PromptKD: Unsupervised Prompt Distillation for Vision-Language Models

Zheng Li[1], Xiang Li[2,1*], Xinyi Fu[3], Xin Zhang[1], Weiqiang Wang[3], Shuo Chen[4], Jian Yang[1*]
[1] PCA Lab, VCIP, College of Computer Science, Nankai University
[2] NKIARI, Shenzhen Futian, [3] Tiansuan Lab, Ant Group, [4] RIKEN
{zhengli97, zhasion}@mail.nankai.edu.cn, {xiang.li.implus, csjyang}@nankai.edu.cn
{fxy122992, weiqiang.wwq}@antgroup.com, shuo.chen.ya@riken.jp

- Problem/Objective
  - CLIP + KD
  - Unsupervised distillation
- Contribution/Key Idea
  - domain-specific prompt-based knowledge distillation
  - reuse pre-stored text features
  - unlabeled로 student모델 학습

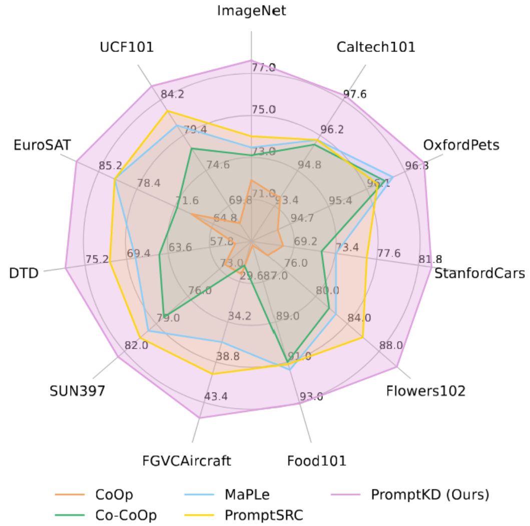


Figure 1. Harmonic mean (HM) comparison on base-to-novel generalization. All methods adopt the **ViT-B/16 image encoder** from the pre-trained CLIP model. PromptKD achieves state-of-the-art performance on 11 diverse recognition datasets.

PromptKD: Unsupervised Prompt Distillation for Vision-Language Models

• Remaining Method

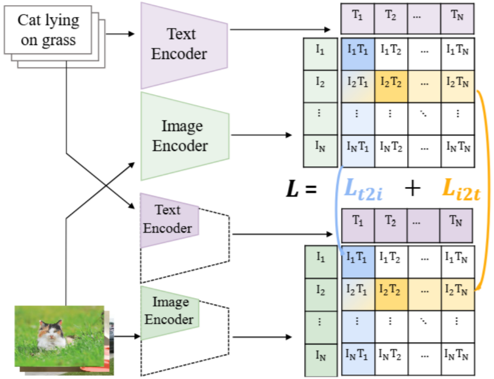


Figure 2. Affinity mimicking for language-image models. The loss includes image-to-text loss (yellow) and text-to-image loss (blue).

TinyCLIP [1]

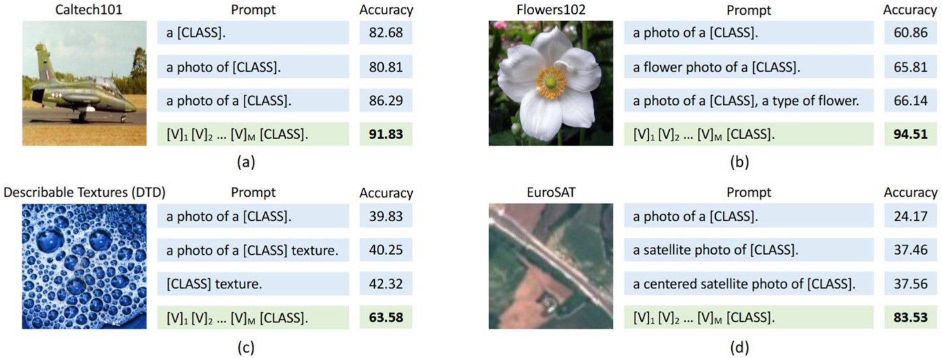


Fig. 1 Prompt engineering vs Context Optimization (CoOp). The former needs to use a held-out validation set for words tuning, which is inefficient; the latter automates the process and requires only a few labeled images for learning.

