

Test-Time Adaptation with CLIP Reward for Zero-Shot Generalization in Vision-Language Models

Shuai Zhao, Xiaohan Wang, Linchao Zhu, Yi Yang
University of Technology Sydney, Zhejiang University



Task and Problem

Task: TTA with VLMs in 0-shot cases

Problem: popular entropy minimization methods are stick to incorrect predictions



Feedback in TTA

Solution: feedback mechanism

Feedback source: CLIP! 1) No need for label. 2) CLIP is reliable.

$$\text{CLIP}(t, v) = \cos(h(t), g(v))$$

Reinforcement Learning with CLIP Feedback (RLCF)

Goal: given a VLM f_θ and a single input image v or text t , learn $P = f_\theta(v)$ or $f_\theta(t)$

$$\max_{\theta} \mathbb{E}_{t \sim P(\cdot | v, \theta)} \mathcal{R}(t, v).$$

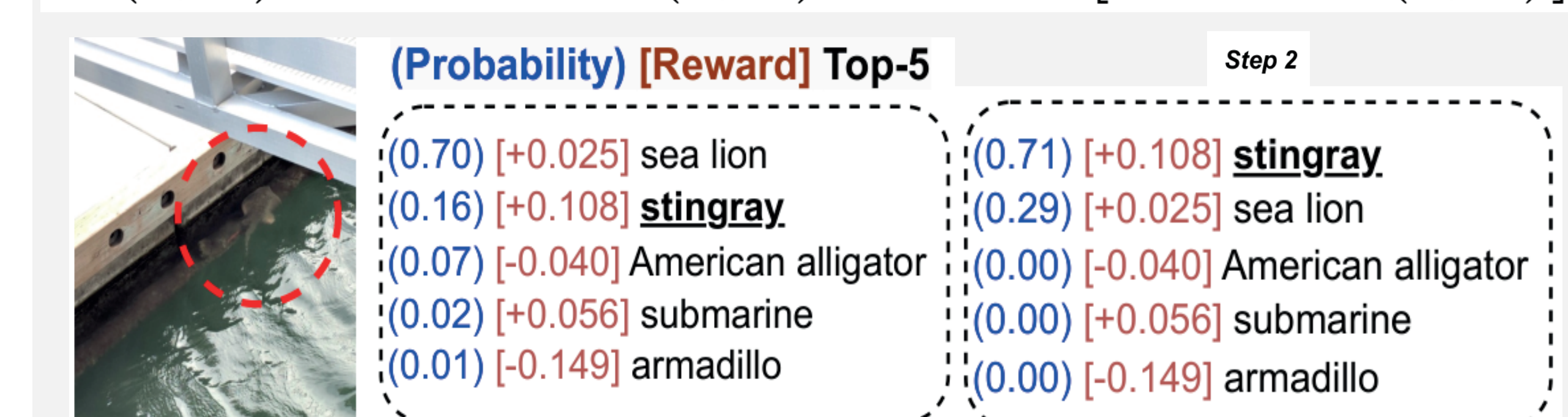
Policy gradient with REINFORCE:

$$\nabla_{\theta} \mathbb{E}_{t \sim P} [\mathcal{R}(t, v)] = \mathbb{E}_{t \sim P} [\mathcal{R}(t, v) \nabla_{\theta} \log P(t | v; \theta)].$$

