Transductive Zero-Shot & Few-Shot Adaptation of Large Medical Vision Language Models


Ismail Ben Ayed


Julio Silva-Rodriguez


Maxime Zanella


Yunshi Huang


Jose Dolz


Houda Bahig

Fereshteh Shakeri
Jan 2026

ÉTS
Le génie pour l'industrie

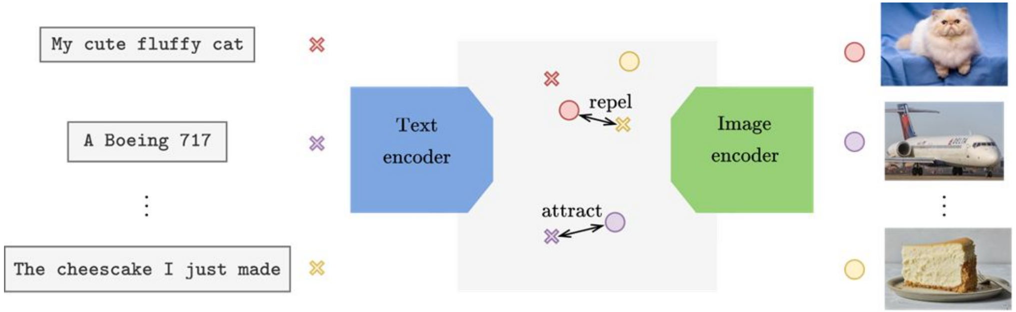
CRCHUM
CENTRE DE RECHERCHE

ILLS
International Laboratory
on Learning Systems

Lots of different Vision-Language Models out there

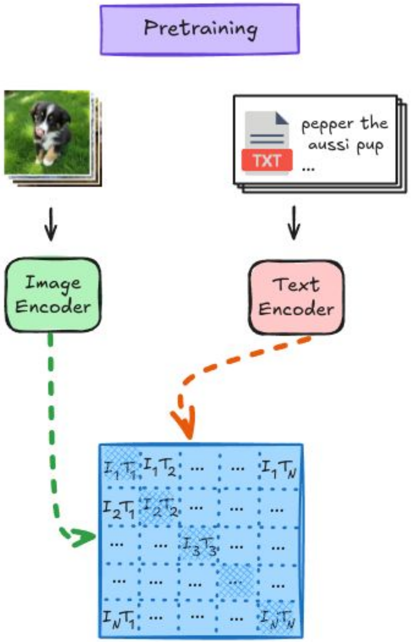
Among them…

- **CLIP: generic text-image pairs of internet**



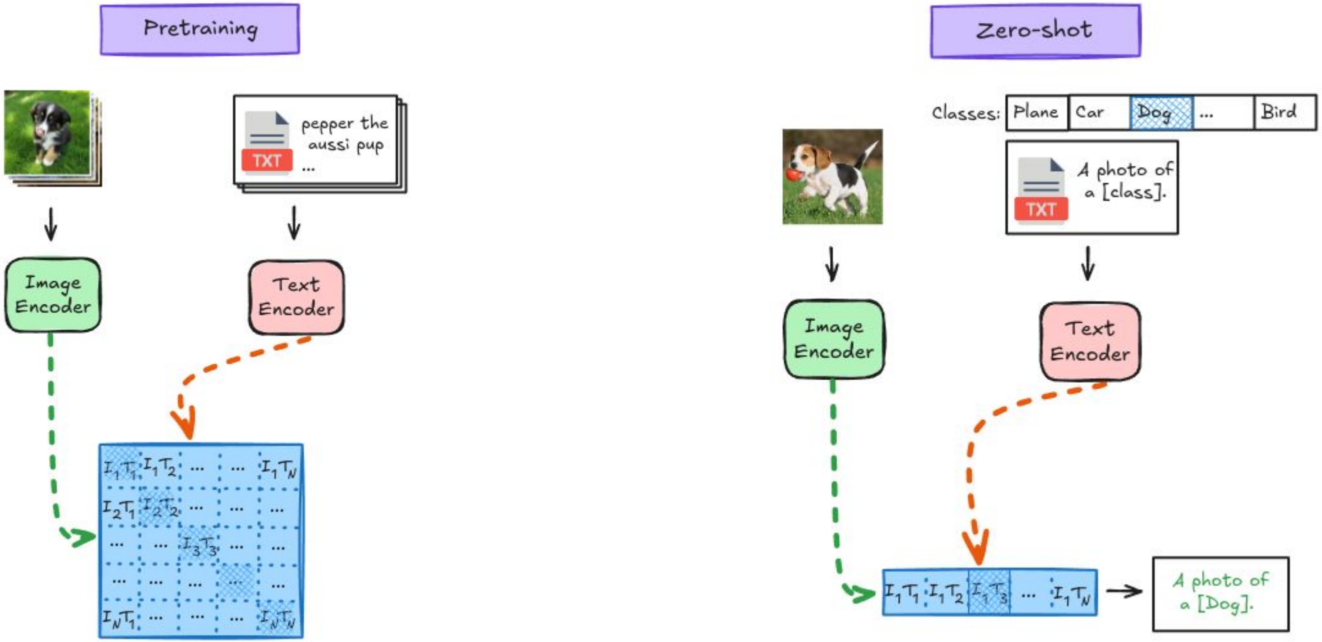
CLIP: Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748-8763). PMLR.

Generalist Vision-language Models (CLIP)



Radford, Alec, et al. "Learning transferable visual models from natural language supervision." PMLR, 2021.

Generalist Vision-language Models (CLIP)

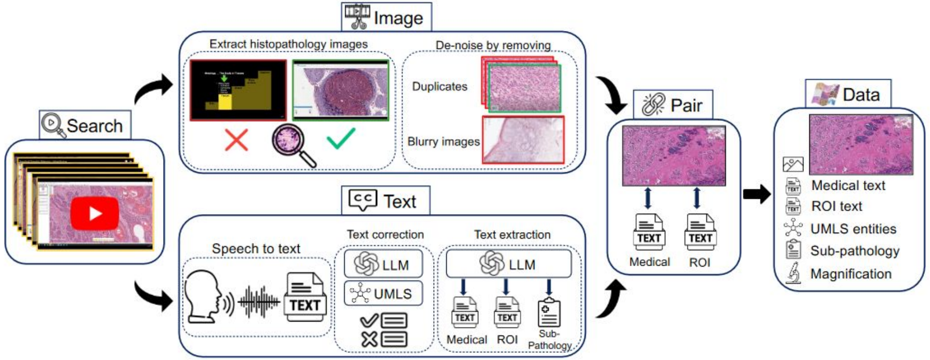


Radford, Alec, et al. "Learning transferable visual models from natural language supervision." PMLR, 2021.

