NEURAL INFORMATION PROCESSING SYSTEMS

CLAP4CLIP: Continual Learning with Probabilistic Finetuning for Vision-Language Models

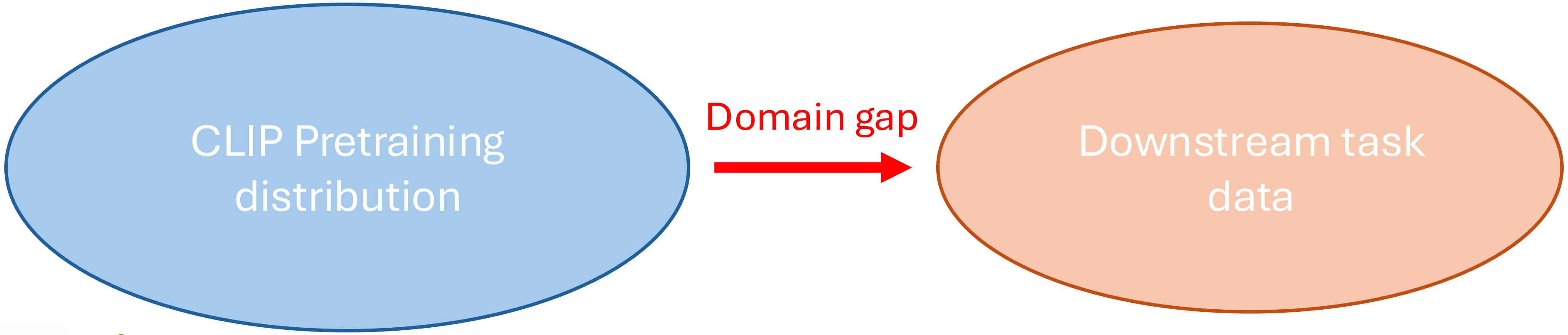
UNSW THE UNIVERSITY OF NEW SOUTH WALES

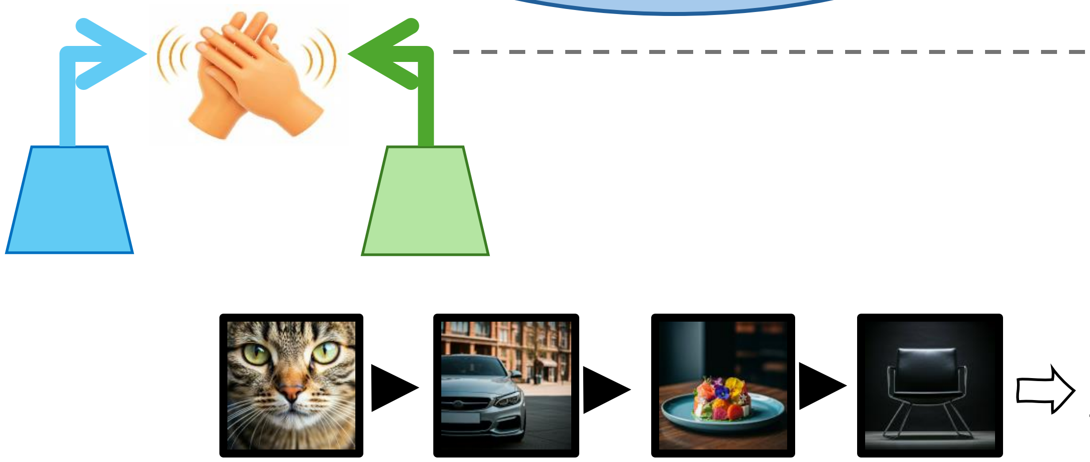
Saurav Jha, Dong Gong, Lina Yao

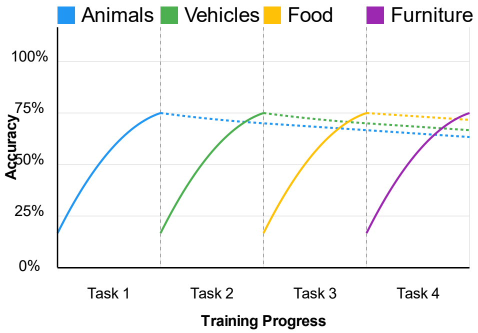
DATA 61 CSIRO

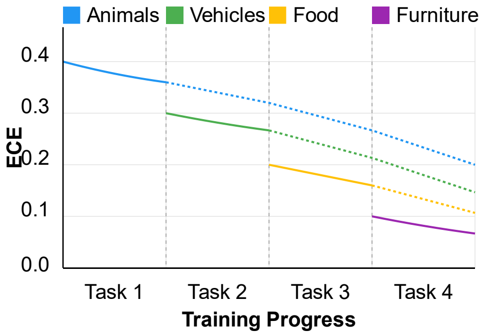
Code: https://github.com/srvCodes/clap4clip

(Continual) Finetuning motivation









Existing finetuning approaches are deterministic

Visual variations for "dog" class

"A photo of a dog"