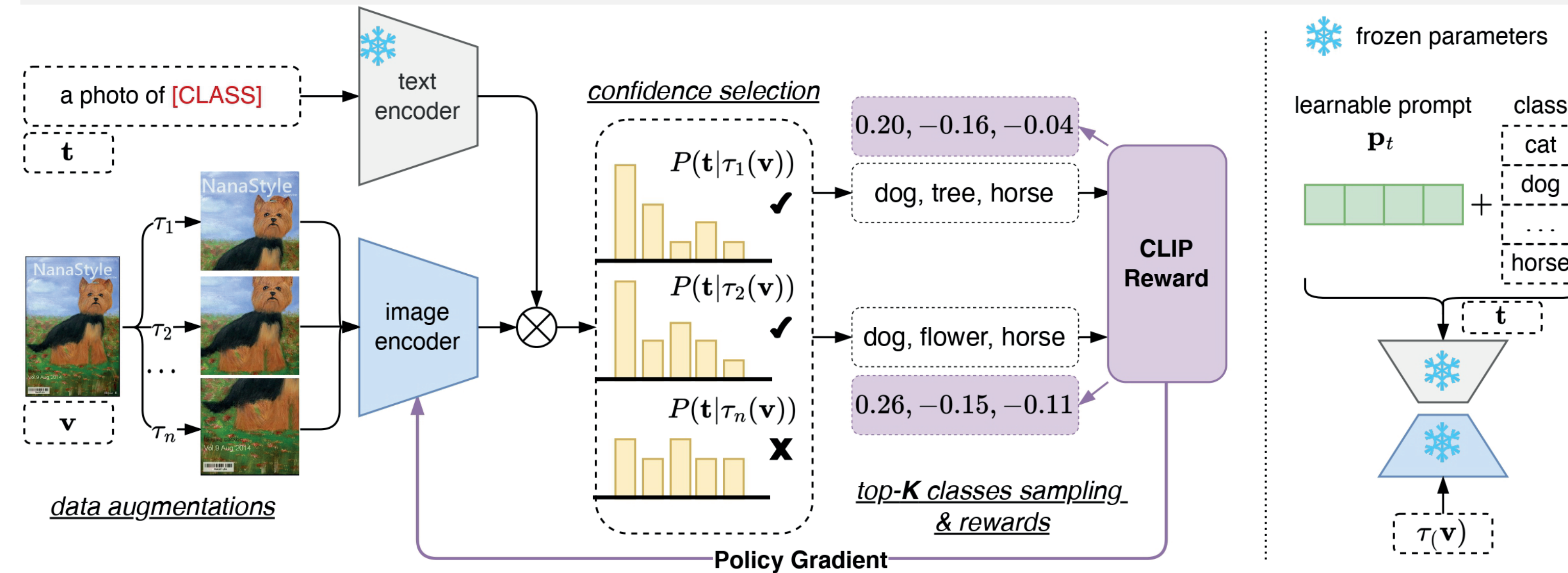
CLIP Reward with baseline:

$$\text{CLIP-S}(t, v) = w \times \max(\text{CLIP}(t, v), 0),$$

$$\mathcal{R}(t, v) = \text{CLIP-S}(t, v) - \mathbb{E}_{t \sim P} [\text{CLIP-S}(t, v)].$$