Prompt Learning [2]

- **기존 CLIP + KD 연구** → CLIP의 image/text encoder 다시 학습 → Cost ↑
  - → Distillation의 Cost를 줄여보자
- **기존 Prompt 연구** → Hard-to-soft Label로의 전환 등 prompt를 설계/학습
  - → prompt를 distillation 도구로 사용해보자

[1] Wu, Kan, et al. "Tinyclip: Clip distillation via affinity mimicking and weight inheritance." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.
[2] Zhou, Kaiyang, et al. "Learning to prompt for vision-language models." International Journal of Computer Vision 130.9 (2022): 2337-2348.

김범준

PromptKD: Unsupervised Prompt Distillation for Vision-Language Models

CVPR 2024

● Classic KD vs PromptKD

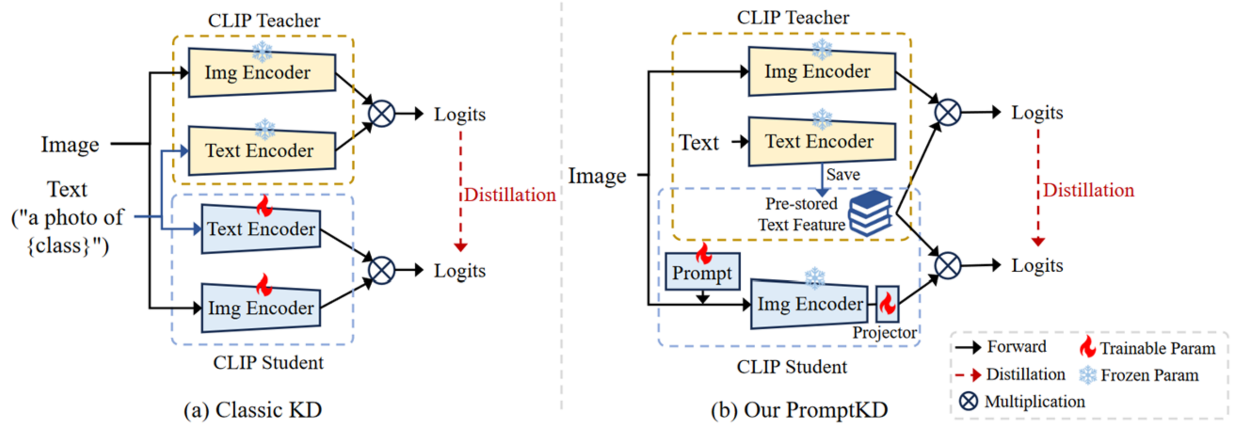
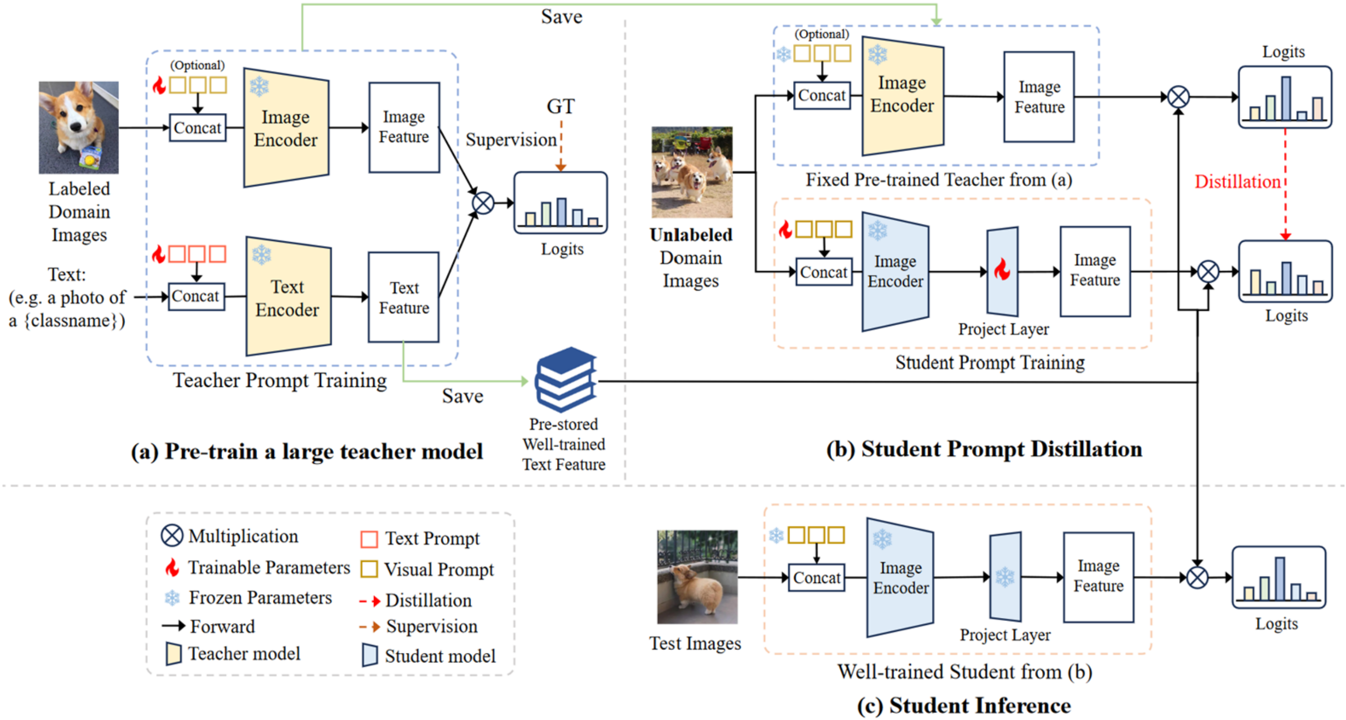