"A dog playing at the beach"

"A close-up of a dog indoors"

Textual variations for "dog" class

Deterministic finetuning approaches risk:

- Overfitting to specific combinations
- Loss of generalizable knowledge

Probabilistic finetuning approaches

- Model the distribution of image/text cues
- Sampling from such distribution can help capture various image-text interactions, and hence generalize better
- Probabilistic finetuning approaches however sacrifice in-domain performance [1]:

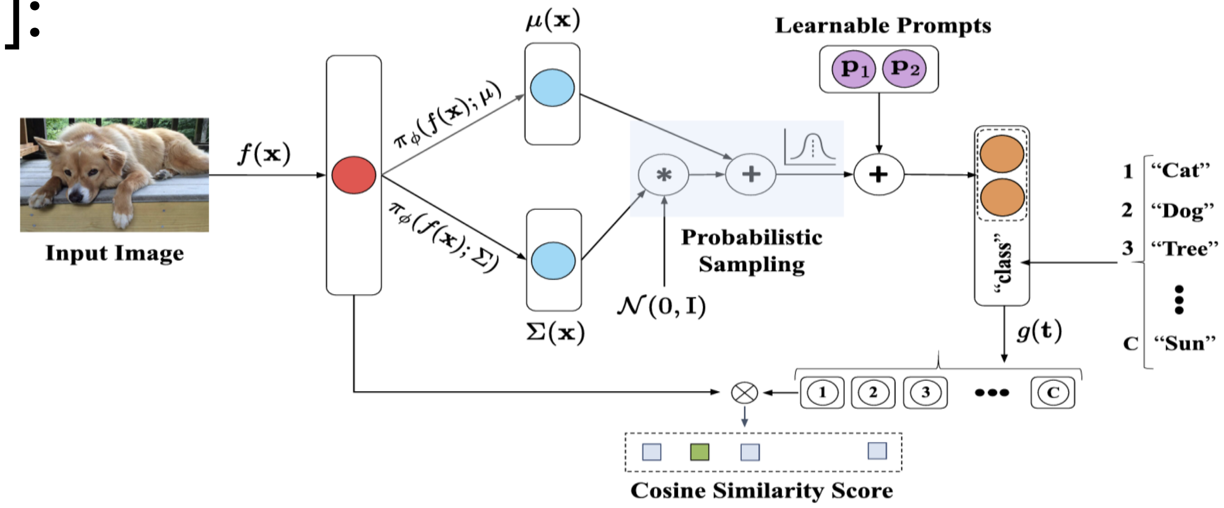
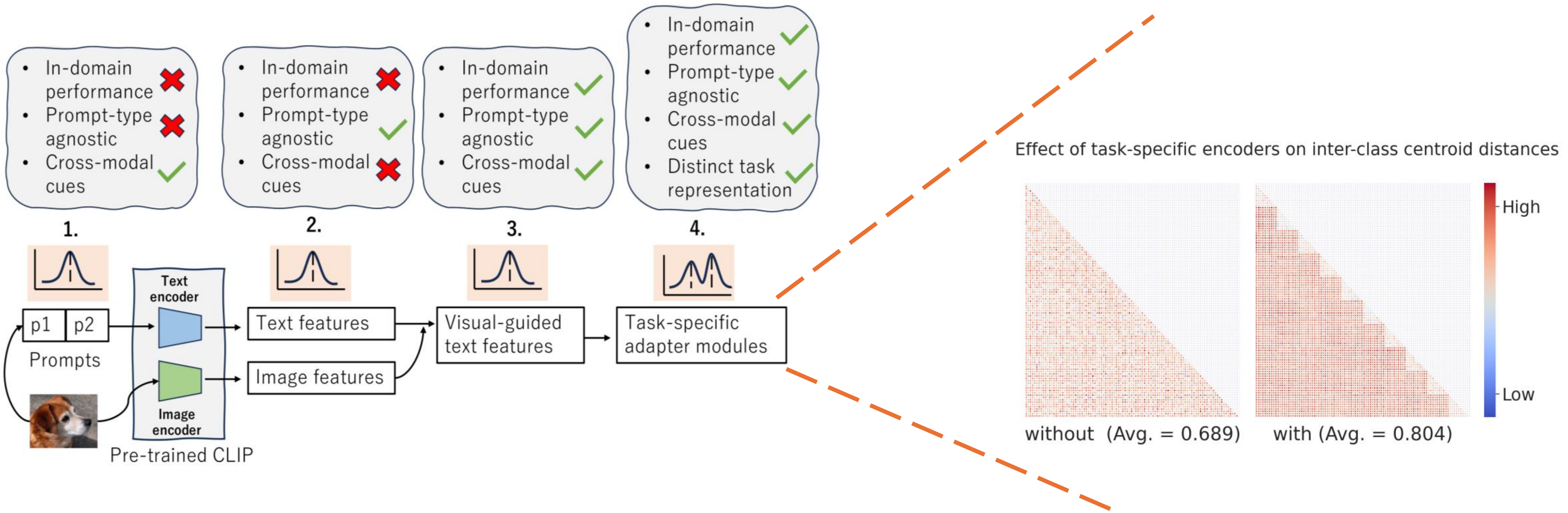