Lots of different Vision-Language Models out there

Among them…

- CLIP: generic text-image pairs of internet
- **Quilt: from Youtube**



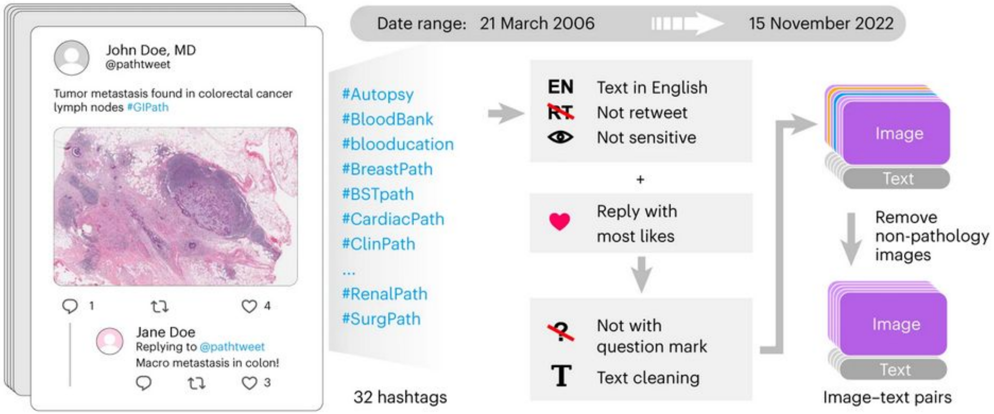
CLIP: Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748-8763). PMLR.

Quilt: Ikezogwo, W., Seyfioglu, S., Ghezloo, F., Geva, D., Sheikh Mohammed, F., Anand, P. K., ... & Shapiro, L. (2024). Quilt-1m: One million image-text pairs for histopathology. *Advances in neural information processing systems*, *36*.

Lots of different Vision-Language Models out there

Among them…

- CLIP: generic text-image pairs of internet
- Quilt: from Youtube
- **PLIP: from Twitter**



CLIP: Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748-8763). PMLR.

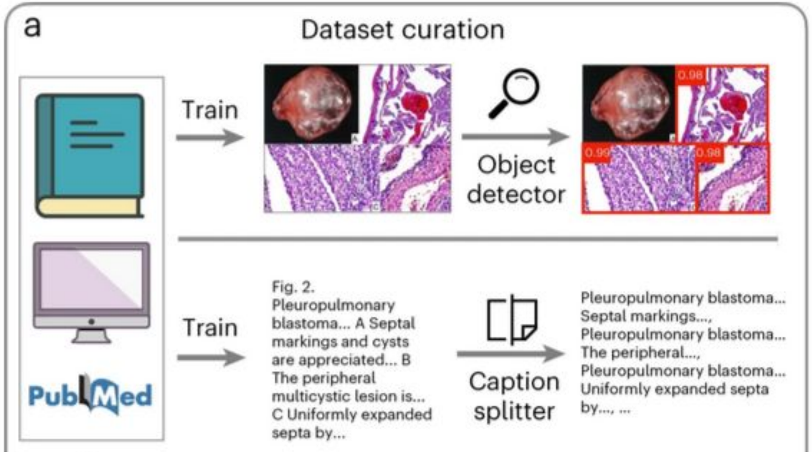
Quilt: Ikezogwo, W., Seyfioglu, S., Ghezloo, F., Geva, D., Sheikh Mohammed, F., Anand, P. K., ... & Shapiro, L. (2024). Quilt-1m: One million image-text pairs for histopathology. *Advances in neural information processing systems*, *36*.

PLIP: Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T. J., & Zou, J. (2023). A visual–language foundation model for pathology image analysis using medical twitter. *Nature medicine*, *29*(9), 2307-2316.

Lots of different Vision-Language Models out there

Among them…

- CLIP: generic text-image pairs of internet
- Quilt: from Youtube
- PLIP: from Twitter
- **CONCH : from PubMed**



CLIP: Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748-8763). PMLR.

Quilt: Ikezogwo, W., Seyfioglu, S., Ghezloo, F., Geva, D., Sheikh Mohammed, F., Anand, P. K., ... & Shapiro, L. (2024). Quilt-1m: One million image-text pairs for histopathology. *Advances in neural information processing systems*, *36*.

PLIP: Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T. J., & Zou, J. (2023). A visual–language foundation model for pathology image analysis using medical twitter. *Nature medicine*, *29*(9), 2307-2316.

CONCH:Lu, M. Y., Chen, B., Williamson, D. F., Chen, R. J., Liang, I., Ding, T., ... & Mahmood, F. (2024). A visual-language foundation model for computational pathology. *Nature Medicine*, *30*(3), 863-874.

Lots of different Vision-Language Models out there

Among them…

- CLIP: generic text-image pairs of internet
- Quilt: from Youtube
- PLIP: from Twitter
- CONCH : from PubMed

How to leverage their representational power?

…

Boosting Vision-Language Models for Histopathology Classification: Predict all at once

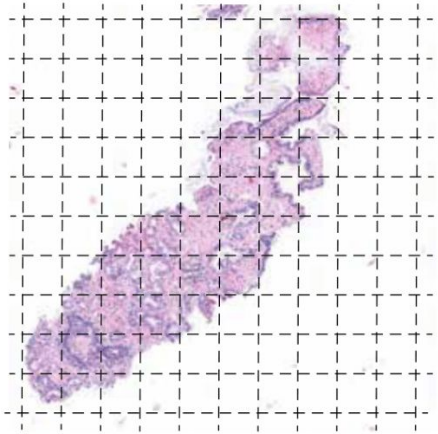
CLIP: Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748-8763). PMLR.

Quilt: Ikezogwo, W., Seyfioglu, S., Ghezloo, F., Geva, D., Sheikh Mohammed, F., Anand, P. K., ... & Shapiro, L. (2024). Quilt-1m: One million image-text pairs for histopathology. *Advances in neural information processing systems*, *36*.