Figure 2. Architecture comparison between classic KD paradigm for CLIP (likewise CLIP-KD [44]) and our prompt distillation framework. (a) Classic KD methods perform distillation between independent teacher and student models. Students are typically fully fine-tuned by teachers' soft labels. (b) PromptKD breaks the rules of teacher-student independence. We propose to reuse the previously well-trained text features from the teacher pre-training stage and incorporate them into the student image encoder for both distillation and inference.

- Cost 감소
    - Unsupervised
    - Text feature saving
- Prompt 활용
    - Img Encoder Frozen → visual prompt + project

김범준

- **Method**

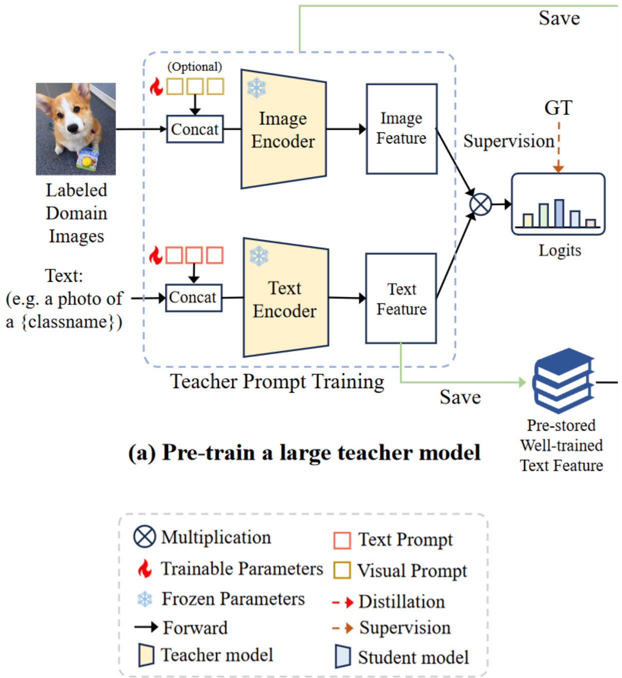


- 2 Stage Training

PromptKD: Unsupervised Prompt Distillation for Vision-Language Models

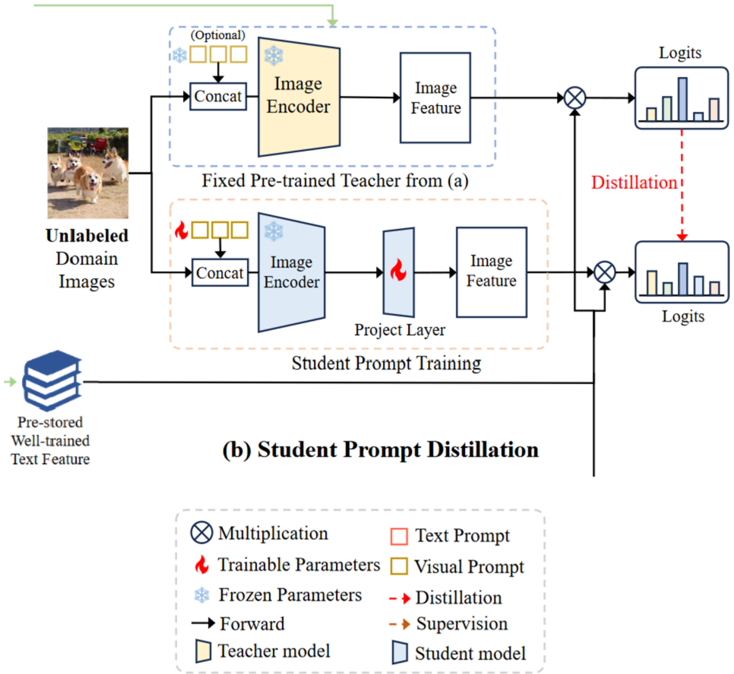
CVPR 2024

- **Method - Stage 1 (Teacher model Training)**



- Large CLIP 모델 → pretrain
- CLIP → **(+ domain-specific few-shot fine-tuning)** → teacher 완성
  - 16 few-shot 사용
- 각 class 이름 → prompt에 넣어서 class 별 feature vector 추출/저장

Method - Stage 2 (Student prompt distillation)



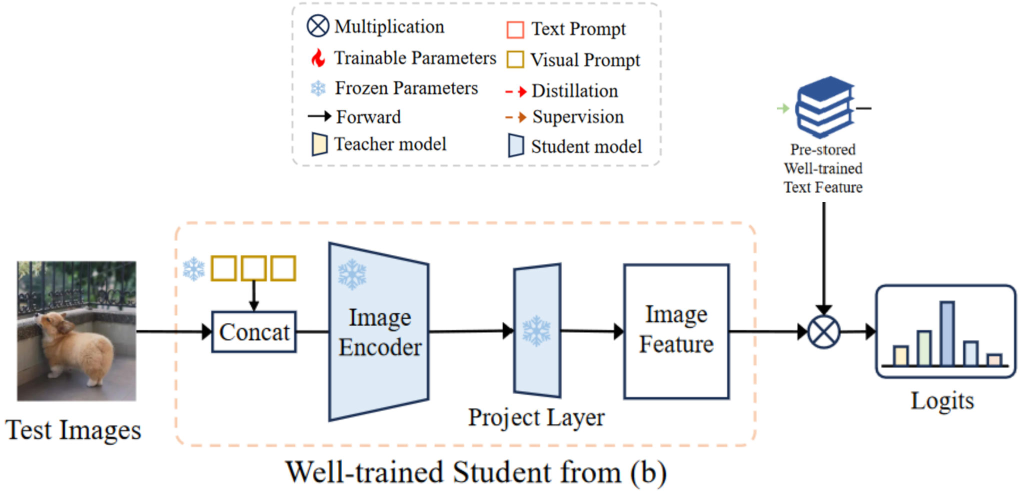
- Frozen Image encoder
  - 앞에 visual prompt 학습 (visual soft prompt)
  - 뒤에 project layer 학습 (MLP)
- Text r 사용

