



SCHOOL OF
**COMPUTING &
DATA SCIENCE**
The University of Hong Kong



VaMP: Variational Multi-Modal Prompt Learning for Vision-Language Models

Silin Cheng, Kai Han

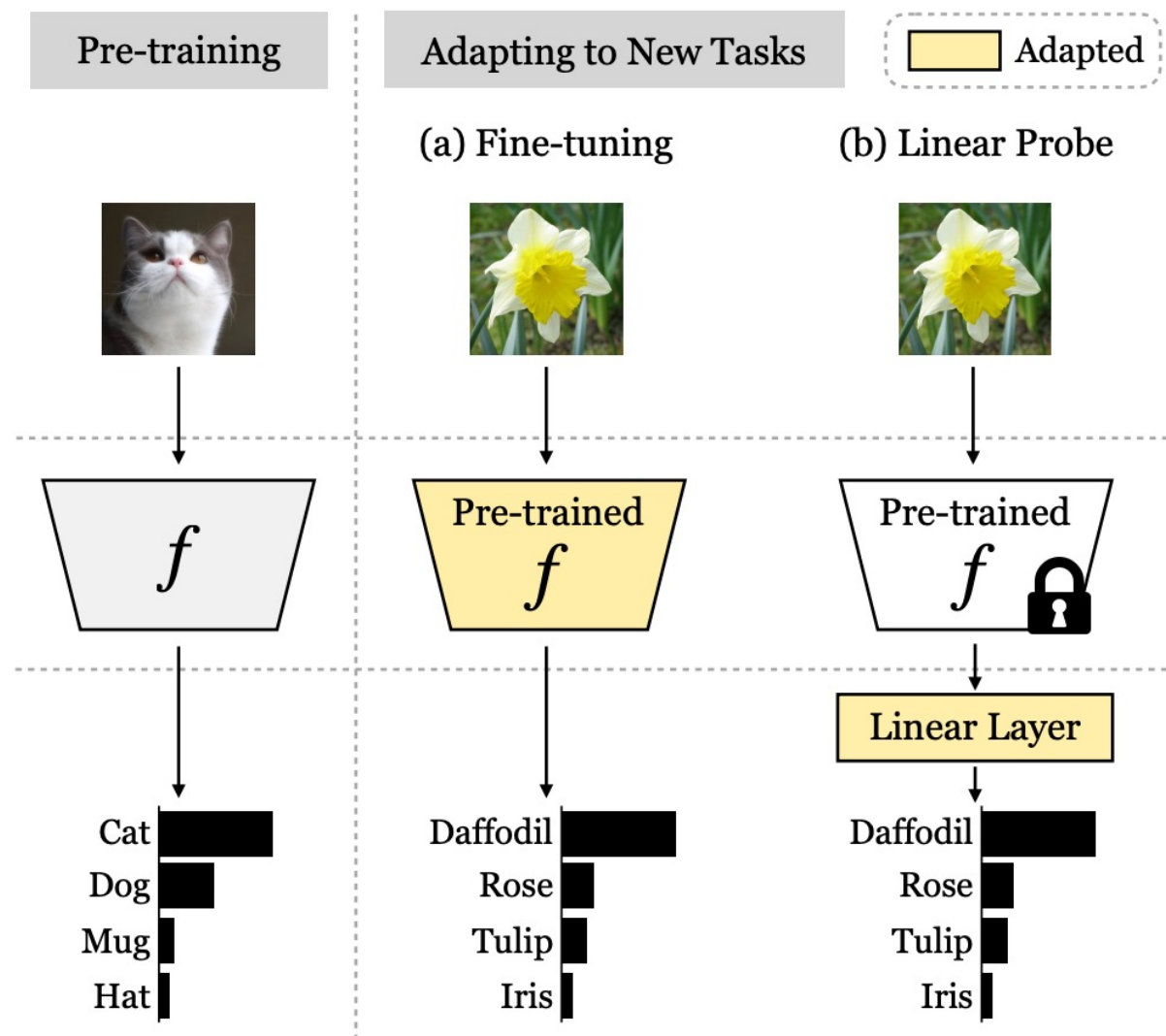
Visual AI Lab, The University of Hong Kong

NeurIPS 2025

Project Page: <https://visual-ai.github.io/vamp>


The Need for Prompt Learning

- Large VLMs (e.g., CLIP) are powerful but expensive to adapt
- **Full Fine-Tuning** is parameter-inefficient, prone to overfitting, and requires storing a full model copy for every task




The Need for Prompt Learning

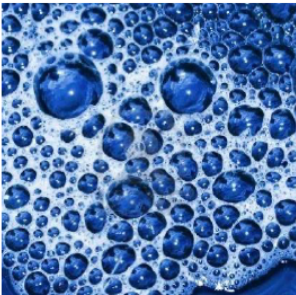
- **Manual Prompting** ("a photo of a...") is brittle
 - A small change in wording can cause huge performance shifts

Caltech101	Prompt	Accuracy
	a [CLASS].	82.68
	a photo of [CLASS].	80.81
	a photo of a [CLASS].	86.29

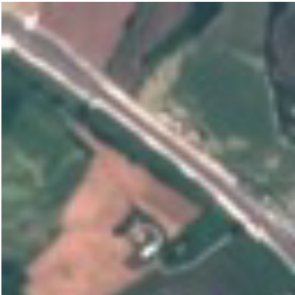
(a)

Flowers102	Prompt	Accuracy
	a photo of a [CLASS].	60.86
	a flower photo of a [CLASS].	65.81
	a photo of a [CLASS], a type of flower.	66.14

(b)

Describable Textures (DTD)	Prompt	Accuracy
	a photo of a [CLASS].	39.83
	a photo of a [CLASS] texture.	40.25
	[CLASS] texture.	42.32


(c)

EuroSAT	Prompt	Accuracy
	a photo of a [CLASS].	24.17
	a satellite photo of [CLASS].	37.46
	a centered satellite photo of [CLASS].	37.56


(d)

The Need for Prompt Learning

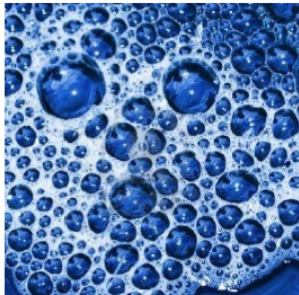
- **Prompt Learning** is the parameter-efficient solution
 - freeze the VLM and **learn** the best prompts from data