PLIP: Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T. J., & Zou, J. (2023). A visual–language foundation model for pathology image analysis using medical twitter. *Nature medicine*, *29*(9), 2307-2316.

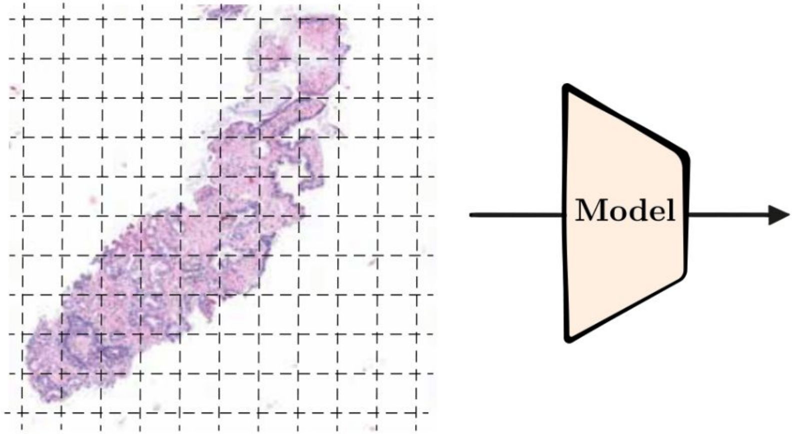
CONCH:Lu, M. Y., Chen, B., Williamson, D. F., Chen, R. J., Liang, I., Ding, T., ... & Mahmood, F. (2024). A visual-language foundation model for computational pathology. *Nature Medicine*, *30*(3), 863-874.

Inductive inference



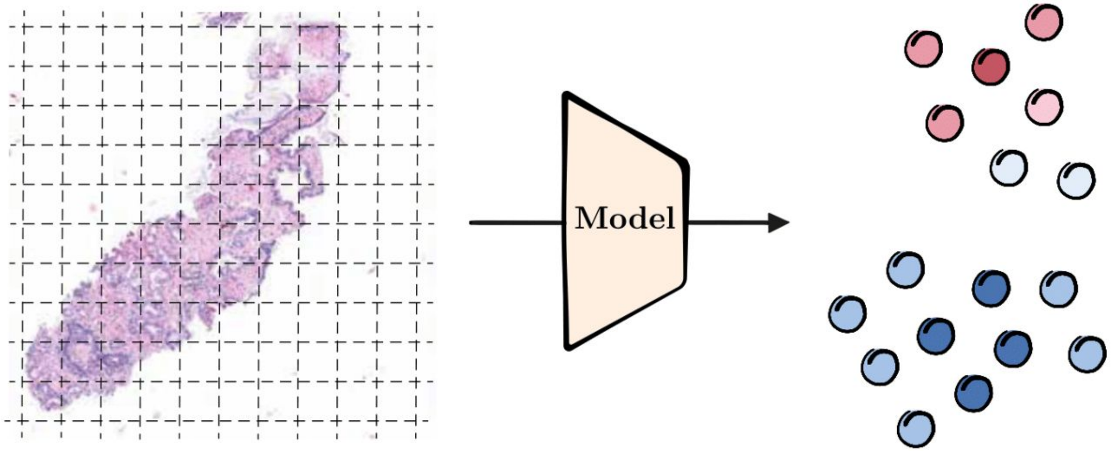
Large whole slides divided per patches, large-scale database, …

Inductive inference



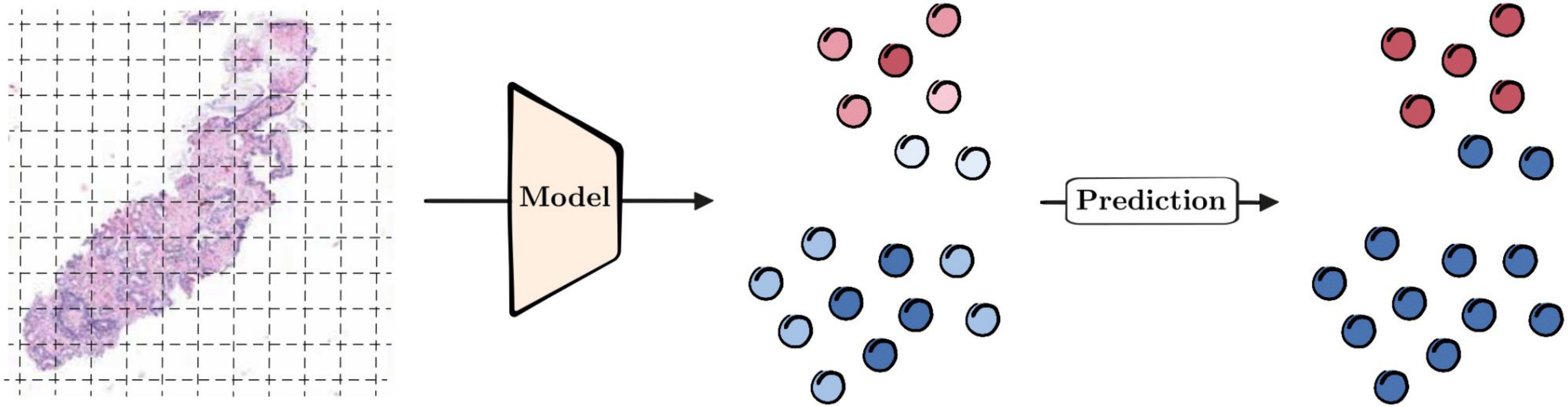
(1) If we have **enough annotated data**, we can train a Model and use it at inference time…