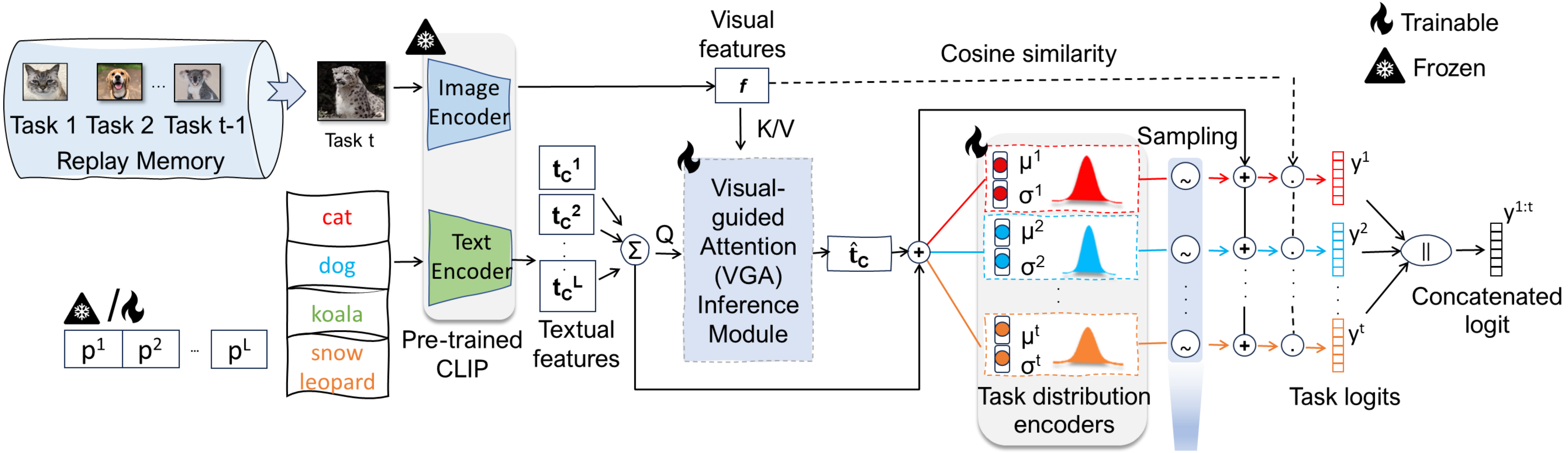
Image source: [1]