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

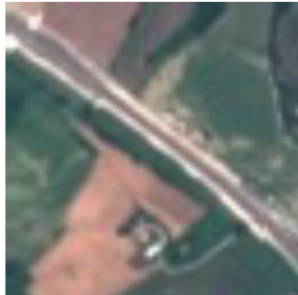
(a)

Flowers102	Prompt	Accuracy
	a photo of a [CLASS].	60.86
	a flower photo of a [CLASS].	65.81
	a photo of a [CLASS], a type of flower.	66.14
	$[V]_1 [V]_2 \dots [V]_M$ [CLASS].	94.51

(b)

Describable Textures (DTD)	Prompt	Accuracy
	a photo of a [CLASS].	39.83
	a photo of a [CLASS] texture.	40.25
	[CLASS] texture.	42.32
	$[V]_1 [V]_2 \dots [V]_M$ [CLASS].	63.58

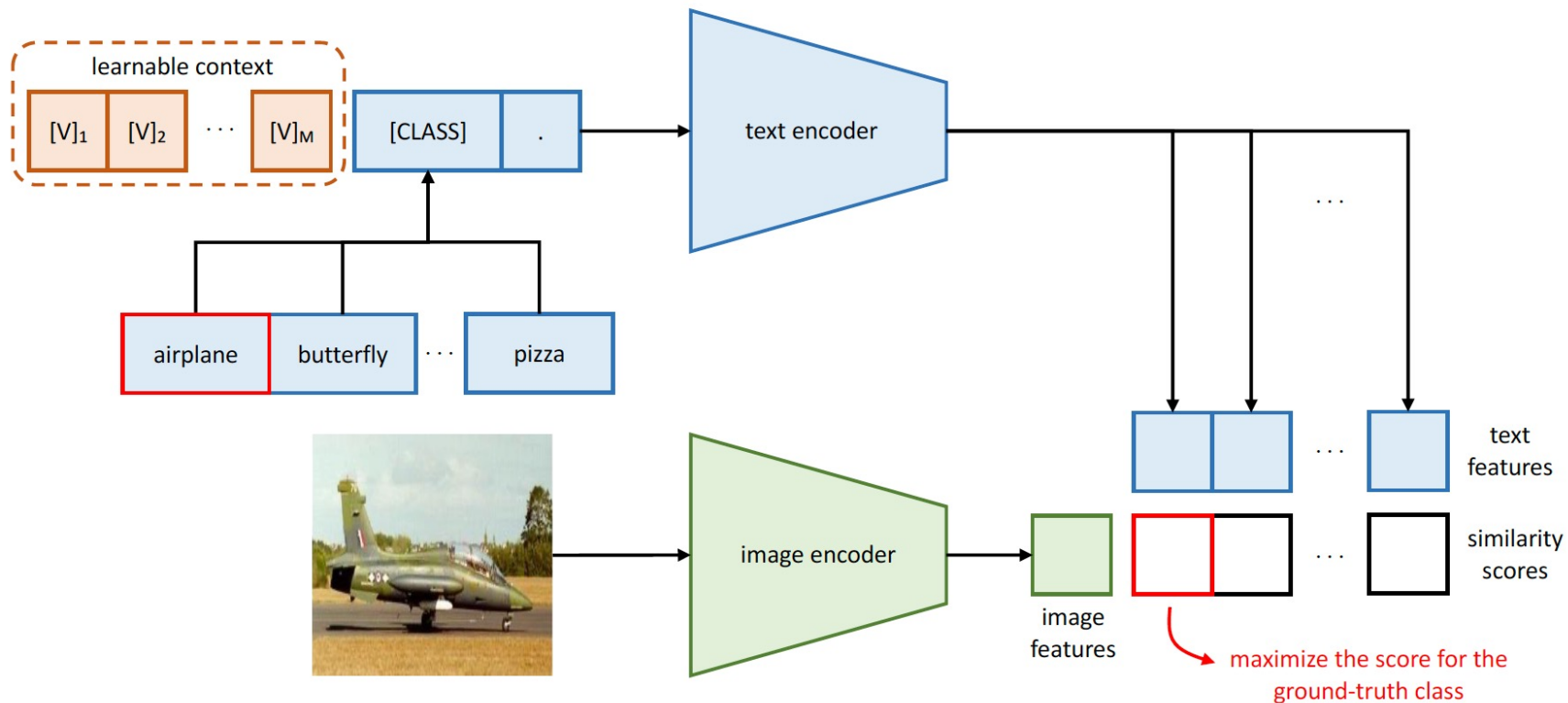
(c)

EuroSAT	Prompt	Accuracy
	a photo of a [CLASS].	24.17
	a satellite photo of [CLASS].	37.46
	a centered satellite photo of [CLASS].	37.56
	$[V]_1 [V]_2 \dots [V]_M$ [CLASS].	83.53

(d)

