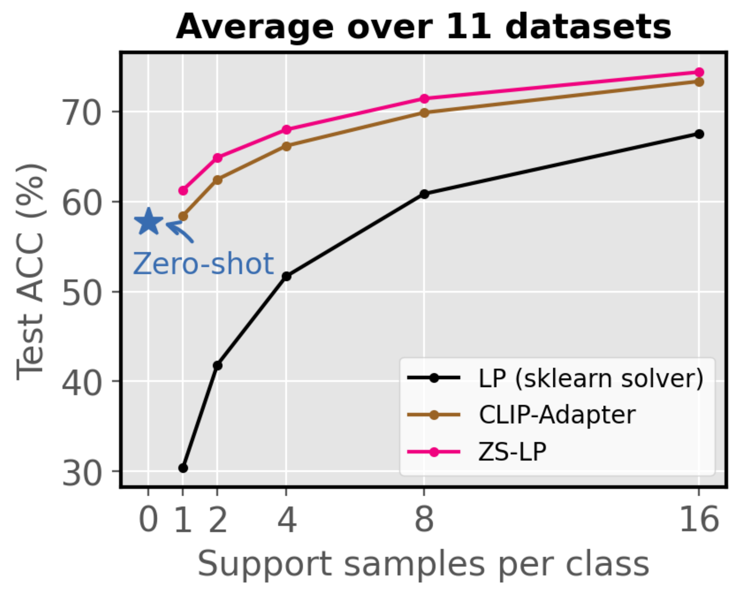
A Closer Look at the Few-Shot Adaptation of Large Vision-Language Models

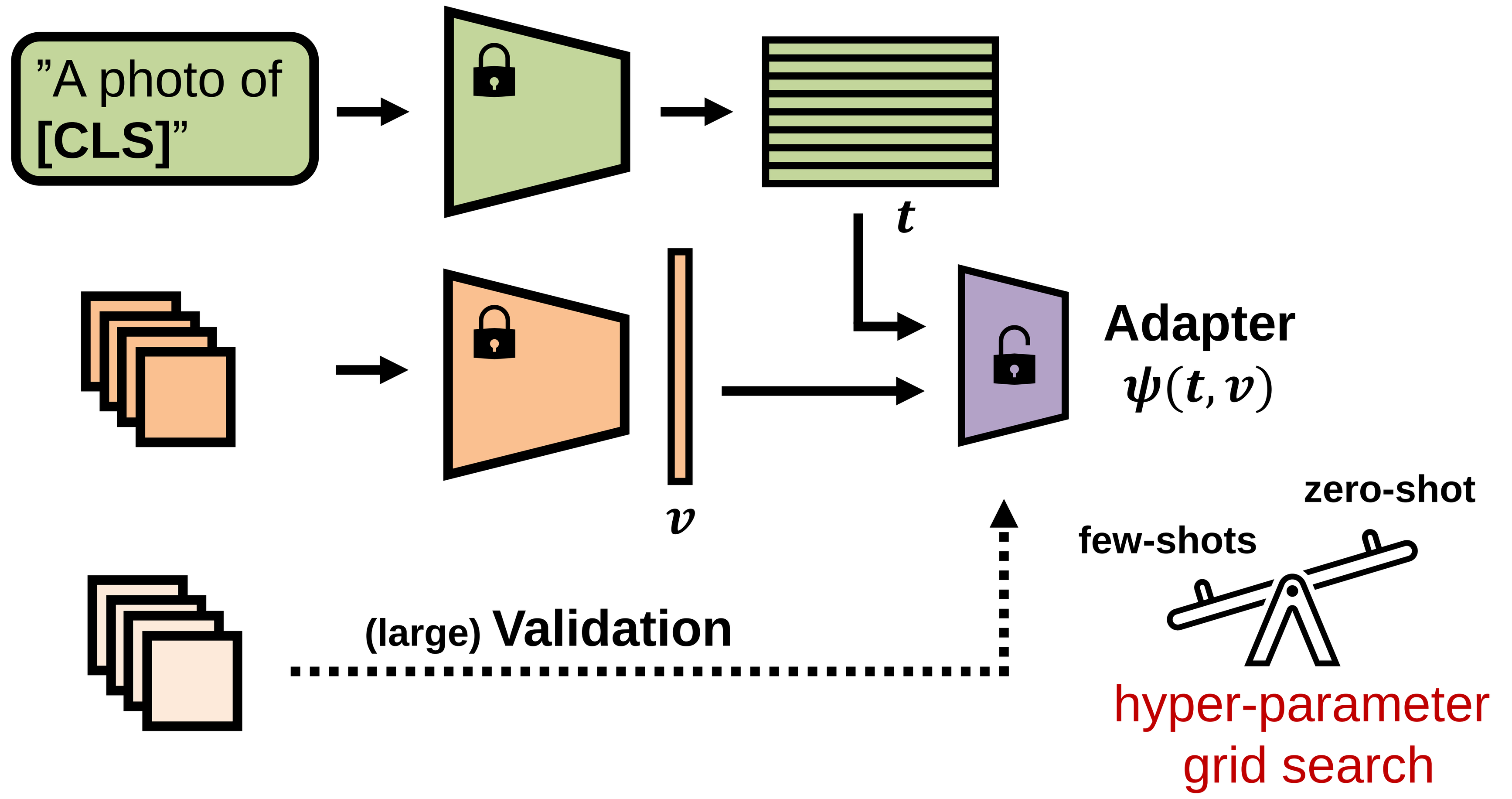
Julio Silva-Rodríguez · Sina Hajimiri · Ismail Ben Ayed · Jose Dolz ÉTS Montreal