[1] Derakhshani *et al.* "Variational Prompt Tuning Improves Generalization of Vision-Language Models"

Candidate spaces for probabilistic modeling



CLAP: Variational modelling over VGA outputs



Why model visual-guided text features?

- We analyze the effect of CL on the spatial geometry of cross-modal features
- The rotation angle arccos<*t*, 1>, where *t* = test features of 1st test task after step *t*

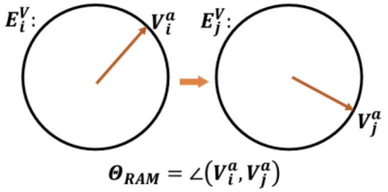
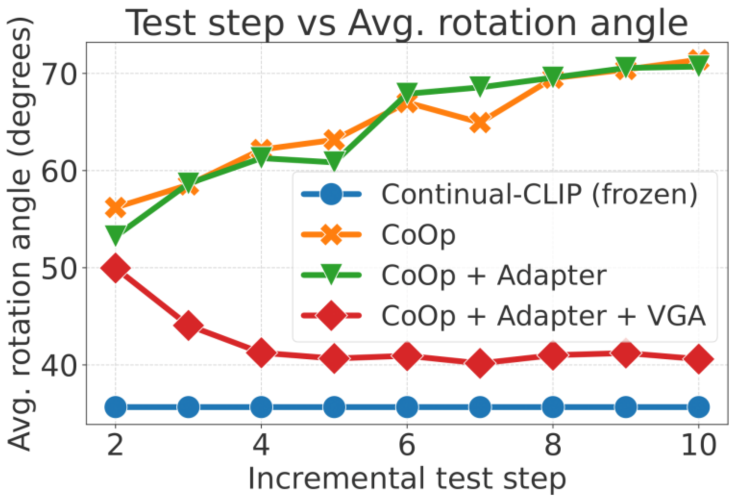


Image source: [1]

- Introducing a Visual-guided Adapter (VGA) module for alignment:



[1] Ni *et al.* "Continual Vision-Language Representation Learning with Off-diagonal Information"

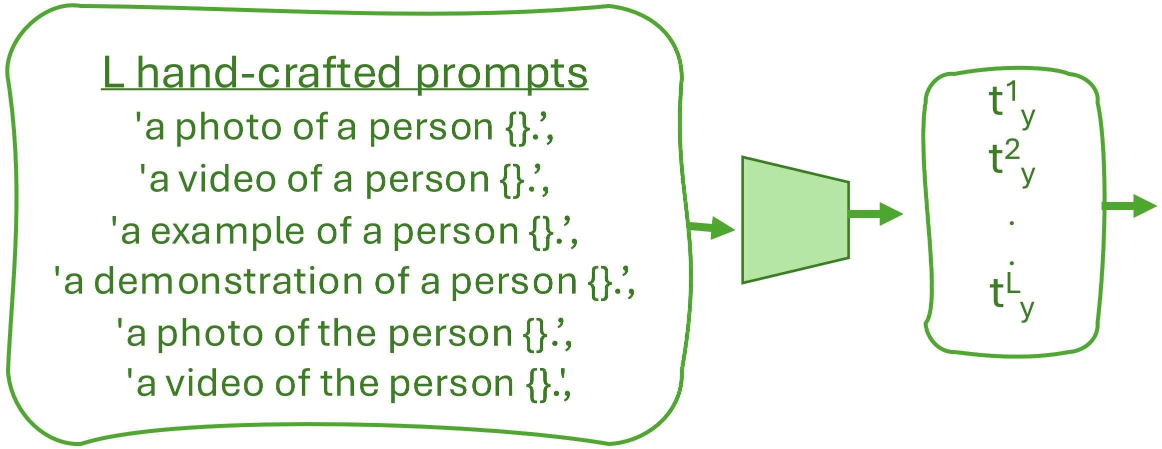
Can we do better against forgetting?