Inductive inference



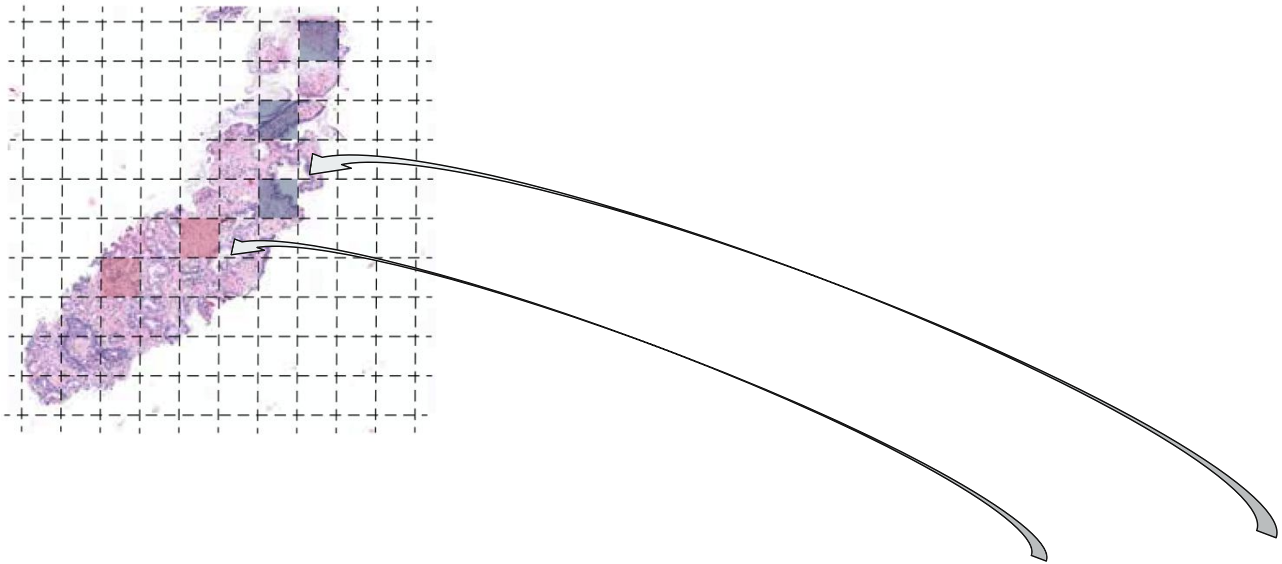
… which gives us more or less confident predictions …

Inductive inference



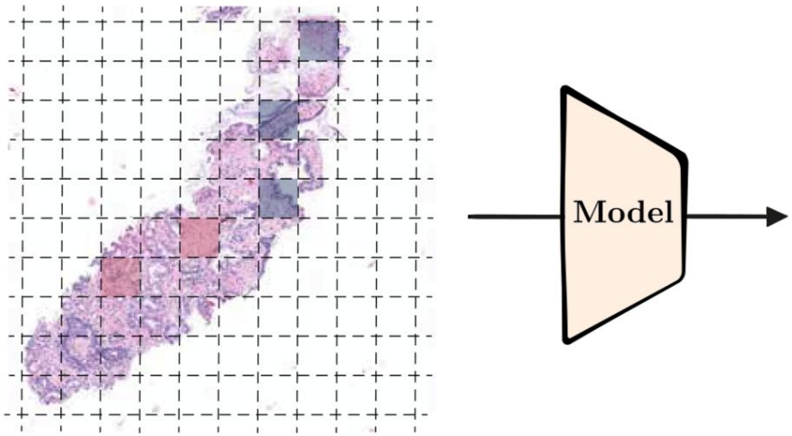
… and we take the most probable class for **each patch independently**.

Few-shot transductive inference



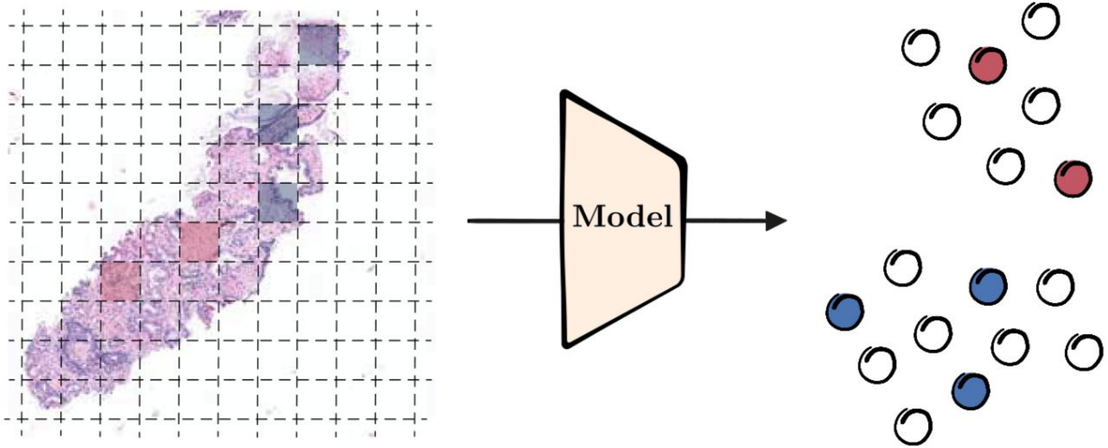
Large whole slides divided per patches, large-scale database, … **+ a few annotated patches**

Few-shot transductive inference



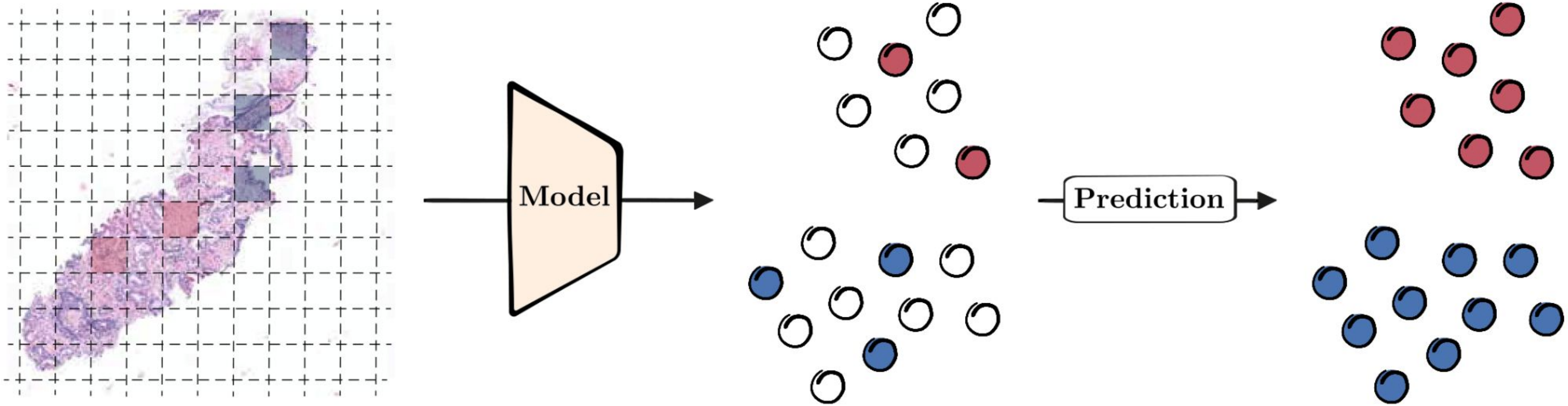
(2) We can use a good embedder (even self-supervised!) …