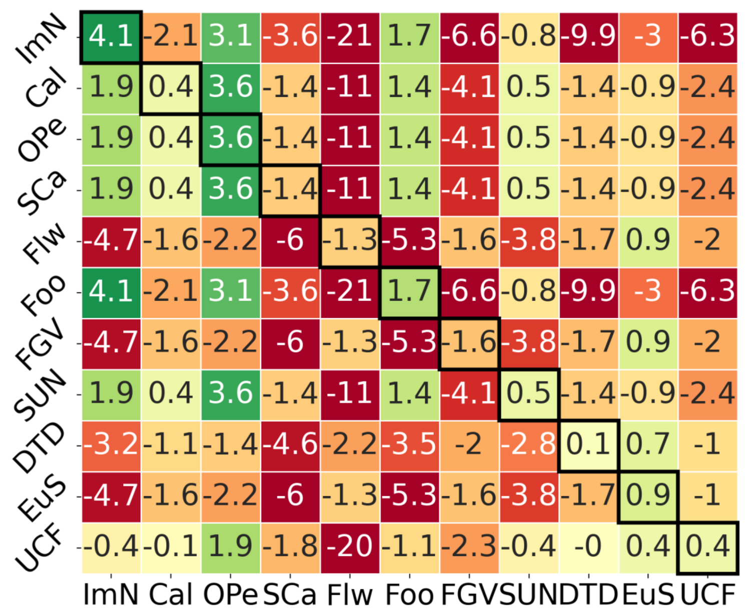
VLMs Adaptation

- VLMs. present **robust zero-shot** performance
- **Few-shot Adapters** enhance the transferability **combining visual and text information**.



Pitfalls on Existing Few-Shot Adapters





CLIP-Adapter *vs.* (ZS) Linear Probe

How realistic is using a validation set during few-shot adaptation?

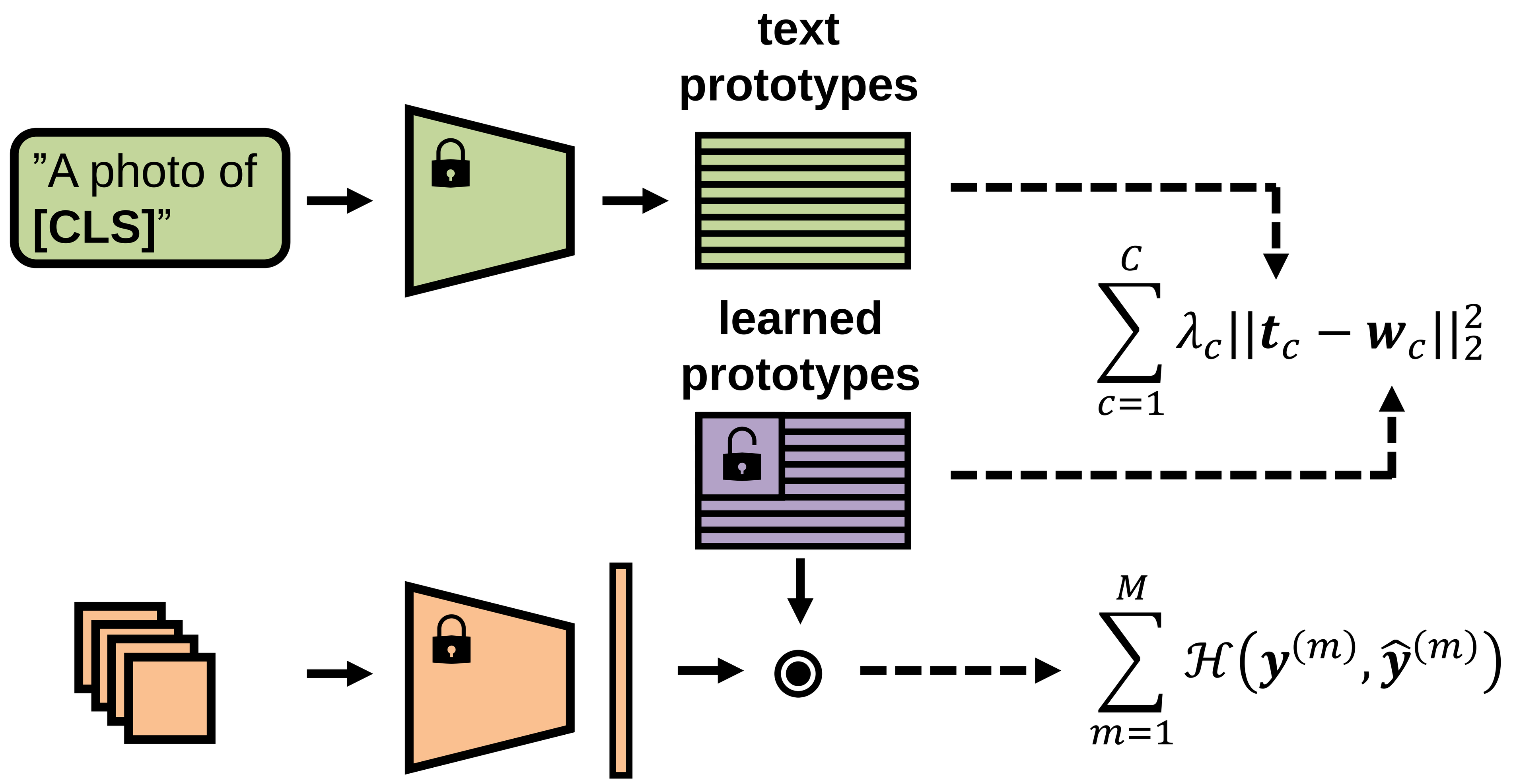
Otherwise, SoTA Adapters struggle to ourperform a simple well-initialized Linear Probe

