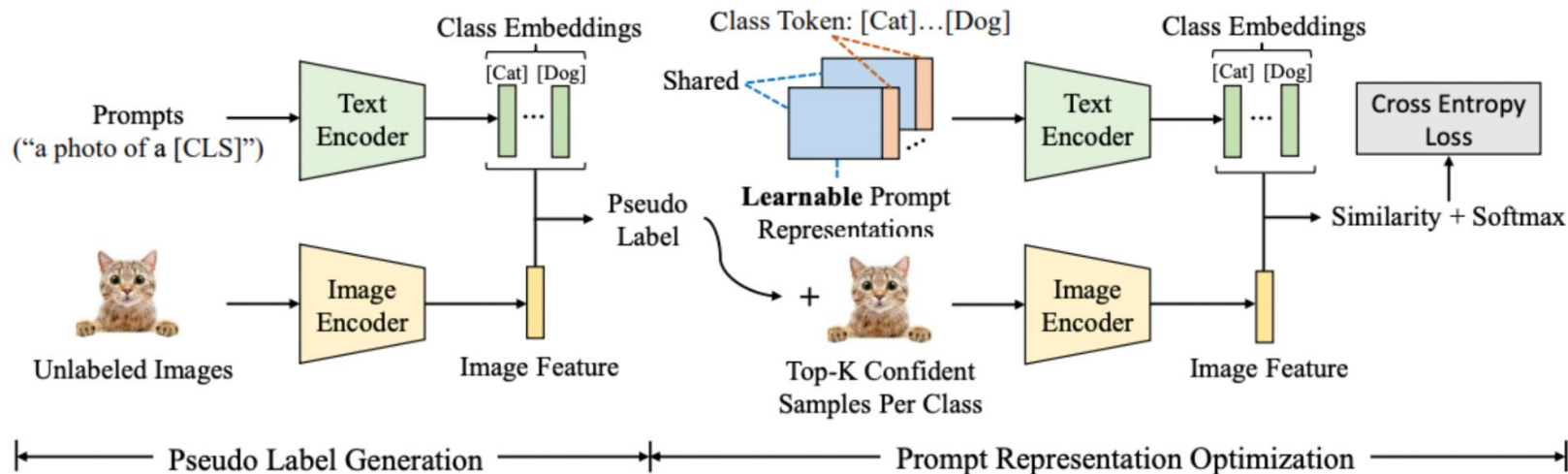


Figure 2: Overview of the proposed unsupervised prompt learning (UPL) framework. Our UPL mainly contains two parts, namely pseudo label generation and prompt representation optimization. We first use CLIP with a simple prompt (e.g., “a photo of a [CLS]”) to generate pseudo labels for target datasets and select top- K confident samples per class for subsequent training. Then we define a learnable prompt representation which is optimized on selected pseudo-labeled samples. For inference, we simply swap out the hand-crafted prompts with the well-optimized prompt representations.