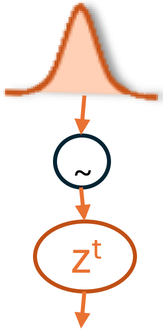
- We know that CLIP comes with rich pre-trained knowledge
- This helps in swift construction of task-specific hand-crafted prompts that perform well in general
- Can we leverage such hand-crafted prompts to counter forgetting?

'a photo of a person {}.',
'a video of a person {}.',
'a example of a person {}.',
'a demonstration of a person {}.',
'a photo of the person {}.',
'a video of the person {}.',

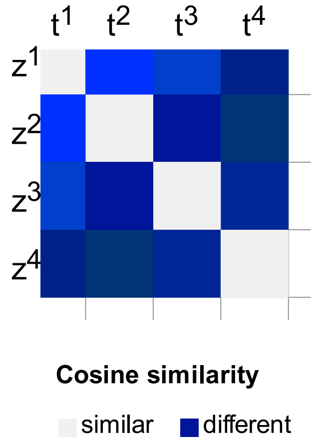
Pretrained language knowledge for countering forgetting

1. Past-task distribution regularization





$$P_{\mathrm{KD}}(y|z^t) = \frac{1}{M}\sum_{m=1}^{M}\frac{1}{L}\sum_{l=1}^{L}\frac{\exp\left(\langle \mathbf{t}_y^{h,l}, z_m^t\rangle\right)}{\sum_{c=1}^{|C^t|}\exp\left(\langle \mathbf{t}_c^{h,l}, z_m^t\rangle\right)}$$



Pretrained language knowledge for countering forgetting

2. Weight initialization for mitigating stability gap [1]

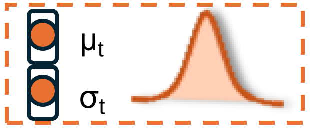
$w\_t = \mathbb{R}^{d \times d}$
$s = \mathbb{R}^{|C^t| \times d}$

s_μ, s_σ = mean, std. dev. of *L* task-specific text features

<u>L hand-crafted prompts</u>
'a photo of a person {}.',
'a video of a person {}.',
'a example of a person {}.',
'a demonstration of a person {}.',
'a photo of the person {}.',
'a video of the person {}.',

t^1_y
t^2_y
.
.
t^L_y

$$\mathbf{w}_t^{\mu} = \frac{1}{d}\langle \mathbf{s}_\mu^T, \mathbf{s}_\mu \rangle, \quad \mathbf{w}_t^{\sigma} = \frac{1}{d}\langle \mathbf{s}_\sigma^T, \mathbf{s}_\sigma \rangle$$