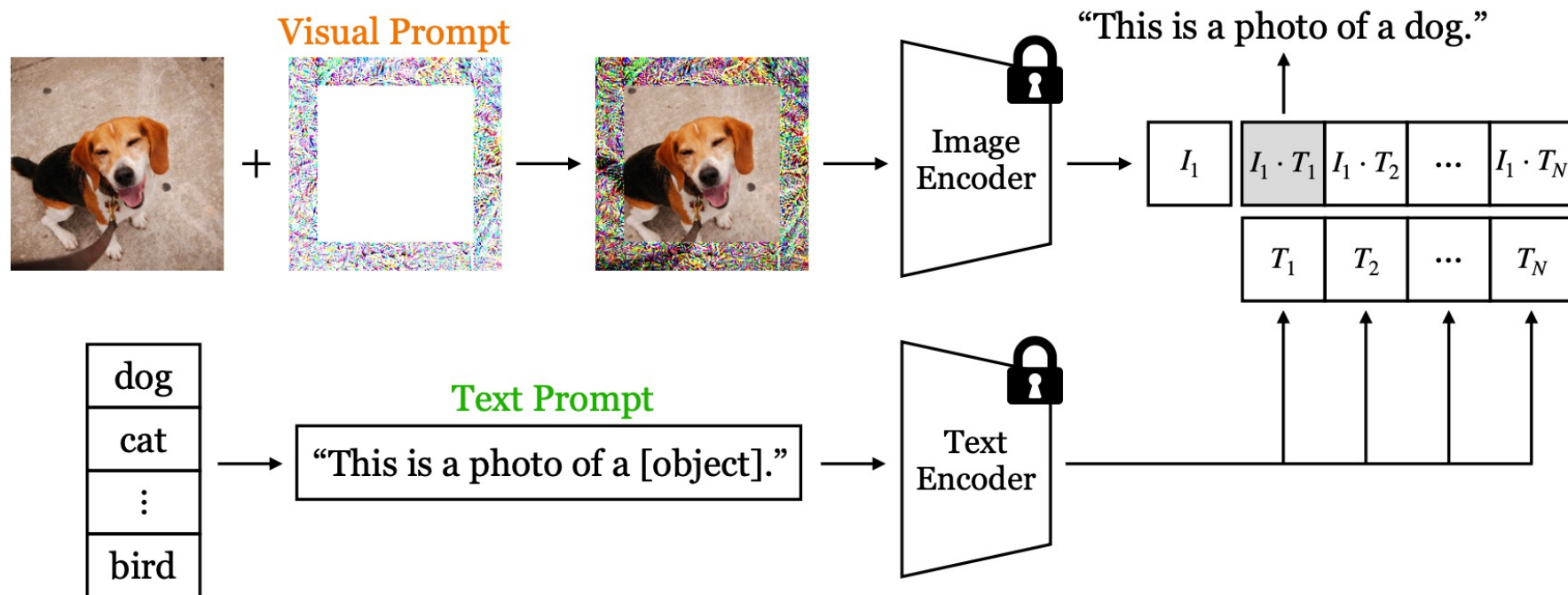
Approach 1: Textual Prompt Learning

- Learns continuous vectors (prompts) in the text embedding space
- These learnable prompts are fed **only** to the Text Encoder
- The Image Encoder is completely frozen and unchanged
- **Limitation:** This uni-modal approach ignores the vision branch entirely during adaptation



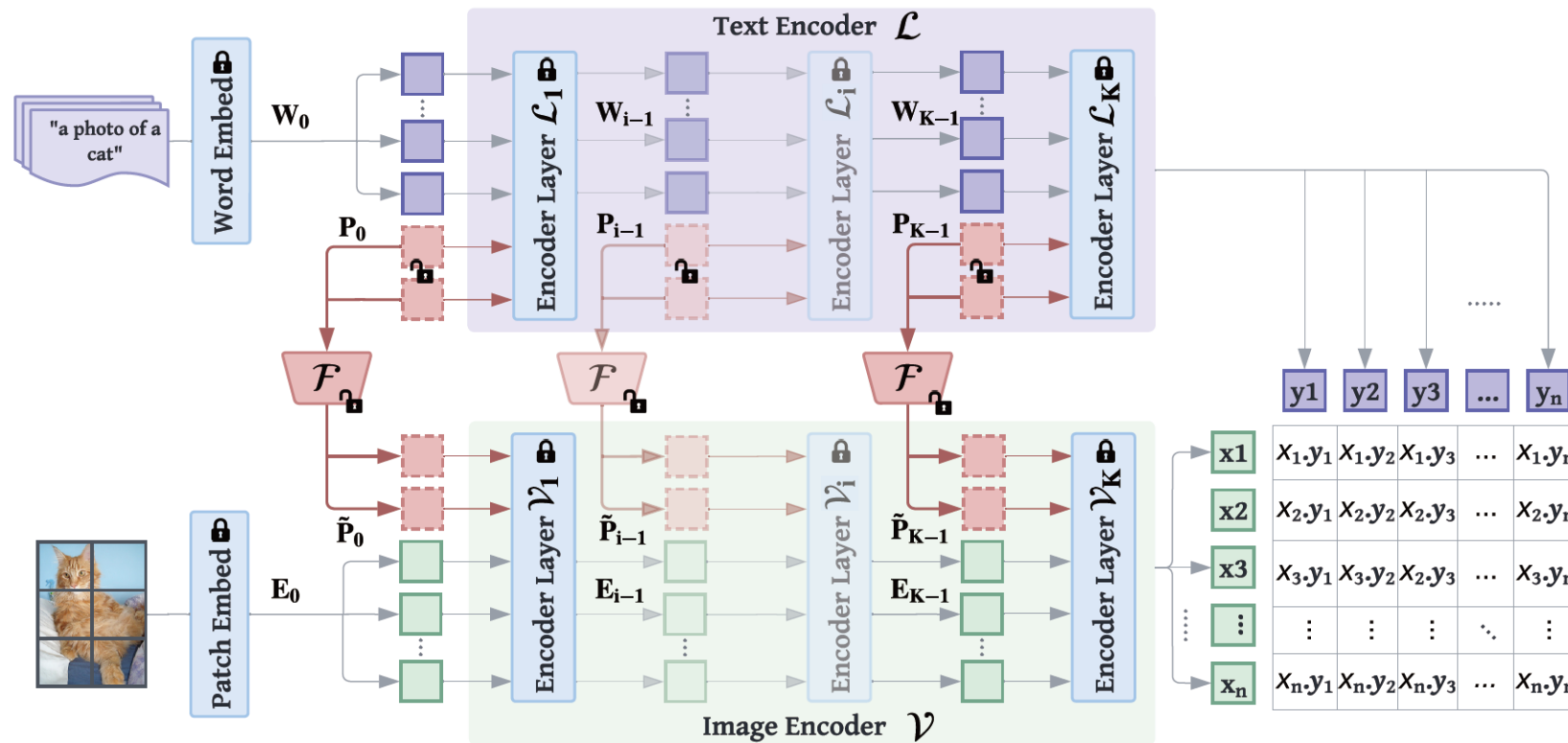
Approach 2: Visual Prompt Learning

- Learns continuous vectors (prompts) in the **image** patch space.
- These visual prompts are prepended to the image patch tokens and fed **only** to the Image Encoder.
- The Text Encoder is completely frozen and unchanged
- **Limitation:** This uni-modal approach ignores the language branch entirely during adaptation



Approach 3: Multi-Modal Prompt Learning

- Learns prompts for **both** the Text and Image Encoders
- A "Coupling Function" maps the learned language prompts to the visual prompts
- This allows for joint, synergistic tuning to improve vision-language alignment
- **Limitation:** The prompts are still **static** and **deterministic**