Few-shot transductive inference



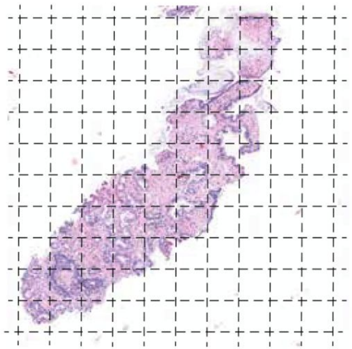
… which gives a good representation …

Few-shot transduction inference



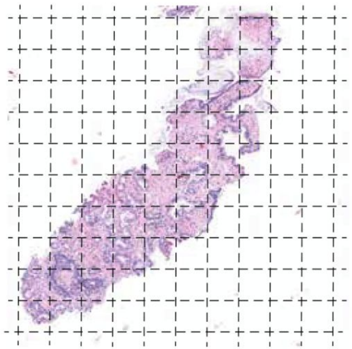
… we can leverage **data structure** and the few-labeled patches !

Unsupervised transductive inference



Large whole slides divided per patches, large-scale database, … ~~+ a few annotated patches~~

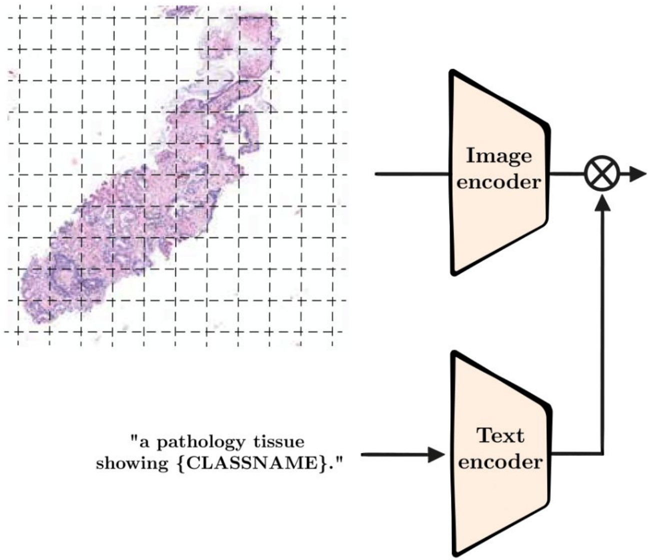
Unsupervised transductive inference



"a pathology tissue
showing {CLASSNAME}."

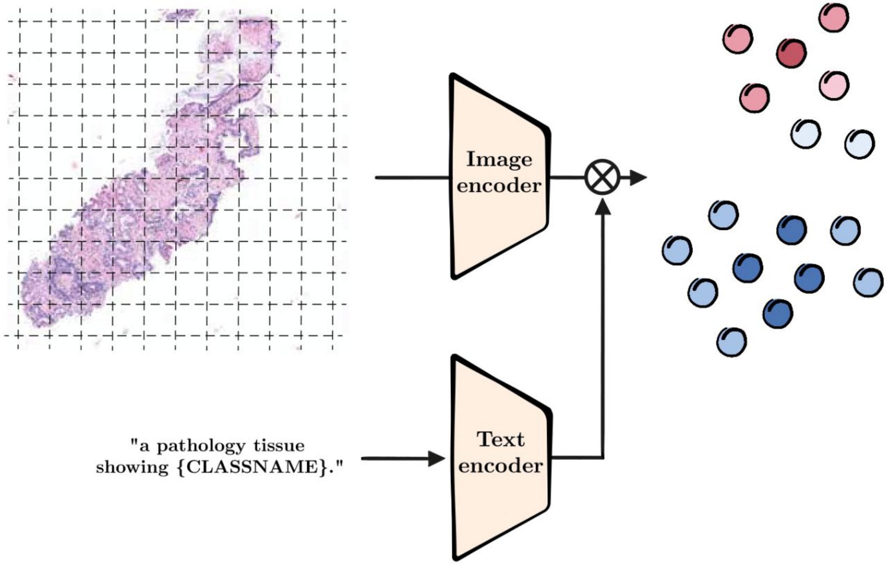
Large whole slides divided per patches, large-scale database, … **+ text description for each class**

Unsupervised transductive inference



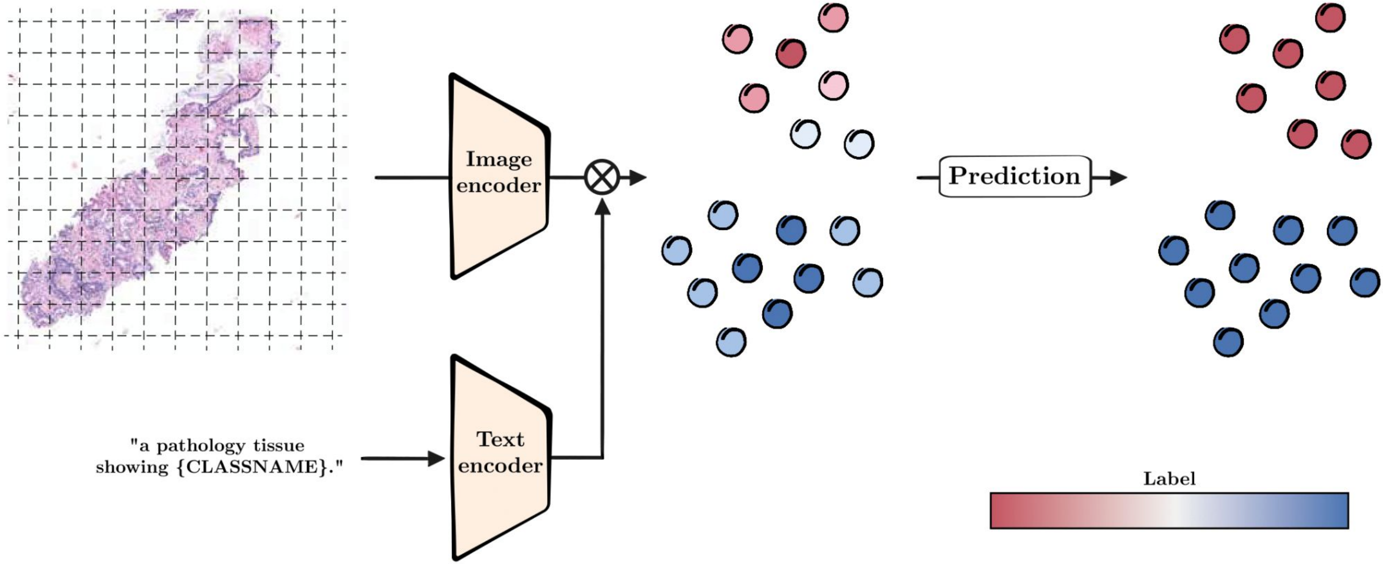
(3) we leverage **Vision-Language Models** (VLMs) trained on large-scale unsupervised text-image data…