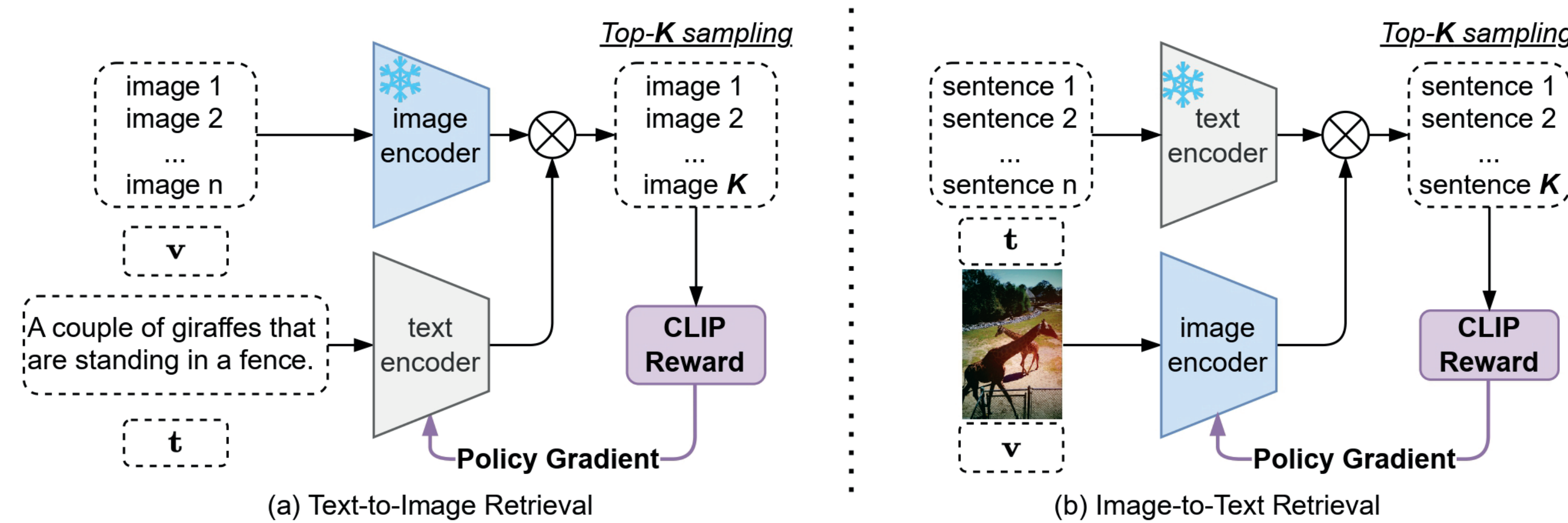


RLCF for zero-shot classification



Pipeline: data augmentations $\rightarrow$ confidence selection $\rightarrow$ top-K sampling $\rightarrow$ CLIP Reward $\rightarrow$ gradient updating.

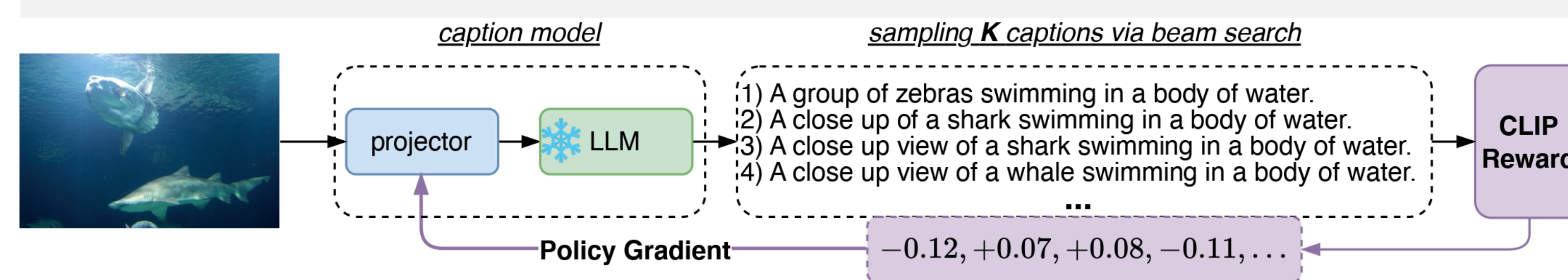
RLCF for zero-shot retrieval



Pipeline: query $\rightarrow$ top-K sampling $\rightarrow$ CLIP Reward $\rightarrow$ gradient updating.

We only update the parameters *w.r.t.* the query.

RLCF for image captioning



Pipeline: image $\rightarrow$ projection $\rightarrow$ LLM generation $\rightarrow$ beam search $\rightarrow$ CLIP Reward $\rightarrow$ gradient updating.

We only update the parameters of the projector.

Captioning model: CapDec and ClipCap.

Experiments

RLCF: CLIP-L/14 as the reward model.

RLCF-S: CLIP-L/14-336, CLIP-L/14, CLIP-RN50x64

RLCF-S-M: + momentum buffer for continual learning

Table 1: Top-1 accuracy of zero-shot image classification. The best and second-best results.