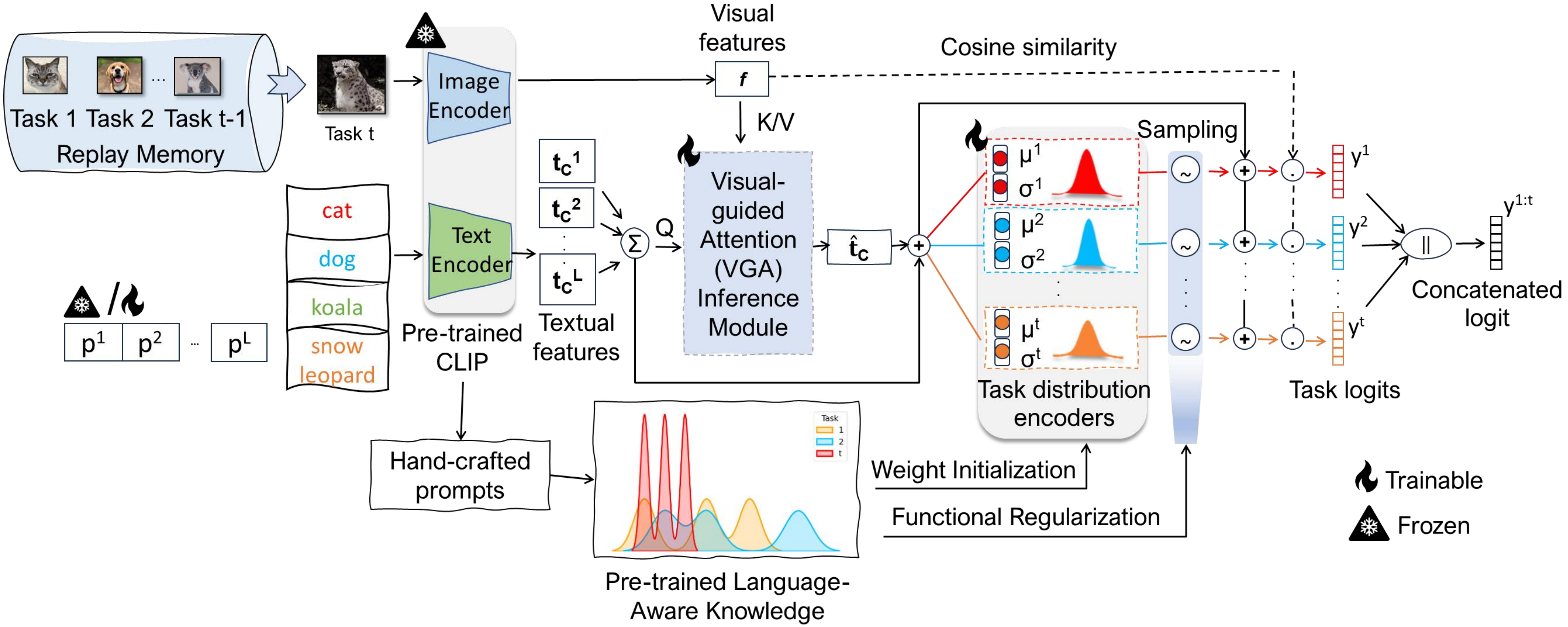


Initialize t-th task adapter weights

---

[1] Harun *et al.* “Overcoming the stability gap in continual learning”

CLAP with pre-trained knowledge



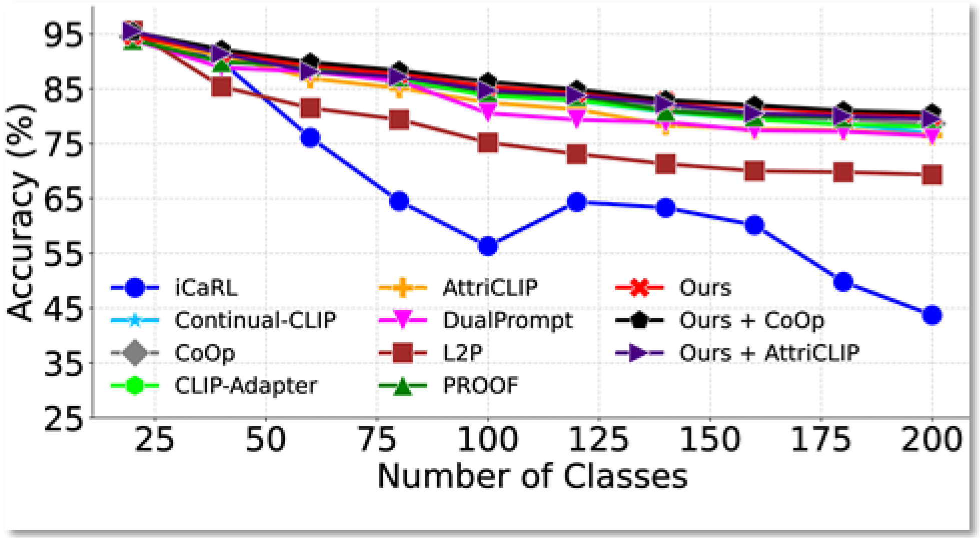
Evaluation

- On five class-incremental dataset setups
- We incorporate CLAP with:
    1. Hand-crafted prompts (Continual-CLIP)
    2. Task-conditioned learnable prompts (CoOp)
    3. Instance-conditioned learnable prompts (AttriCLIP)
    4. Multimodal prompts (MaPLe)