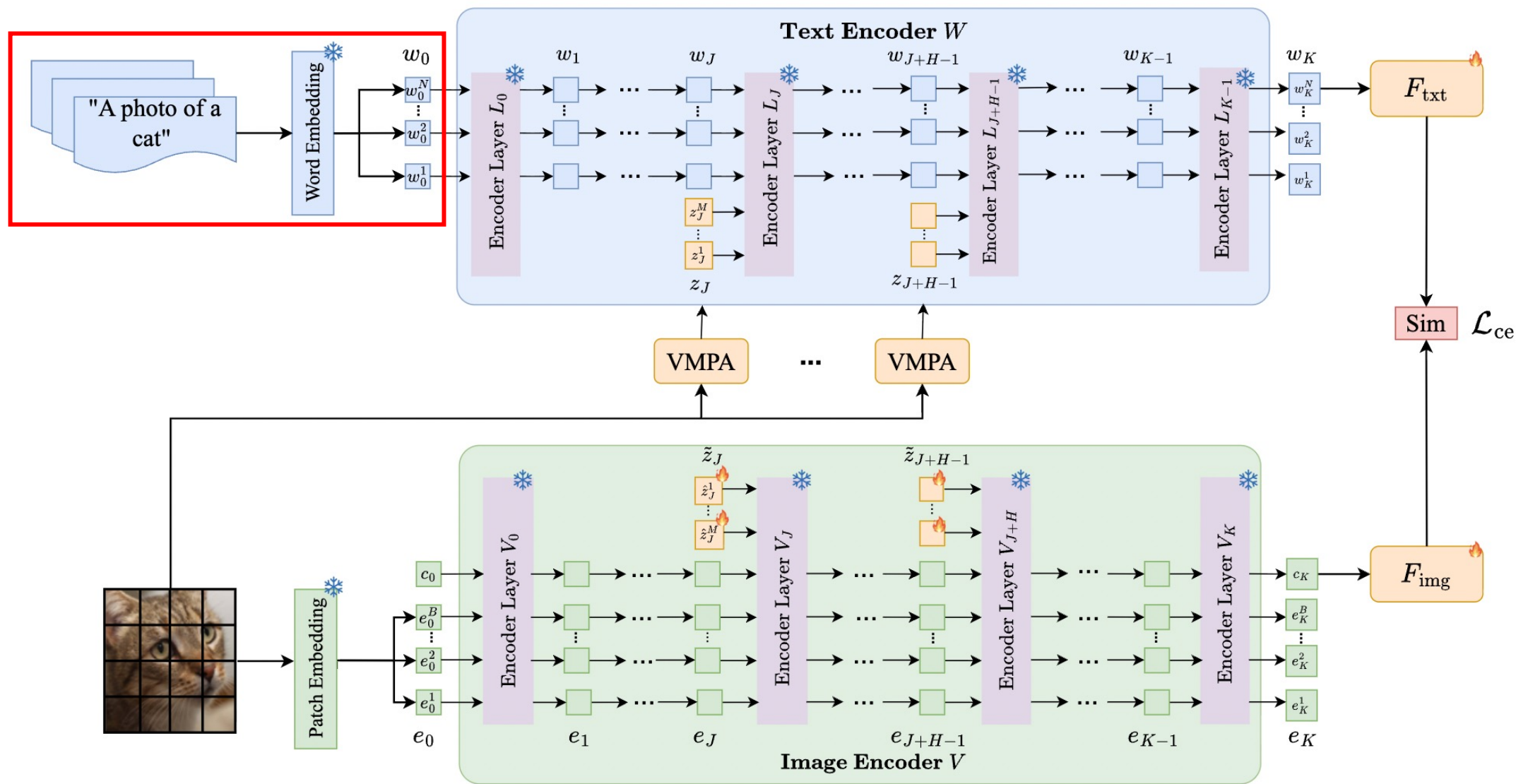


Limitations of Existing Methods

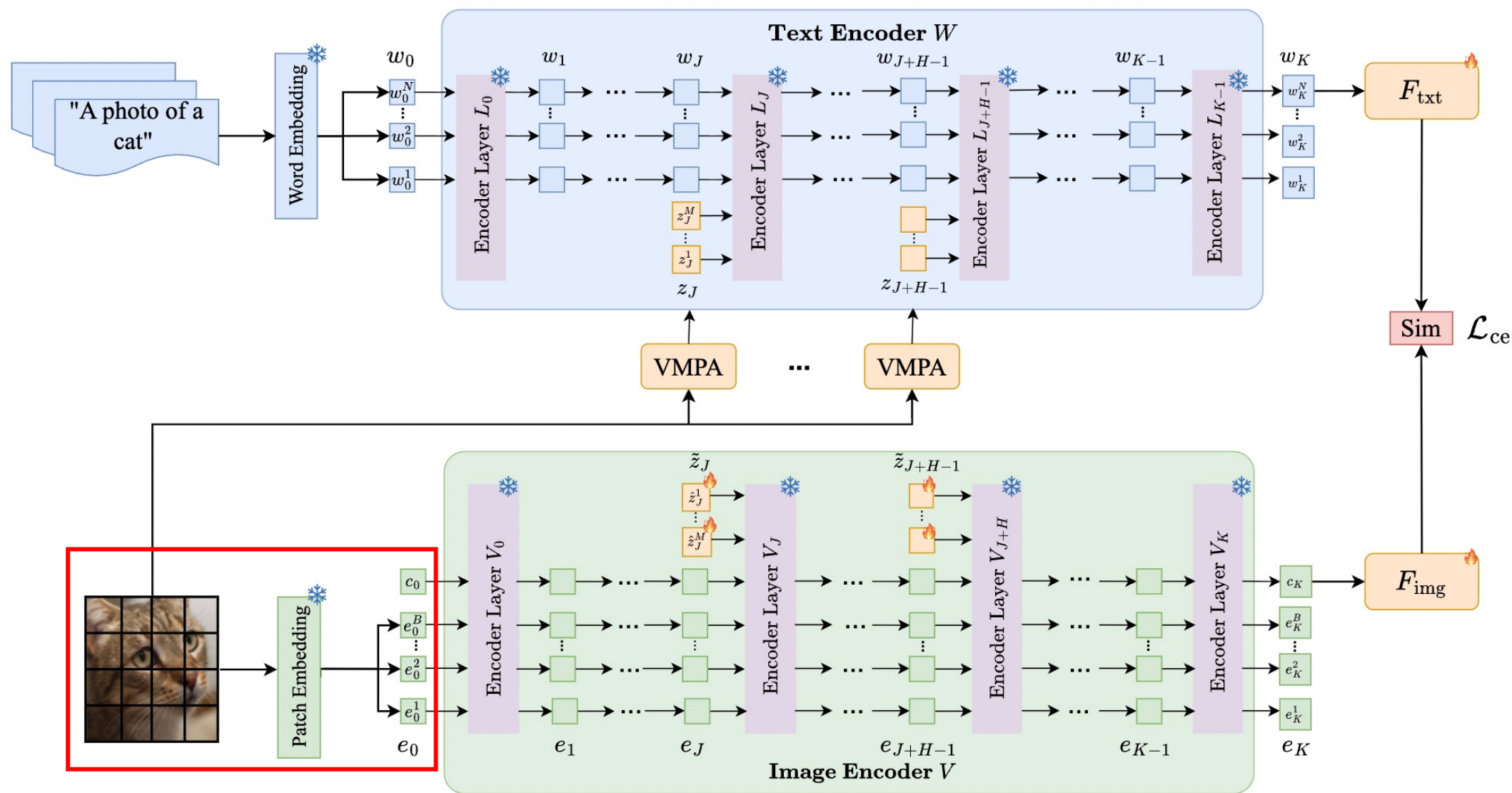
- **Static & Deterministic:** Methods like CoOp, VPT, and MaPLe learn a single, **fixed** set of prompts that are applied uniformly to all samples
- **No Instance-Level Adaptation:** They lack the flexibility to adapt to instance-level variations (e.g., a "photo of a dog" vs. a "sketch of a dog")
- **No Uncertainty Modeling:** They are deterministic and fail to capture model uncertainty, which limits robustness to new domains