Algorithm 1 Pseudocode of PromptKD in PyTorch.

```
# tea_t: text encoder of teacher CLIP
# tea_i: image encoder of teacher CLIP
# stu_i: image encoder of student CLIP
# l_tea: teacher output logits
# l_stu: student output logits
# Proj: Feature Projector

# init
f_txt_t = tea_t(txt_of_all_classes)

# forward
for img in unlabeled_dataset:
    f_img_t = tea_i(img)
    f_img_s = stu_i(img)

    f_img_s = Proj(f_img_s)

    # get output predictions
    l_tea = f_img_t * f_txt_t.t()
    l_stu = f_img_s * f_txt_t.t()

    # calculate distillation loss
    loss = KLDivergence(l_stu, l_tea)
    loss.backward()
```

[PromptKD: Unsupervised Prompt Distillation for Vision-Language Models]

- # Method - Stage 3 (Inference)



(c) Student Inference

PromptKD: Unsupervised Prompt Distillation for Vision-Language Models

CVPR 2024

- **Experiment**

ViT-B/16	Base	Novel	HM
CLIP	69.34	74.22	71.70
CoOp	82.69	63.22	71.66
CoCoOp	80.47	71.69	75.83
MaPLe	82.28	75.14	78.55
PromptSRC	84.26	76.10	79.97
PromptKD	86.96	80.73	83.73
Δ	+2.70	+4.63	+3.76

(a) Average over 11 datasets.

ViT-B/16	Base	Novel	HM
CLIP	72.43	68.14	70.22
CoOp	76.47	67.88	71.92
CoCoOp	75.98	70.43	73.10
MaPLe	76.66	70.54	73.47
PromptSRC	77.60	70.73	74.01
PromptKD	80.83	74.66	77.62
Δ	+3.23	+3.93	+3.61

(b) ImageNet

ViT-B/16	Base	Novel	HM
CLIP	96.84	94.00	95.40
CoOp	98.00	89.81	93.73
CoCoOp	97.96	93.81	95.84
MaPLe	97.74	94.36	96.02
PromptSRC	98.10	94.03	96.02
PromptKD	98.91	96.65	97.77
Δ	+0.81	+2.62	+1.75

(c) Caltech101

ViT-B/16	Base	Novel	HM
CLIP	91.17	97.26	94.12
CoOp	93.67	95.29	94.47
CoCoOp	95.20	97.69	96.43
MaPLe	95.43	97.76	96.58
PromptSRC	95.33	97.30	96.30
PromptKD	96.30	98.01	97.15
Δ	+0.97	+0.71	+0.85

(d) OxfordPets

ViT-B/16	Base	Novel	HM
CLIP	63.37	74.89	68.65
CoOp	78.12	60.40	68.13
CoCoOp	70.49	73.59	72.01
MaPLe	72.94	74.00	73.47
PromptSRC	78.27	74.97	76.58
PromptKD	82.80	83.37	83.13
Δ	+4.53	+8.40	+6.55

(e) StanfordCars

ViT-B/16	Base	Novel	HM
CLIP	72.08	77.80	74.83
CoOp	97.60	59.67	74.06
CoCoOp	94.87	71.75	81.71
MaPLe	95.92	72.46	82.56
PromptSRC	98.07	76.50	85.95
PromptKD	99.42	82.62	90.24
Δ	+1.35	+6.12	+4.29

(f) Flowers102

ViT-B/16	Base	Novel	HM
CLIP	90.10	91.22	90.66
CoOp	88.33	82.26	85.19
CoCoOp	90.70	91.29	90.99
MaPLe	90.71	92.05	91.38
PromptSRC	90.67	91.53	91.10
PromptKD	92.43	93.68	93.05
Δ	+1.76	+2.15	+1.95