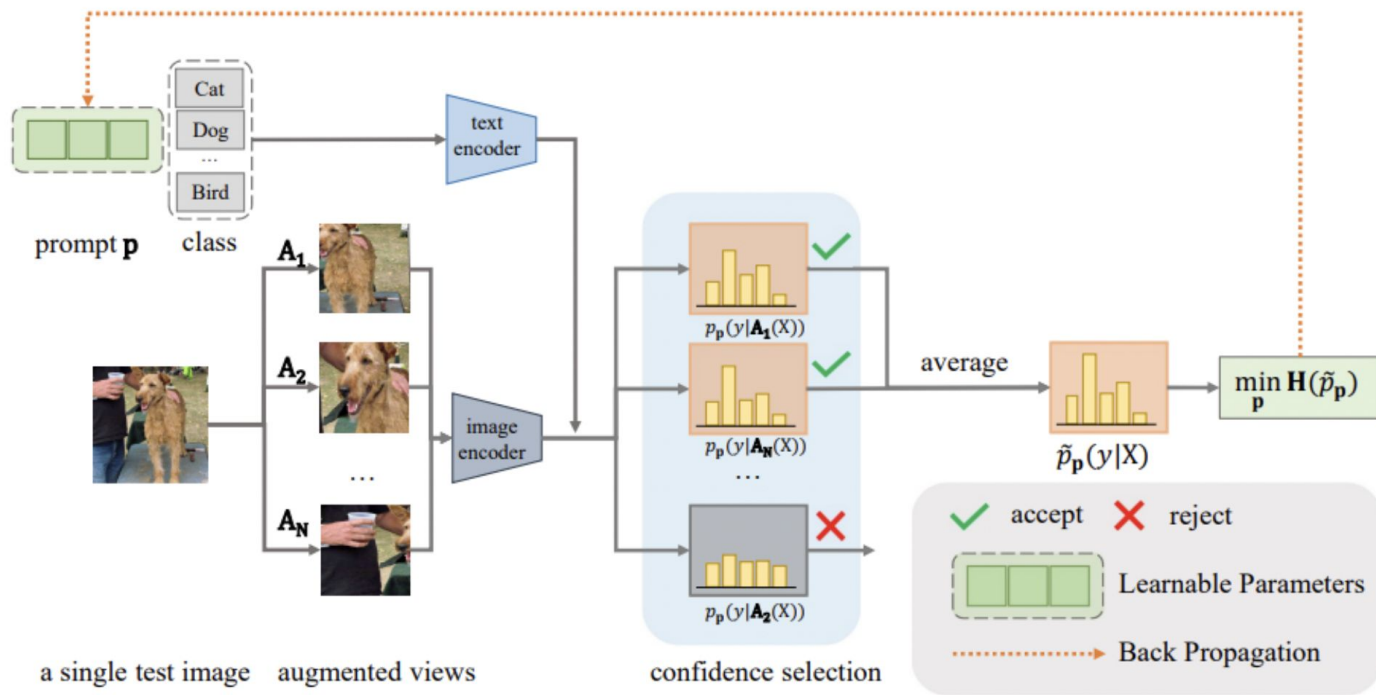
1. Text PT

[13] TPT (NeurIPS 2022)

- TPT = Test-time PT
- Goal: 개별 instance에 대한 adaptive prompt를 배우기!
- How? Confidence selection을 사용하여 entropy minimization
(= 동일 instance에 대한 multiple view는 서로 consistent해야!)

1. Text PT

[13] TPT (NeurIPS 2022)



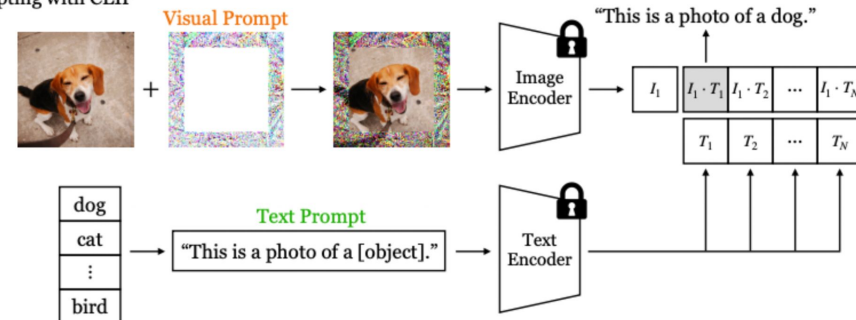
2. Visual PT

[1] VP (arxiv 2022)

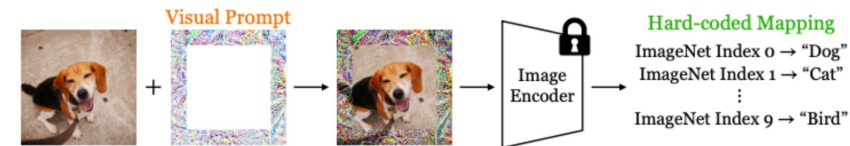
- VP = Visual prompting
- Learnable “image perturbation”