Average incremental accuracy

- T = number of tasks, C/T = number of classes per task

Model	**CIFAR-100 (10 T, 10 C/T)**	**ImageNet-R (10 T, 20 C/T)**	**V-TAB (5 T, 10 C/T)**
CODA-P	85.19	82.06	87.5
Continual-CLIP	78.65	84.43	68.5
+ Ours	**86.13**	85.77	91.37
CoOp	81.17	84.7	87.06
+ Ours	85.71	85.32	**92.51**
MaPLe	82.74	85.28	83.91
+ Ours	86.06	86.25	90.97
AttriCLIP	79.31	83.09	71.84
+ Ours	78.06	**86.35**	74.84



Further robust evaluations

- Calibration (Expected Calibration Error)

Model	ImageNet-R	V-TAB
CoOp	0.191	0.191
+ Ours	0.207	**0.136**

- Forgetting (Backward transfer)

Model	ImageNet-R	V-TAB
CoOp	-0.12	-0.007
+ Ours	**-0.112**	**0.011**

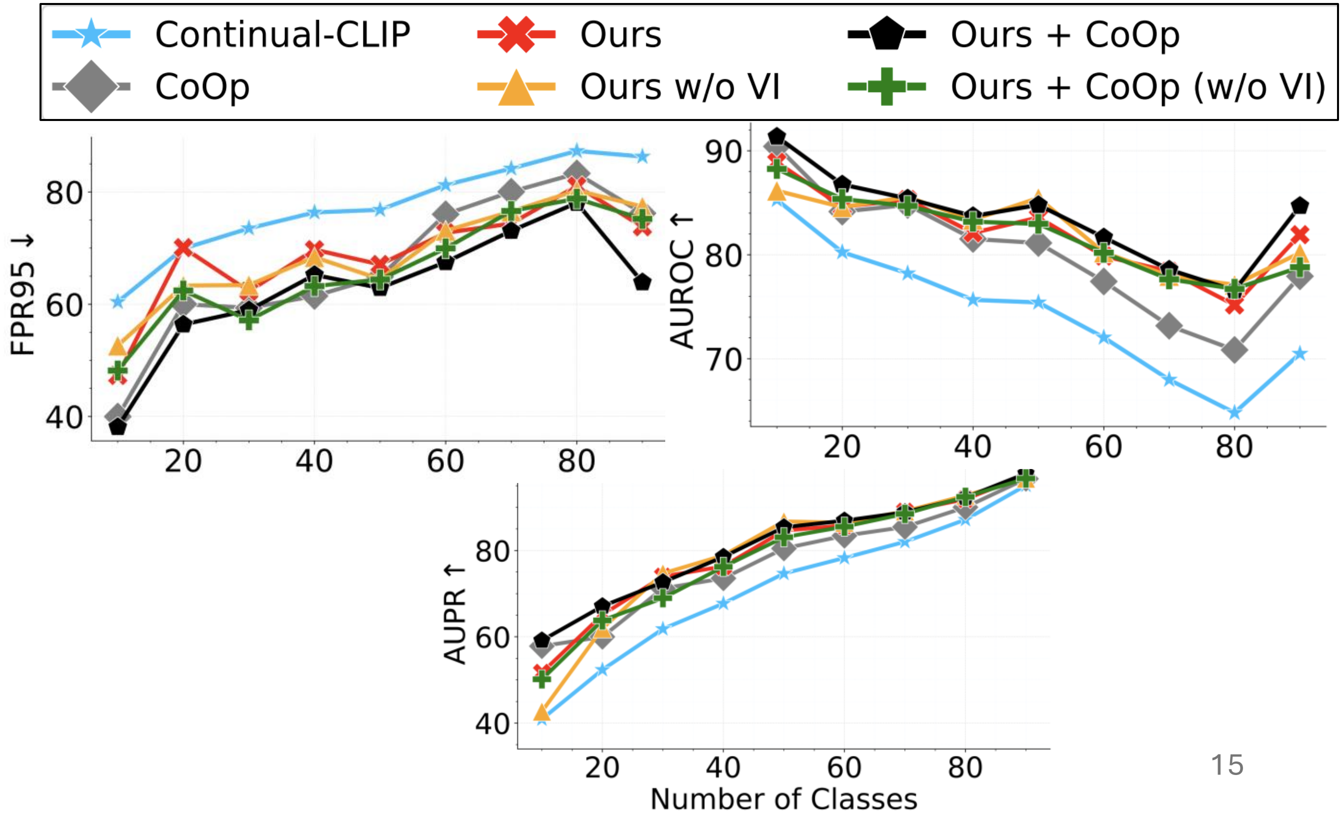
- Generalization (Forward transfer)