Unsupervised transductive inference



… text-image similarities give us (noisy) **zero-shot predictions** for each patch …

Unsupervised transductive inference



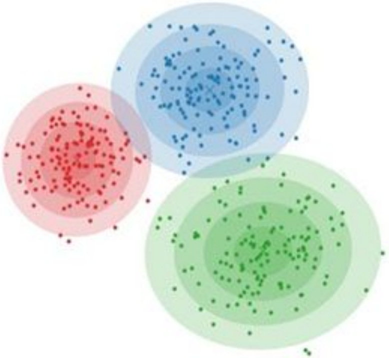
… these **zero-shot predictions** can be used in combination with the **data structure!**

Our algorithm (in short)

$$\mathcal{L}_{\text{Zero-Shot}}(\mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \underbrace{-\frac{1}{|\mathcal{Q}|}\sum_{i \in \mathcal{Q}} \mathbf{z}_i^\top \log(\mathbf{p}_i)}_{\text{GMM clustering}} \underbrace{- \sum_{i \in \mathcal{D}} \sum_{j \in \mathcal{D}} w_{ij} \mathbf{z}_i \mathbf{z}_j}_{\text{Laplacian reg.}} + \underbrace{\sum_{i \in \mathcal{Q}} \mathrm{KL}_\lambda(\mathbf{z}_i || }_{\text{Text knowledge}}$$

Our algorithm (in short)

$$\mathcal{L}_{\text{Zero-Shot}}(\mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \underbrace{-\frac{1}{|\mathcal{Q}|}\sum_{i\in\mathcal{Q}} \mathbf{z}_i^\top \log(\mathbf{p}_i)}_{\text{GMM clustering}} \underbrace{- \sum_{i\in\mathcal{D}}\sum_{j\in\mathcal{D}} w_{ij}\mathbf{z}_i\mathbf{z}_j}_{\text{Laplacian reg.}} + \underbrace{\sum_{i\in\mathcal{Q}} \mathrm{KL}_\lambda(\mathbf{z}_i||\hat{\mathbf{y}}_i)}_{\text{Text knowledge}}$$



We model features as a balanced Gaussian Mixture Model (**GMM**).

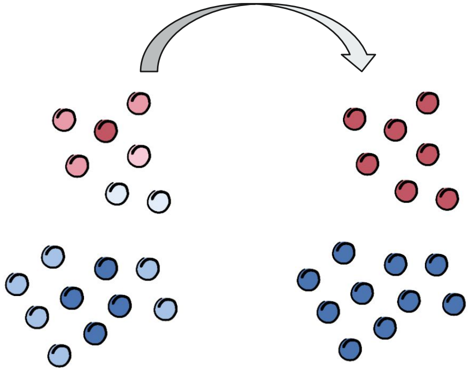
Our algorithm (in short)

$$\mathcal{L}_{\text{Zero-Shot}}(\mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \underbrace{-\frac{1}{|\mathcal{Q}|}\sum_{i\in\mathcal{Q}} \mathbf{z}_i^\top \log(\mathbf{p}_i)}_{\text{GMM clustering}} \underbrace{- \sum_{i\in\mathcal{D}}\sum_{j\in\mathcal{D}} w_{ij}\mathbf{z}_i\mathbf{z}_j}_{\text{Laplacian reg.}} + \underbrace{\sum_{i\in\mathcal{Q}} \mathrm{KL}_\lambda(\mathbf{z}_i || \hat{\mathbf{y}}_i)}_{\text{Text knowledge}}$$

Similar patches should have **similar predictions**!

Our algorithm (in short)

$$\mathcal{L}_{\text{Zero-Shot}}(\mathbf{z},\boldsymbol{\mu},\boldsymbol{\Sigma}) = \underbrace{-\frac{1}{|\mathcal{Q}|}\sum_{i\in\mathcal{Q}} \mathbf{z}_i^\top \log(\mathbf{p}_i)}_{\text{GMM clustering}} \underbrace{- \sum_{i\in\mathcal{D}}\sum_{j\in\mathcal{D}} w_{ij}\mathbf{z}_i\mathbf{z}_j}_{\text{Laplacian reg.}} + \underbrace{\sum_{i\in\mathcal{Q}} \mathrm{KL}_\lambda(\mathbf{z}_i||\hat{\mathbf{y}}_i)}_{\text{Text knowledge}}$$



New predictions should remain **as close as possible to the initial (zero-shot)** ones (while minimizing the 2 other terms!)

Our algorithm (in short)

Our algorithm (in short)

Solving procedure of 3 update equations

$\longrightarrow$ just a **few lines of code**!

function Histo-TransCLIP(f, t, τ)

// Text-based pseudo-labels $\hat{\mathbf{y}}$

$\hat{\mathbf{y}}_i = \text{softmax}(\tau \mathbf{f}_i^T \mathbf{t}) \quad \forall i$

// Initialize z, μ, Σ

$\mathbf{z}_i = \hat{\mathbf{y}}_i \quad \forall i$

$\boldsymbol{\mu}_k = \text{top\_confident\_average}(\mathbf{f}, \hat{\mathbf{y}}) \quad \forall k$