$$x^I \text{ by } x^I + v$$

(a) Prompting with CLIP



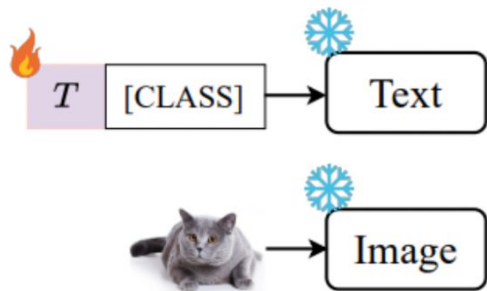
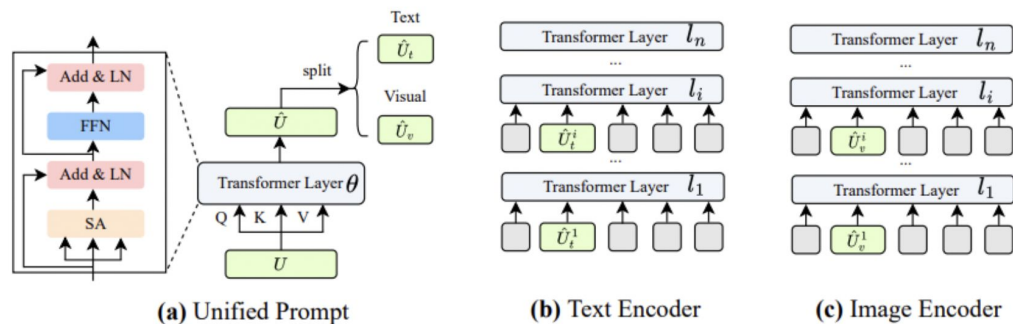
(b) Prompting (adversarial reprogramming) with vision models



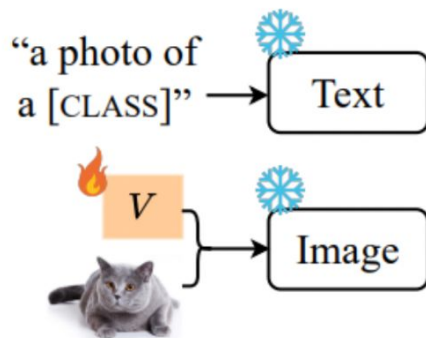
3. Text-Visual PT

[1] UPT (arxiv 2022)

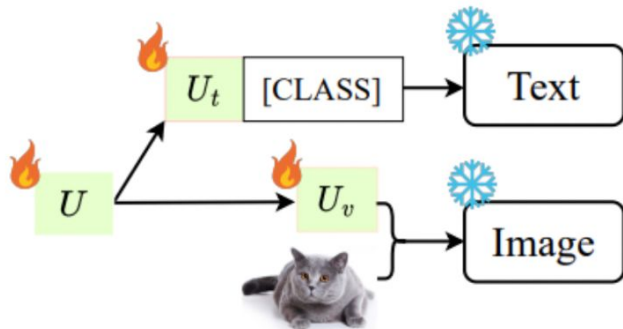
- U = Unified (V&L)
- Learn tiny NN to jointly optimize prompts across different modalities



(a) Text Prompt - CoOp



(b) Visual Prompt - VPT



(c) Unified Prompt - Ours

3. Text-Visual PT

[2] MVLPT (WACV 2024)