We need prompts that are: Sample-Specific and Uncertainty-Aware !

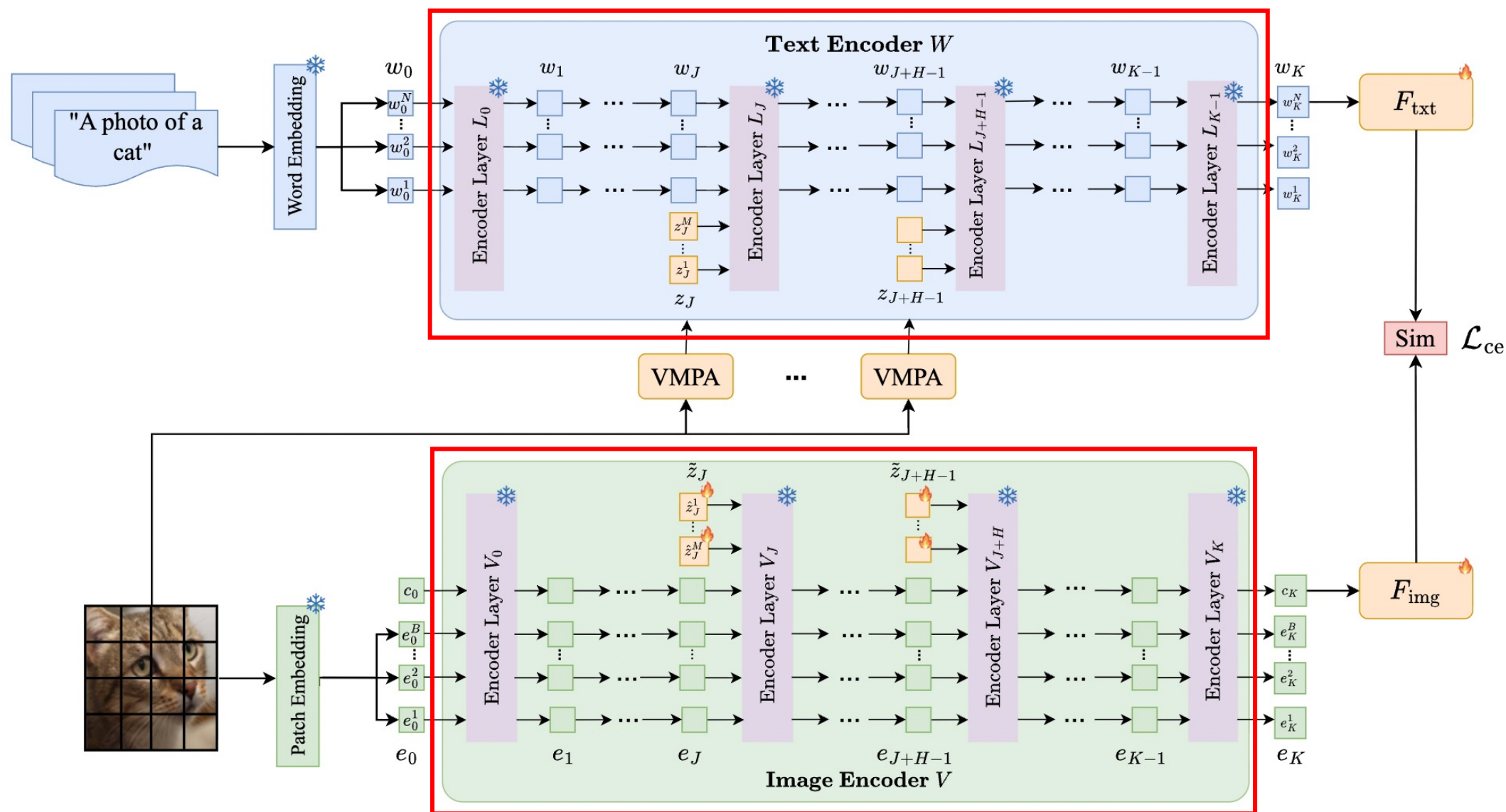
The VaMP Framework



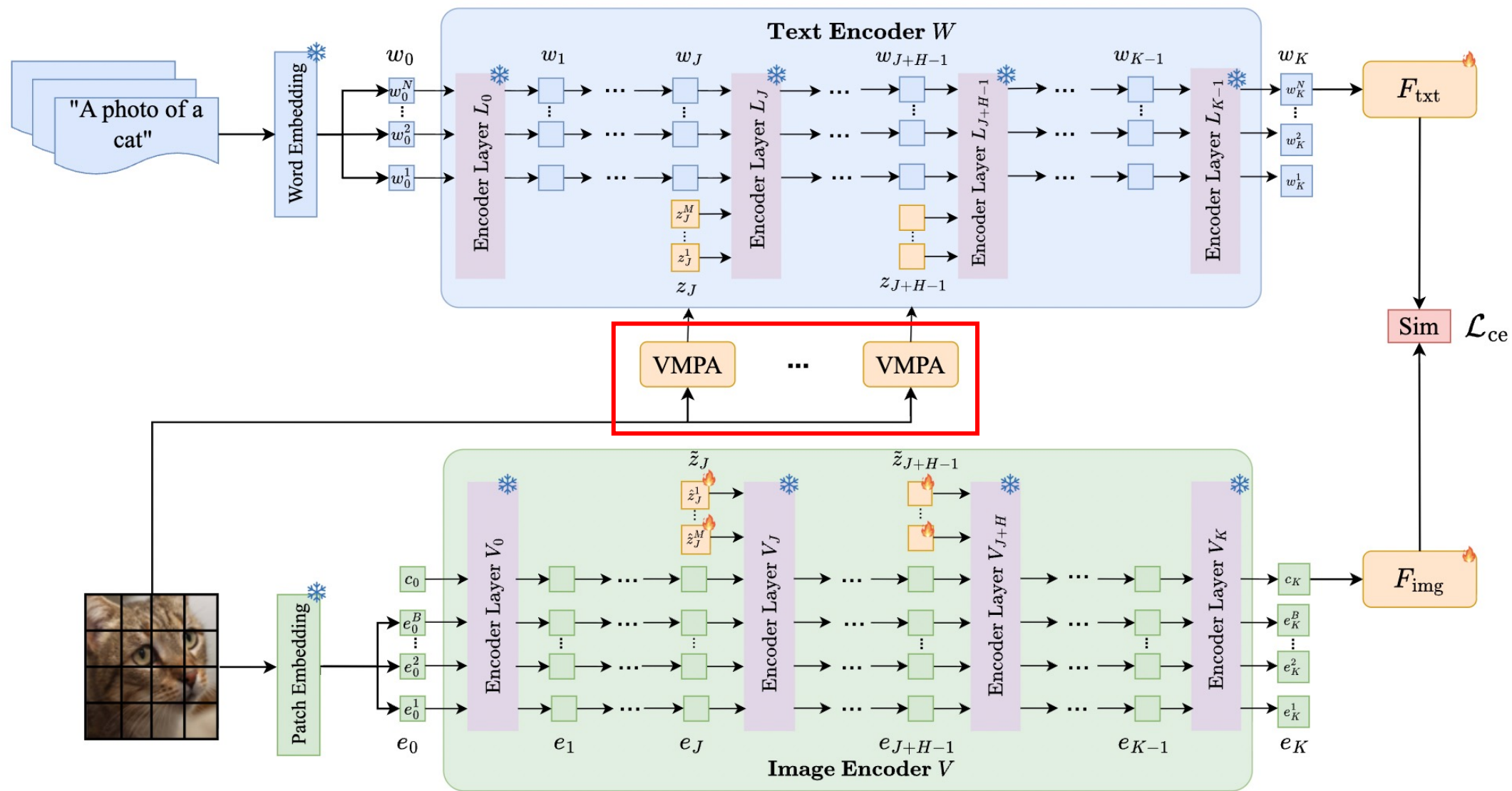
The VaMP Framework



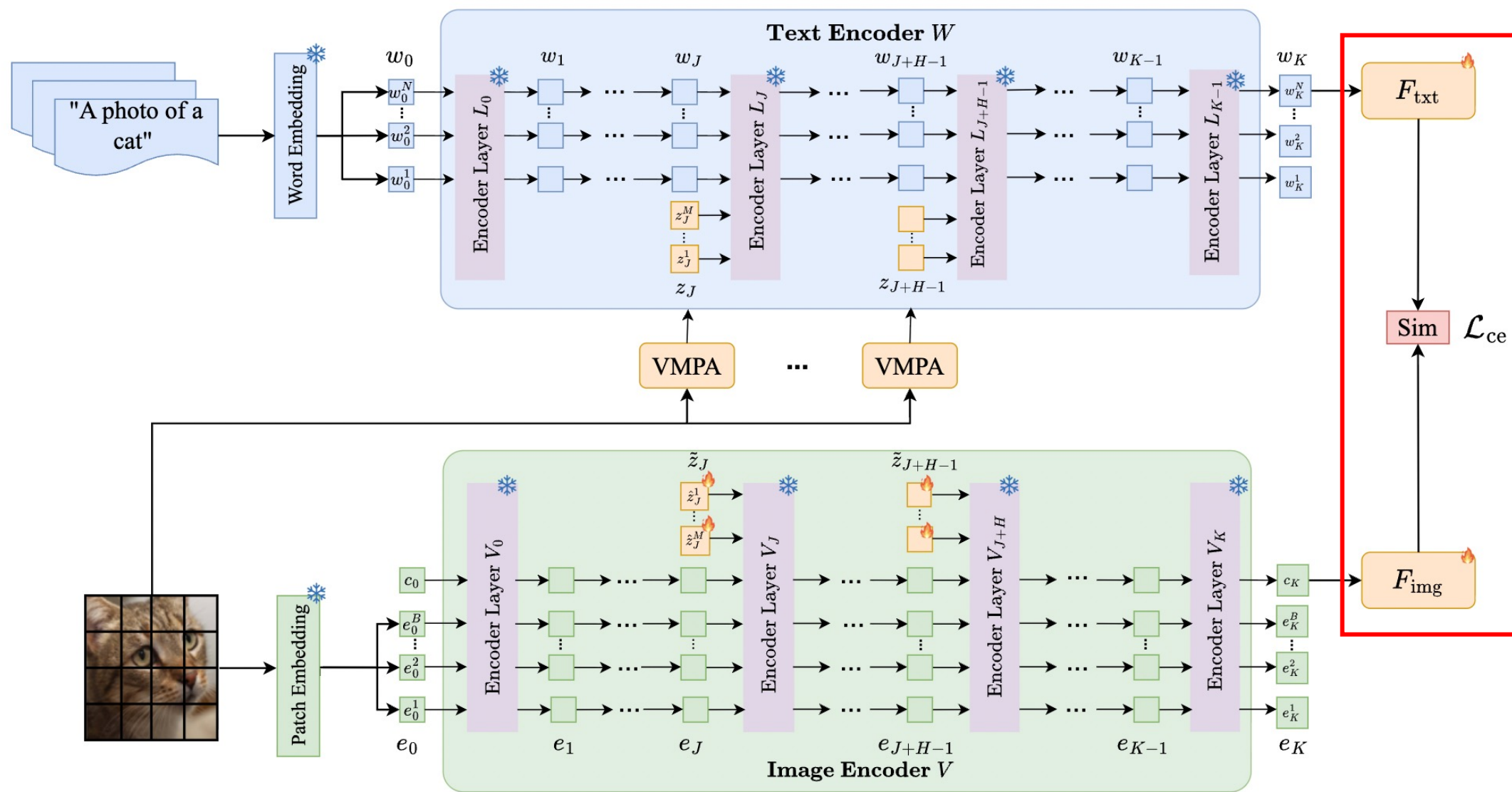
The VaMP Framework



The VaMP Framework

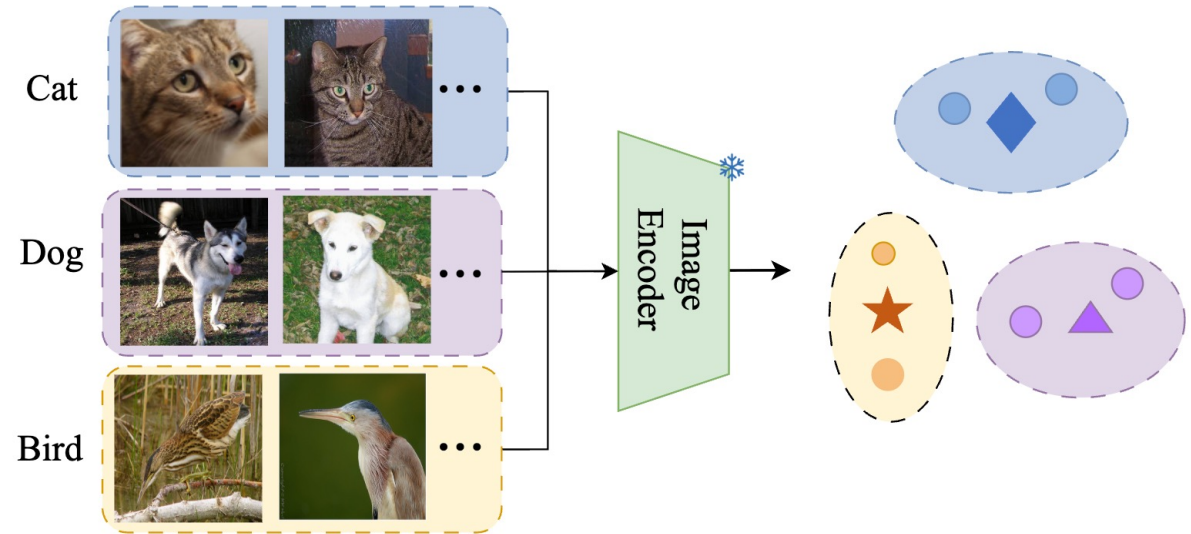


The VaMP Framework



Class-Aware Prior

- **Problem:** A standard prior $N(0, I)$ is uninformative and lacks semantic structure
- **Solution:** Introduce a structured prior conditioned on class semantics
 - **Step 1:** Pre-compute a class prototype for each class by averaging its training sample features
 - **Step 2:** Learn a prior network that maps this prototype to a Gaussian prior distribution

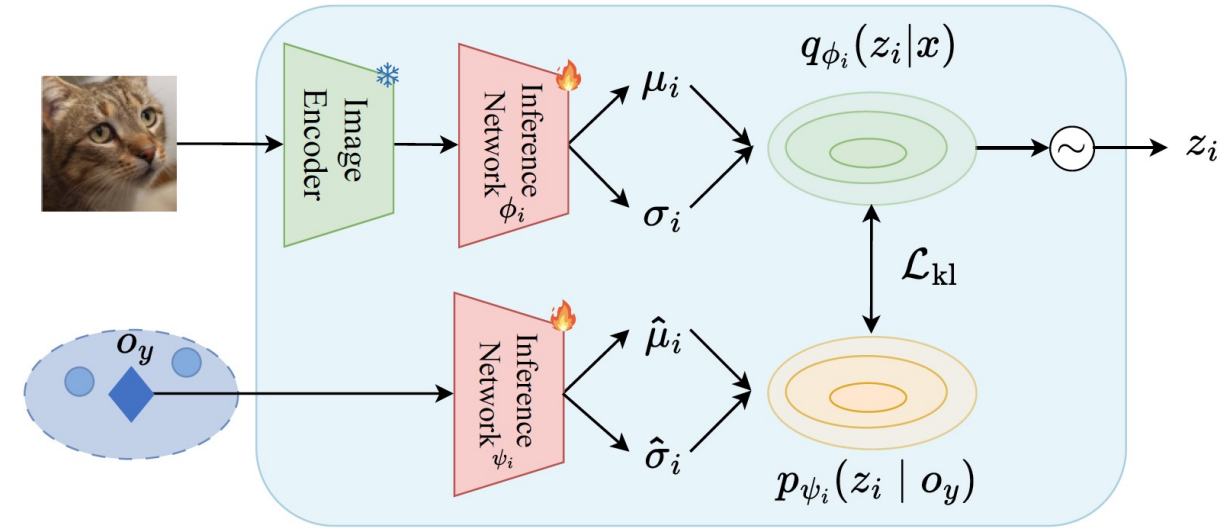


- Regularizes the latent space, pulling prompts from the same class closer together