(g) Food101

ViT-B/16	Base	Novel	HM
CLIP	27.19	36.29	31.09
CoOp	40.44	22.30	28.75
CoCoOp	33.41	23.71	27.74
MaPLe	37.44	35.61	36.50
PromptSRC	42.73	37.87	40.15
PromptKD	49.12	41.81	45.17
Δ	+6.39	+3.94	+5.02

(h) FGVCAircraft

ViT-B/16	Base	Novel	HM
CLIP	69.36	75.35	72.23
CoOp	80.60	65.89	72.51
CoCoOp	79.74	76.86	78.27
MaPLe	80.82	78.70	79.75
PromptSRC	82.67	78.47	80.52
PromptKD	83.69	81.54	82.60
Δ	+1.02	+3.07	+2.08

(i) SUN397

ViT-B/16	Base	Novel	HM
CLIP	53.24	59.90	56.37
CoOp	79.44	41.18	54.24
CoCoOp	77.01	56.00	64.85
MaPLe	80.36	59.18	68.16
PromptSRC	83.37	62.97	71.75
PromptKD	85.84	71.37	77.94
Δ	+2.47	+8.40	+6.19

(j) DTD

ViT-B/16	Base	Novel	HM
CLIP	56.48	64.05	60.03
CoOp	92.19	54.74	68.69
CoCoOp	87.49	60.04	71.21
MaPLe	94.07	73.23	82.35
PromptSRC	92.90	73.90	82.32
PromptKD	97.54	82.08	89.14
Δ	+4.64	+8.18	+6.82

(k) EuroSAT

ViT-B/16	Base	Novel	HM
CLIP	70.53	77.50	73.85
CoOp	84.69	56.05	67.46
CoCoOp	82.33	73.45	77.64
MaPLe	83.00	78.66	80.77
PromptSRC	87.10	78.80	82.74
PromptKD	89.71	82.27	86.10
Δ	+2.61	+3.47	+3.36

(l) UCF101

$$\mathrm{HM} = \frac{2 \cdot \mathrm{Base} \cdot \mathrm{Novel}}{\mathrm{Base} + \mathrm{Novel}}$$

김범준

[PromptKD: Unsupervised Prompt Distillation for Vision-Language Models]

- # Experiment

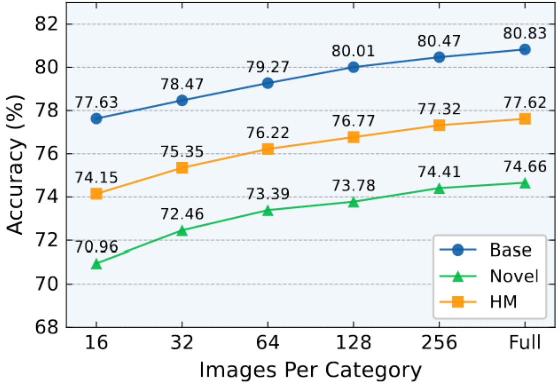


Figure 4. Improved ImageNet classification accuracy of the student model with increasing numbers of unlabeled images per category used for distillation.

- Unlabeled 수 많을수록 good, 일정해짐

Experiment

ZSL	ViT-B/16	Caltech 101	Oxford Pets	Standford Cars	Flowers 102	Food101	FGVC Aircraft	SUN397	DTD	Euro SAT	UCF101	Avg.
		Target Dataset										
In-ductive	CoOp	93.70	89.14	64.51	68.71	85.30	18.47	64.15	41.92	46.39	66.55	63.88
	CoCoOp	**94.43**	90.14	65.32	71.88	86.06	22.94	67.36	45.73	45.37	68.21	65.74
	MaPLe	93.53	90.49	65.57	72.23	86.20	24.74	67.01	46.49	48.06	68.69	66.30
	PromptSRC	93.60	90.25	65.70	70.25	86.15	23.90	67.10	46.87	45.50	68.75	65.81
Trans-ductive	PromptKD	93.61	**91.59**	**73.93**	**75.33**	**88.84**	**26.24**	**68.57**	**55.08**	**63.74**	**76.39**	**71.33**
	Δ	+0.01	+1.34	+8.23	+5.08	+2.69	+2.34	+1.47	+8.21	+18.24	+7.64	+5.52

Table 2. Comparison of PromptKD with existing advanced approaches on cross-dataset benchmark evaluation. Based on our pipeline, we perform unsupervised prompt distillation using the unlabeled domain data respectively (i.e., the transductive setting). The source model is trained on ImageNet [4]. "ZSL" denotes the setting type for Zero-Shot Learning. PromptKD achieves better results on 9 of 10 datasets.