$\text{diag}(\boldsymbol{\Sigma}) = \frac{1}{n\_features}$

// Iterative procedure

while not_converged **do**

for l = 1:... **do**

$$\mathbf{z}_i^{(l+1)} = \frac{\hat{\mathbf{y}}_i \odot \exp(\log(\mathbf{p}_i) + \sum_{j \in \mathcal{Q}} w_{ij} \mathbf{z}_j^{(l)})}{(\hat{\mathbf{y}}_i \odot \exp(\log(\mathbf{p}_i) + \sum_{j \in \mathcal{Q}} w_{ij} \mathbf{z}_j^{(l)}))^\top \mathbb{1}_K} \quad \forall i$$

$$\boldsymbol{\mu}_k = \frac{\sum_{i \in \mathcal{Q}} z_{i,k} \mathbf{f}_i}{\sum_{i \in \mathcal{Q}} z_{i,k}} \quad \forall k$$

$$\text{diag}(\boldsymbol{\Sigma}) = \frac{1}{|\mathcal{Q}|} \sum_{i \in \mathcal{Q}} \sum_k z_{i,k} (\mathbf{f}_i - \boldsymbol{\mu}_k)^2$$

return $\mathbf{z}$

Our algorithm (in short)

Solving procedure of 3 update equations
$\longrightarrow$ just a **few lines of code**!

Really fast to solve
$\longrightarrow$ just a **few seconds for 100,000 patches**

function Histo-TransCLIP(f, t, τ)

// Text-based pseudo-labels $\hat{\mathbf{y}}$

$\hat{\mathbf{y}}_i = \text{softmax}(\tau \mathbf{f}_i^T \mathbf{t}) \quad \forall i$

// Initialize z, μ, Σ

$\mathbf{z}_i = \hat{\mathbf{y}}_i \quad \forall i$

$\boldsymbol{\mu}_k = \text{top\_confident\_average}(\mathbf{f}, \hat{\mathbf{y}}) \quad \forall k$

$\text{diag}(\boldsymbol{\Sigma}) = \frac{1}{n\_features}$

// Iterative procedure

while not_converged **do**

for l = 1:... **do**

$\mathbf{z}_i^{(l+1)} = \frac{\hat{\mathbf{y}}_i \odot \exp(\log(\mathbf{p}_i) + \sum_{j \in \mathcal{Q}} w_{ij} \mathbf{z}_j^{(l)})}{(\hat{\mathbf{y}}_i \odot \exp(\log(\mathbf{p}_i) + \sum_{j \in \mathcal{Q}} w_{ij} \mathbf{z}_j^{(l)}))^\top \mathbb{1}_K} \quad \forall i$

$\boldsymbol{\mu}_k = \frac{\sum_{i \in \mathcal{Q}} z_{i,k} \mathbf{f}_i}{\sum_{i \in \mathcal{Q}} z_{i,k}} \quad \forall k$

$\text{diag}(\boldsymbol{\Sigma}) = \frac{1}{|\mathcal{Q}|} \sum_{i \in \mathcal{Q}} \sum_k z_{i,k} (\mathbf{f}_i - \boldsymbol{\mu}_k)^2$

return z

Our algorithm (in short)

Solving procedure of 3 update equations
$\longrightarrow$ just a **few lines of code**!

Really fast to solve
$\longrightarrow$ just a **few seconds for 100,000 patches**

#Patches	Features	Histo-TransCLIP
10^2	$\sim$ 1 sec.	$\sim$ 0.1 sec.
10^3	$\sim$ 4 sec.	$\sim$ 0.2 sec.
10^4	$\sim$ 28 sec.	$\sim$ 0.4 sec.
10^5	$\sim$ 5 min.	$\sim$ 6 sec.

function Histo-TransCLIP(f, t, τ)

// Text-based pseudo-labels $\hat{\mathbf{y}}$

$\hat{\mathbf{y}}_i = \mathrm{softmax}(\tau \mathbf{f}_i^T \mathbf{t}) \quad \forall i$

// Initialize z, μ, Σ

$\mathbf{z}_i = \hat{\mathbf{y}}_i \quad \forall i$

$\boldsymbol{\mu}_k = \mathrm{top\_confident\_average}(\mathbf{f}, \hat{\mathbf{y}}) \quad \forall k$

$\mathrm{diag}(\boldsymbol{\Sigma}) = \frac{1}{n\_features}$

// Iterative procedure

while not_converged **do**

for l = 1:... **do**

$$\mathbf{z}_i^{(l+1)} = \frac{\hat{\mathbf{y}}_i \odot \exp(\log(\mathbf{p}_i) + \sum_{j \in \mathcal{Q}} w_{ij} \mathbf{z}_j^{(l)})}{(\hat{\mathbf{y}}_i \odot \exp(\log(\mathbf{p}_i) + \sum_{j \in \mathcal{Q}} w_{ij} \mathbf{z}_j^{(l)}))^\top \mathbb{1}_K} \quad \forall i$$

$$\boldsymbol{\mu}_k = \frac{\sum_{i \in \mathcal{Q}} z_{i,k} \mathbf{f}_i}{\sum_{i \in \mathcal{Q}} z_{i,k}} \quad \forall k$$

$$\mathrm{diag}(\boldsymbol{\Sigma}) = \frac{1}{|\mathcal{Q}|} \sum_{i \in \mathcal{Q}} \sum_k z_{i,k} (\mathbf{f}_i - \boldsymbol{\mu}_k)^2$$

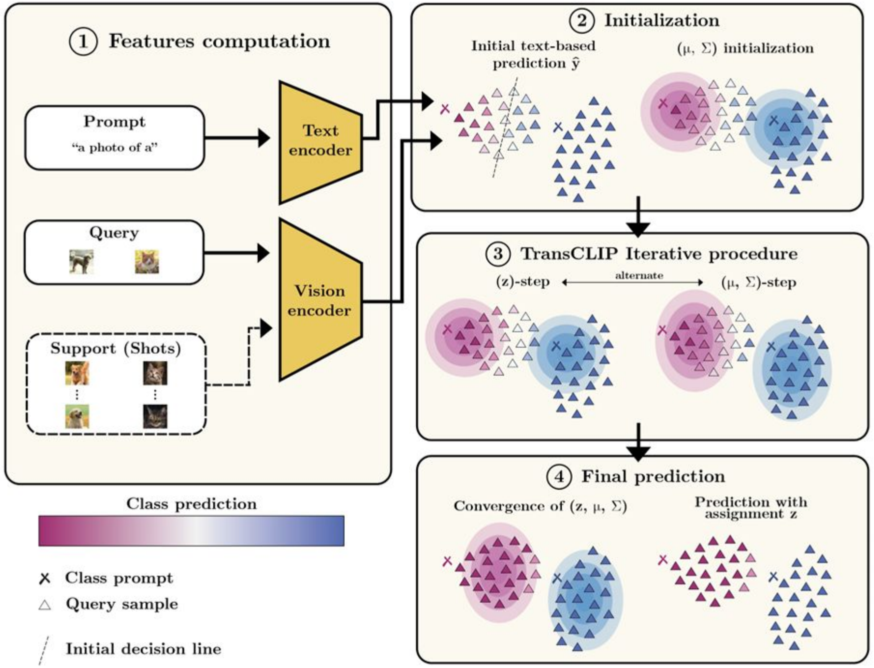
return z

Our algorithm (in short)

Solving procedure of 3 update equations
$\longrightarrow$ just a **few lines of code**!