Variational Multi-Modal Prompt Adaptation

The core engine of VaMP, which works in two streams:

- **Posterior (Sample-Specific):** An "Inference Network" ϕ_i uses the input image feature x to predict a posterior distribution $q_{\phi_i}(z_i|x)$
- **Prior (Class-Aware):** A "Prior Network" ψ_i uses the class prototype o_y to generate our structured prior distribution $p_{\psi_i}(z_i|x)$



Training: A KL divergence loss L_{kl} forces the image-specific posterior $q_{\phi_i}(z_i|x)$ to be close to $p_{\psi_i}(z_i|x)$

Results: Base-to-Novel Generalization

Method	Average			ImageNet			Caltech101			OxfordPets		
	Base	Novel	H	Base	Novel	H	Base	Novel	H	Base	Novel	H
CLIP [1]	69.34	74.22	71.70	72.43	68.14	70.22	96.84	94.00	95.40	91.17	97.26	94.12
CoOp [2]	82.69	63.22	71.66	76.47	67.88	71.92	98.00	89.81	93.73	93.67	95.29	94.47
CoOpOp [3]	80.47	71.69	75.83	75.98	70.43	73.10	97.96	93.81	95.84	95.20	97.69	96.43
ProDA [38]	81.56	72.30	76.65	75.40	70.23	72.72	98.27	93.23	95.68	95.43	97.83	96.62
KgCoOp [41]	80.73	73.60	77.00	75.83	69.96	72.78	97.72	94.39	96.03	94.65	97.76	96.18
MaPLe [5]	82.28	75.14	78.55	76.66	70.54	73.47	97.74	94.36	96.02	95.43	97.76	96.58
PromptSRC [6]	84.26	76.10	79.97	77.60	70.73	74.01	98.10	94.03	96.02	95.33	97.30	96.30
TCP [42]	84.13	75.36	79.51	77.27	69.87	73.38	98.23	94.67	96.42	94.67	97.20	95.92
MMA [70]	83.20	76.80	79.87	77.31	71.00	74.02	98.40	94.00	96.15	95.40	98.07	96.72
2SFS [87]	85.55	75.48	80.20	77.71	70.99	74.20	98.71	94.43	96.52	95.32	97.82	96.55
SkipT [88]	85.04	77.53	81.11	77.73	70.40	73.89	98.50	95.33	96.89	95.70	97.87	96.77
MMRL [8]	85.68	77.16	81.20	77.90	71.30	74.45	98.97	94.50	96.68	95.90	97.60	96.74
VaMP	86.45	78.67	82.37	78.98	73.45	76.11	98.95	95.96	97.43	96.95	95.24	96.08