Method	IN	IN-A	IN-V2	IN-R	IN-Sketch	OOD Avg.
<i>Zero-shot baseline</i>						
CLIP-ViT-B/16	66.73	47.87	60.86	73.98	46.09	57.20
CLIP-ViT-L/14	73.44	68.82	67.80	85.40	57.84	69.97
Ensemble (B/16 + L/14)	75.09	65.94	69.02	85.92	57.98	69.72
<i>Prompt tuning for CLIP-ViT-B/16</i>						
TPT + CoOp	73.61	57.95	66.83	77.27	49.29	62.84
RLCF	73.23	65.45	69.77	83.35	54.74	68.33
RLCF + CoOp	76.05	69.74	70.62	84.51	56.49	70.34
RLCF-S + CoOp	76.50	71.11	70.92	84.73	56.97	70.93
<i>Image encoder tuning for CLIP-ViT-B/16</i>						
ATKD	70.51	70.66	65.54	85.12	53.56	68.72
RLCF	74.85	73.71	69.77	86.19	57.10	71.69
RLCF-S	75.34	75.00	70.08	86.97	57.75	72.45
RLCF-S-M	75.48	75.16	70.42	87.23	57.73	72.64

Table 2: TTA for zero-shot text-image retrieval. Improvement in ($\uparrow$ blue).

Method	MS-COCO (5K test images)				Flickr30K (1K test images)			
	text-to-image	image-to-text	text-to-image	image-to-text	text-to-image	image-to-text	text-to-image	image-to-text
	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
<i>Zero-shot baseline</i>								
CLIP-ViT-B/16	33.0	58.2	52.5	76.8	62.2	85.7	81.2	96.4
CLIP-ViT-L/14	36.1	60.9	56.2	78.9	64.6	87.1	85.3	97.2
CLIP-ViT-L/14-336	36.6	60.9	57.3	80.6	67.1	88.9	86.6	98.0
<i>TTA for CLIP-ViT-B/16</i>								
RLCF	37.3 ($\uparrow$ 4.3)	62.7	59.1 ($\uparrow$ 6.6)	80.1	67.1 ($\uparrow$ 4.9)	89.1	87.3 ($\uparrow$ 6.1)	97.2
RLCF-S	38.3 ($\uparrow$ 5.3)	63.4	60.8 ($\uparrow$ 8.3)	80.8	68.5 ($\uparrow$ 6.3)	90.0	88.3 ($\uparrow$ 7.1)	97.7
RLCF-S-M	38.4 ($\uparrow$ 5.4)	63.5	60.8 ($\uparrow$ 8.3)	80.5	68.5 ($\uparrow$ 6.3)	90.2	88.1 ($\uparrow$ 6.9)	97.7

Table 3: TTA for image captioning. BLEU@4, CIDEr (gain in ($\uparrow$ blue)), and SPICE.

Method	MS-COCO $\Rightarrow$ NoCaps								
	in domain			near domain			out domain		
	B@4	C	S	B@4	C	S	B@4	C	S
<i>TTA for CapDec (zero-shot)</i>									
CapDec	32.4	62.6	10.3	29.2	54.0	9.6	17.2	31.7	6.4
CapDec + RLCF	33.3	68.0 ($\uparrow$ 5.3)	10.7	30.3	57.9 ($\uparrow$ 3.9)	10.3	17.6	35.5 ($\uparrow$ 3.8)	6.9
<i>TTA for CLIPCap (cross-domain)</i>									
CLIPCap	36.3	76.9	11.9	34.8	73.5	11.0	22.5	54.6	8.6
CLIPCap + RLCF	38.6	84.0 ($\uparrow$ 7.1)	12.5	36.1	79.6 ($\uparrow$ 6.1)	11.8	24.7	63.8 ($\uparrow$ 9.2)	9.6