Our Few-Shot Adapter: CLAP

- We propose a novel and simple approach that meets challenges of real-world scenarios: **not requiring hyper-parameter tuning**.
- We introduce **CL**ass-**A**daptive linear **P**robe (**CLAP**), a linear classifier with **prototypes constrained to remain close to the initial, robust zero-shot prototypes.**

$$\min_{\mathbf{w}} \underbrace{\sum_{m=1}^{M} \mathcal{H}\left(\mathbf{y}^{(m)}, \hat{\mathbf{y}}^{(m)}\right)}_{\text{Cross-entropy on few shots}} + \underbrace{\sum_{c=1}^{C} \lambda_c ||\mathbf{t}_c - \mathbf{w}_c||_2^2}_{\text{Learned prototypes constrained to zero-shot}}$$

- For each class, $\boldsymbol{\lambda_c}$ **is fixed using zero-shot performance on support samples.** Thus, better performance → larger $\boldsymbol{\lambda_c}$.



SoTA Adapters Comparisons

Validation-free comparison

Method	K=1	K=2	K=4	K=8	K=16
Prompt-learning methods					
CoOp IJCV'22 [46]	59.56	61.78	66.47	69.85	73.33
ProGrad ICCV'23 [13]	62.61	64.90	68.45	71.41	74.28
PLOT ICLR'23 [6]	62.59	65.23	68.60	71.23	73.94
Efficient transfer learning - *a.k.a Adapters*					
Zero-Shot ICML'21 [30]	57.71	57.71	57.71	57.71	57.71
Rand. Init LP ICML'21 [30]	30.42	41.86	51.69	60.84	67.54
CLIP-Adapter IJCV'23 [11]	58.43	62.46	66.18	69.87	73.35
TIP-Adapter ECCV'22 [42]	58.86	60.33	61.49	63.15	64.61
TIP-Adapter(f) ECCV'22 [42]	60.29	62.26	65.32	68.35	71.40
CrossModal-LP CVPR'23 [24]	62.24	64.48	66.67	70.36	73.65
TaskRes(e) CVPR'23 [40]	61.44	65.26	68.35	71.66	74.42
ZS-LP	61.28	64.88	67.98	71.43	74.37
CLAP	**62.79**	**66.07**	**69.13**	**72.08**	**74.57**

Using a few-shot validation set

Method	K=1	K=2	K=4	K=8	K=16
Protocol in [24]: K-shots for train + $min(K, 4)$ for validation					
TIP-Adapter [42]	63.3	65.9	69.0	72.2	75.1
CrossModal LP [24]	64.1	67.0	70.3	73.0	76.0
CrossModal Adapter [24]	64.4	67.6	70.8	73.4	75.9
CrossModal PartialFT [24]	64.7	67.2	70.5	**73.6**	**77.1**
Ours: using $K + min(K, 4)$ shots for training					
ZS-LP	64.9	68.0	71.4	73.1	75.0
CLAP	**66.1**	**69.1**	**72.1**	**73.5**	75.1

Conclusions

- Few-shot adapters should include **model selection strategies based on support data**.
- **CLAP** is largely competitive and **does not require ad-hoc adjustments per dataset.**

Know more!

jusiro/CLAP