Model	ImageNet-R	V-TAB
CoOp	60.93	69.38
+ Ours	**63.44**	**74.1**

Perks of probabilistic modelling

1. Post-hoc Novel Data Detection (PhNDD)
   - At step t, treat all seen (i <= t) test data as in-domain
   - Treat all the future tasks data as novel
   - Energy score of prediction quantifies the model's confidence score

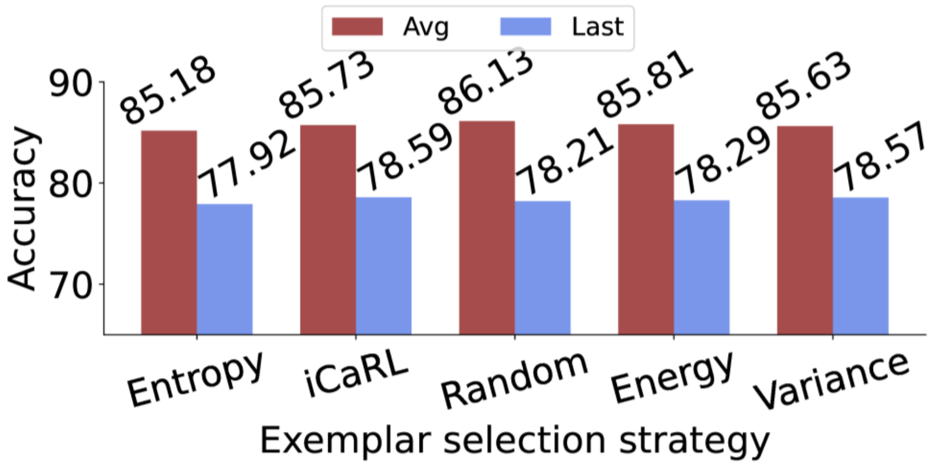
Model	AUROC↑	AUPR↑	FPR95↓
Continual-CLIP	74.46	71.11	77.33
Ours w/o VI	82.29	78.88	68.83
+ CLAP (Ours)	82.21	79.54	68.72
CoOp	80.15	77.62	66.8
+ CLAP w/o VI	81.98	78.88	66.21
+ CLAP	83.73	80.97	62.68



Perks of probabilistic modelling

2. Uncertainty-based exemplar selection
   - Select replay exemplars based on the entropy of CLAP's predictions
   - Deterministic methods are known to perform subpar at this [1]

Model	Avg	Last
CoOp	76.71	64.1
CLIP-Adapter	78.78	68.49
Ours w/o VI	84.44	76.55
Ours	**85.18**	**77.92**



[1] Chaudhry *et al*. "Riemannian walk for incremental learning: Understanding forgetting & intransigence"

Conclusion

- We propose CLAP*4*CLIP, a probabilistic continual finetuning framework for the pre-trained CLIP model
- CLAP supports a diverse range of prompts: hand-crafted, task-conditioned, instance-conditioned, and multi-modal
- For these prompt types, CLAP can help enhance the in-domain performances as well as out-of-domain generalization
- We show out-of-the-box utilities of CLAP's probabilistic nature for post-hoc novel data detection and uncertainty-based exemplar selection