- 도메인 간 Generalization 성능 평가, Source : ImageNet → Target : Each Dataset

- Inductive
  - 모델 학습때 **unlabeled test 데이터 x**
- Transductive
  - 모델 학습때 test 도메인 데이터 이용

- **Experiment**

Method	Domain Data	Base	Novel	HM
CLIP	Zero-shot	72.08	77.80	74.83
PromptSRC	Few-shot	98.07	76.50	85.95
CLIP-PR [13]		65.05	71.13	67.96
UPL [9]	Unlabeled	74.83	78.04	76.40
LaFTer [27]		79.49	82.91	81.16
FPL [26]		97.60	78.27	86.87
IFPL [26]	Few-shot	97.73	80.27	88.14
GRIP [26]	+	97.83	80.87	88.54
PromptKD	Unlabeled	99.42	82.62	90.24
Δ		+1.59	+1.75	+1.70

Table 3. Comparison with existing works using unlabeled data on Flowers102. Our method performs better than previous methods.

Method	Base	Novel	HM
CLIP	72.43	68.14	70.22
Projector Only	78.48	72.79	75.53
Full Fine-tune	75.90	70.95	73.34
w/o Shared Text Feature	78.79	73.37	75.98
PromptKD	79.27	73.39	76.22

Table 5. Ablation study of different distillation ways.

- # Experiment

Role (Method)	Img Backbone	Base	Novel	HM
CLIP	ViT-B/16	72.43	68.14	70.22
PromptSRC	ViT-B/16	77.60	70.73	74.01
Teacher (CLIP)	ViT-L/14	79.18	74.03	76.52
Student	ViT-B/16	76.53	72.58	74.50
Teacher (MaPLe)	ViT-L/14	82.79	76.88	79.73
Student	ViT-B/16	78.43	**73.61**	75.95
Teacher (PromptSRC)	ViT-L/14	83.24	76.83	79.91
Student	ViT-B/16	**79.27**	73.39	**76.22**

Table 6. Comparison of different pre-training methods. Teacher pre-training with PromptSRC brings the best student performance.

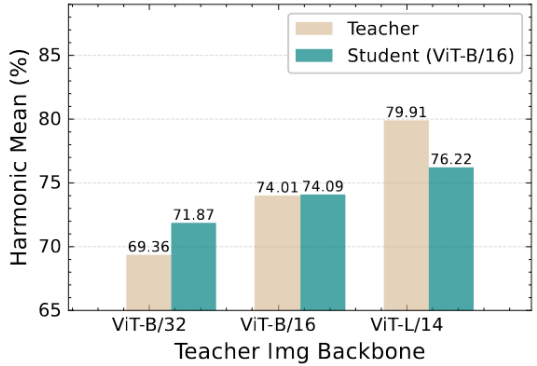


Figure 5. Comparison of distillation results for teachers with different capacities. Better teachers lead to better performance.

- 좋은 모델을 teacher에 사용 → 좋은 result
- 좋은 backbone 사용 → 좋은 result

- **Experiment**

Method	GFLOPs (test)	FPS	HM
CoOp	162.5	1344	71.66
CoCoOp	162.5	15.08	75.83
PromptSRC	162.8	1380	79.97
PromptKD	42.5	1710	83.73

Table 7. Comparison of computation costs among existing methods on the SUN397 dataset. Our PromptKD is more efficient than previous methods during testing.

- PromptKD는 test-time에서 매우 가볍고 빠르며 성능도 좋다