Method	StanfordCars			Flowers102			Food101			FGVCAircraft		
	Base	Novel	H	Base	Novel	H	Base	Novel	H	Base	Novel	H
CLIP [1]	63.37	74.89	68.65	72.08	77.80	74.83	90.10	91.22	90.66	27.19	36.29	31.09
CoOp [2]	78.12	60.40	68.13	97.60	59.67	74.06	88.33	82.26	85.19	40.44	22.30	28.75
CoOpOp [3]	70.49	73.59	72.01	94.87	71.75	81.71	90.70	91.29	90.99	33.41	23.71	27.74
ProDA [38]	74.70	71.20	72.91	97.70	68.68	80.66	90.30	88.57	89.43	36.90	34.13	35.46
KgCoOp [41]	71.76	75.04	73.36	95.00	74.73	83.65	90.50	91.70	91.09	36.21	33.55	34.83
MaPLe [5]	72.94	74.00	73.47	95.92	72.46	82.56	90.71	92.05	91.38	37.44	35.61	36.50
PromptSRC [6]	78.27	74.97	76.58	98.07	76.50	85.95	90.67	91.53	91.10	42.73	37.87	40.15
TCP [42]	80.80	74.13	77.32	97.73	75.57	85.23	90.57	91.37	90.97	41.97	34.43	37.83
MMA [70]	78.50	73.10	75.70	97.77	75.93	85.48	90.13	91.30	90.71	40.57	36.33	38.33
2SFS [87]	82.50	74.80	78.46	98.29	76.17	85.83	89.11	91.34	90.21	47.48	35.51	40.63
SkipT [88]	82.93	72.50	77.37	98.57	75.80	85.70	90.67	92.03	91.34	45.37	37.13	40.84
MMRL [8]	81.30	75.07	78.06	98.97	77.27	86.78	90.57	91.50	91.03	46.30	37.03	41.15
VaMP	83.78	80.14	81.91	98.96	83.97	90.85	92.77	93.16	92.96	46.77	41.13	43.76

Method	SUN397			DTD			EuroSAT			UCF101		
	Base	Novel	H	Base	Novel	H	Base	Novel	H	Base	Novel	H
CLIP [1]	69.36	75.35	72.23	53.24	59.90	56.37	56.48	64.05	60.03	70.53	77.50	73.85
CoOp [2]	80.60	65.89	72.51	79.44	41.18	54.24	92.19	54.74	68.69	84.69	56.05	67.46
CoOpOp [3]	79.74	76.86	78.27	77.01	56.00	64.85	87.49	60.04	71.21	82.33	73.45	77.64
ProDA [38]	78.67	76.93	77.79	80.67	56.48	66.44	83.90	66.00	73.88	85.23	71.97	78.04
KgCoOp [41]	80.29	76.53	78.36	77.55	54.99	64.35	85.64	64.34	73.48	82.89	76.67	79.65
MaPLe [5]	80.82	78.70	79.75	80.36	59.18	68.16	94.07	73.23	82.35	83.00	78.66	80.77
PromptSRC [6]	82.67	78.47	80.52	83.37	62.97	71.75	92.90	73.90	82.32	87.10	78.80	82.74
TCP [42]	82.63	78.20	80.35	82.77	58.07	68.25	91.63	74.73	82.32	87.13	80.77	83.83
MMA [70]	82.27	78.57	80.38	83.20	65.63	73.38	85.46	82.34	83.87	86.23	80.03	82.20
2SFS [87]	82.59	78.91	80.70	84.60	65.01	73.52	96.91	67.09	79.29	87.85	78.19	82.74
SkipT [88]	82.40	79.03	80.68	83.77	67.23	74.59	92.47	83.00	87.48	87.30	82.47	84.81
MMRL [8]	83.20	79.30	81.20	85.67	65.00	73.82	95.60	80.17	87.21	88.10	80.07	83.89
VaMP	83.37	78.95	81.09	86.14	67.20	75.50	95.78	77.21	85.49	88.52	78.99	83.48

Results: Domain Generalization

	Source	Target			
	ImageNet	-V2	-S	-A	-R
CLIP [1]	66.73	60.83	46.15	47.77	73.96
CoOp [2]	71.51	64.20	47.99	49.71	75.21
CoOpOp [3]	71.02	64.07	48.75	50.63	76.18
MaPLe [5]	70.72	64.07	49.15	50.90	76.98
PromptSRC [6]	71.27	64.35	49.55	50.90	77.80
MMA [70]	71.00	64.33	49.13	51.12	77.32
MMRL [8]	72.03	64.47	49.17	51.20	77.53
VaMP	72.83	64.96	49.69	51.97	78.01

Results: Cross-Dataset Generalization

	Source	Target										
	<i>ImageNet</i>	<i>Average</i>	<i>Caltech101</i>	<i>OxfordPets</i>	<i>StanfordCars</i>	<i>Flowers101</i>	<i>Food101</i>	<i>FGVCAircraft</i>	<i>SUN397</i>	<i>DTD</i>	<i>EuroSAT</i>	<i>UCF101</i>
CoOp [2]	71.51	63.88	93.70	89.14	64.51	68.71	85.30	18.47	64.15	41.92	46.39	66.55
CoOpOp [3]	71.02	65.74	94.43	90.14	65.32	71.88	86.06	22.94	67.36	45.73	45.37	68.21
MaPLe [5]	70.72	66.30	93.53	90.49	65.57	72.23	86.20	24.74	67.01	46.49	48.06	68.69
PromptSRC [6]	71.27	65.81	93.60	90.25	65.70	70.25	86.15	23.90	67.10	46.87	45.50	68.75
TCP [42]	71.40	66.29	93.97	91.25	64.69	71.21	86.69	23.45	67.15	44.35	51.45	68.73
MMA [70]	71.00	66.61	93.80	90.30	66.13	72.07	86.12	25.33	68.17	46.57	49.24	68.32
MMRL [8]	72.03	67.25	94.67	91.43	66.10	72.77	86.40	26.30	67.57	45.90	53.10	68.27
VaMP	72.83	67.74	94.96	91.79	66.10	73.18	86.97	26.76	68.04	46.82	53.82	68.93

Thank You!