PI: Prompt Initialization

PA: Prompt Adaptation

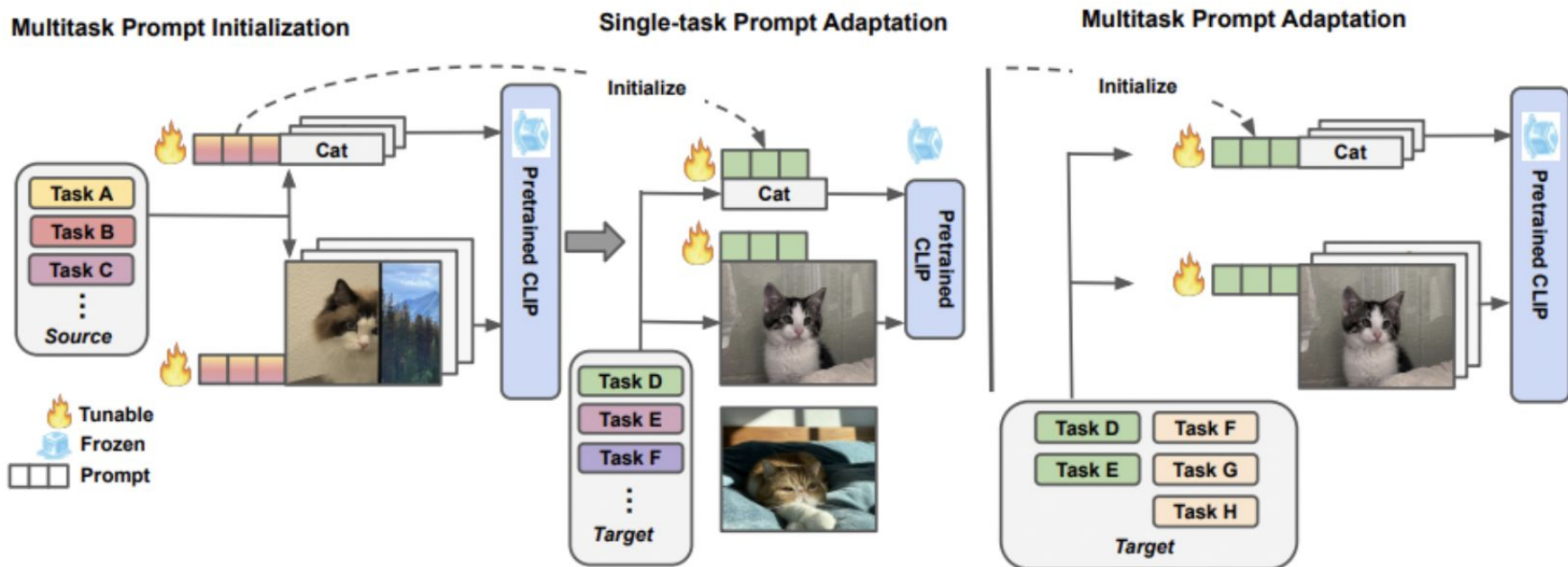
- M = Multi-task
- Cross-task knowledge를 text PT & image PT에 활용하자!
- Details
 - 1) Multi-task PI: single PT로 pretrain (shared)
 - 2-1) Single-task PA
 - 2-2) Multi-task PA

Learnable prompts: $\mathbf{U} = [\mathbf{U}_T, \mathbf{U}_V] \in \mathbb{R}^{d \times n}$ with length n

• where $\mathbf{U}_T \in \mathbb{R}^{d \times n_T}, \mathbf{U}_V \in \mathbb{R}^{d \times n_V}$

3. Text-Visual PT

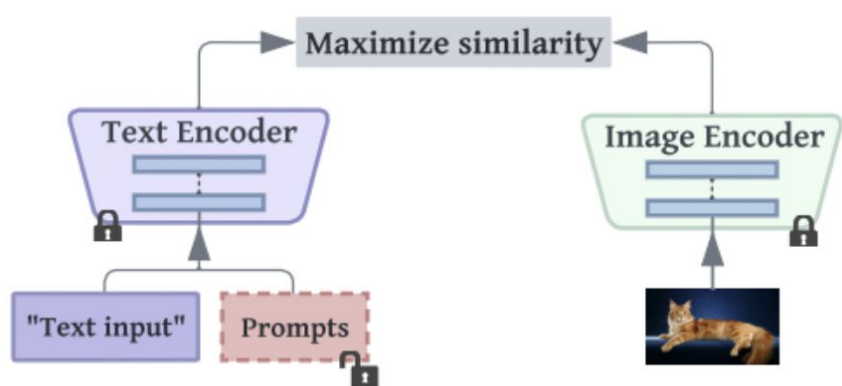
[2] MVLPT (WACV 2024)



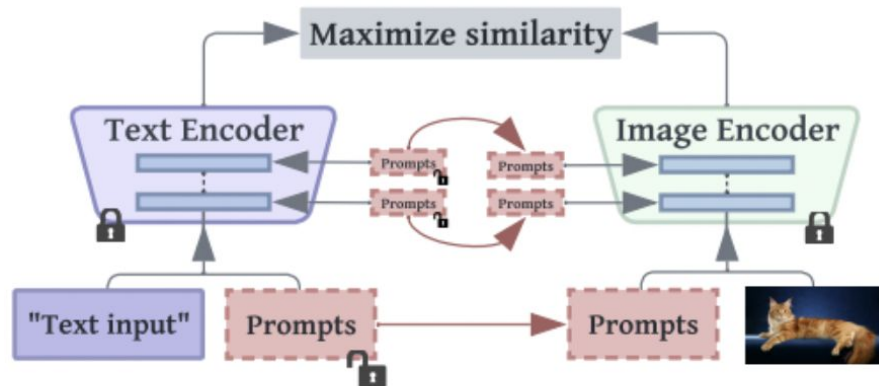
3. Text-Visual PT

[3] MaPLe (CVPR 2023)

- MaPLe = Multi-modal Prompt Learning
- 핵심: Text & Image 사이의 mutual promotion



(a) Existing prompt tuning methods (Uni-modal)



(b) Multi-modal Prompt Learning (MaPLe)

3. Text-Visual PT

[3] MaPLe (CVPR 2023)

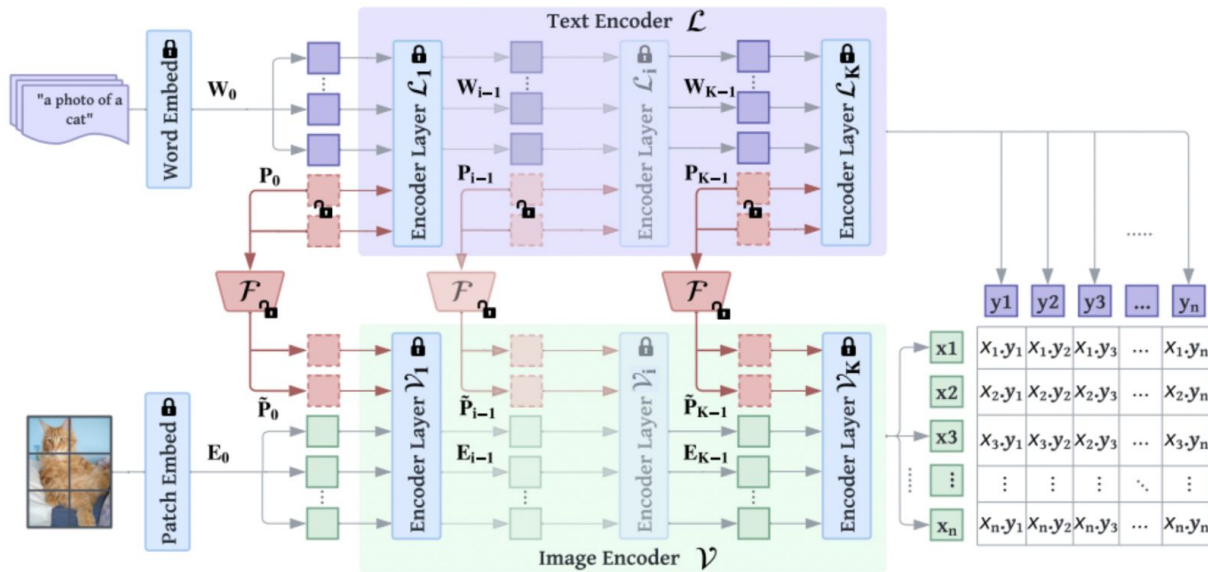


Figure 2. Overview of our proposed MaPLe (Multi-modal Prompt Learning) framework for prompt learning in V-L models. MaPLe tunes both vision and language branches where only the **context prompts** are learned, while the rest of the model is frozen. MaPLe conditions the vision prompts on language prompts via a V-L coupling function \mathcal{F} to induce mutual synergy between the two modalities. Our framework uses deep contextual prompting where separate context prompts are learned across multiple transformer blocks.