Really fast to solve
$\longrightarrow$ just a **few seconds for 100,000 patche**

#Patches	Features	Histo-TransCLIP
10^2	$\sim$ 1 sec.	$\sim$ 0.1 sec.
10^3	$\sim$ 4 sec.	$\sim$ 0.2 sec.
10^4	$\sim$ 28 sec.	$\sim$ 0.4 sec.
10^5	$\sim$ 5 min.	$\sim$ 6 sec.

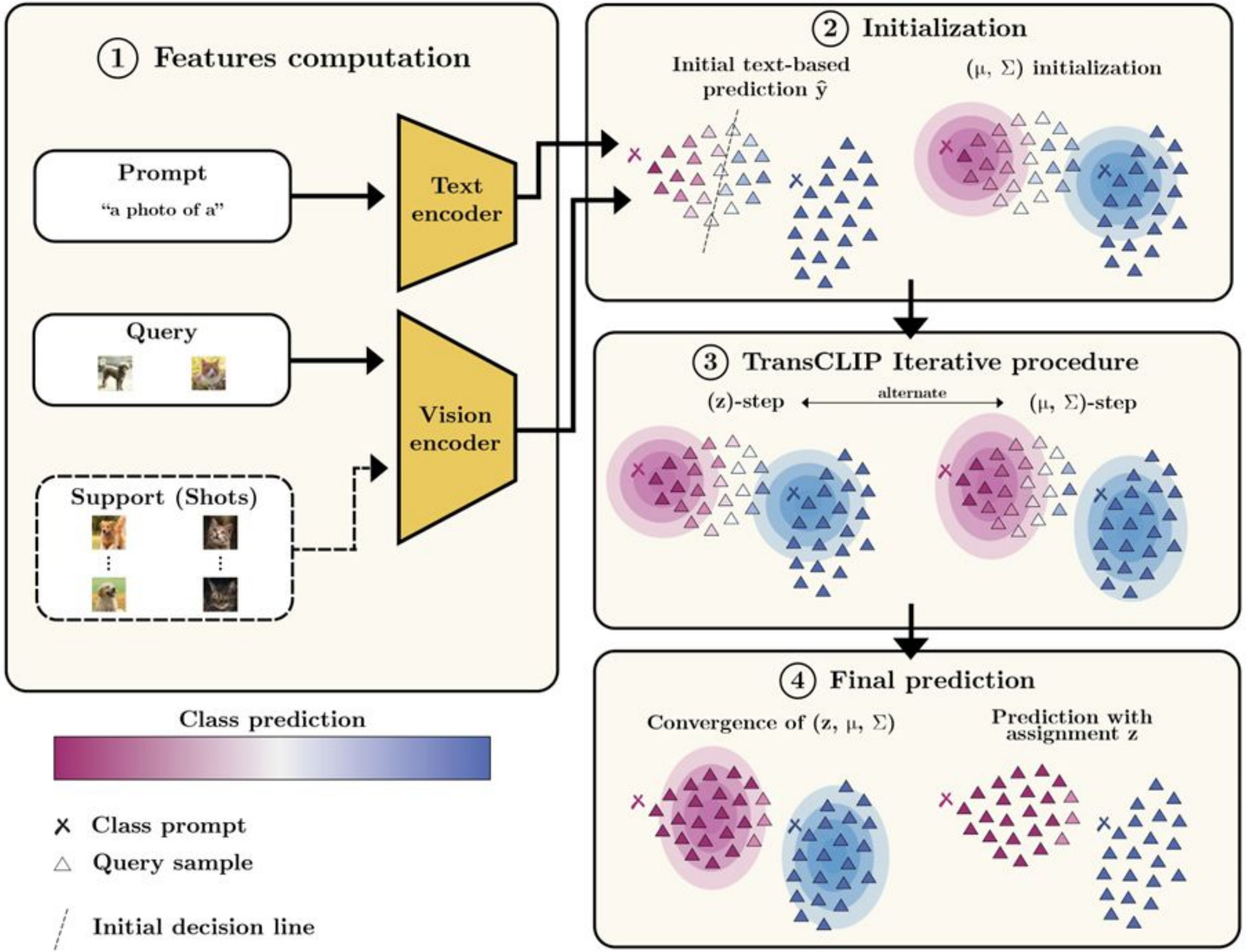


Our algori

Solving pro
→ just a fe

Really fast t
→ just a fe

#I



Results

Dataset	Method	CLIP	Quilt-B16	Quilt-B32	PLIP	CONCH
SICAP-MIL	Zero-shot	**29.85**	40.44	**35.04**	46.84	27.71
	Histo-TransCLIP	24.72	**58.49**	28.18	**53.23**	**32.58**
LC(Lung)	Zero-shot	**31.46**	43.00	76.24	84.96	84.81
	Histo-TransCLIP	25.62	**50.53**	**93.93**	**93.80**	**96.29**
SKINCANCER	Zero-shot	4.20	15.38	39.71	22.90	58.53
	Histo-TransCLIP	**11.46**	**33.33**	**48.80**	**36.72**	**66.22**
NCT-CRC	Zero-shot	25.39	29.61	53.73	63.17	66.27
	Histo-TransCLIP	**39.61**	**48.40**	**58.13**	**77.53**	**70.36**
Average	Zero-shot	22.73	32.1	51.18	54.47	59.33
	Histo-TransCLIP	**25.35**	**47.69**	**57.26**	**65.32**	**66.36**
	$\Delta_{\text{transductive}}$	**+2.62**	**+15.59**	**+6.08**	**+10.85**	**+7.03**

MICCAI2024 Marrakesh MOROCCO

MICCAI

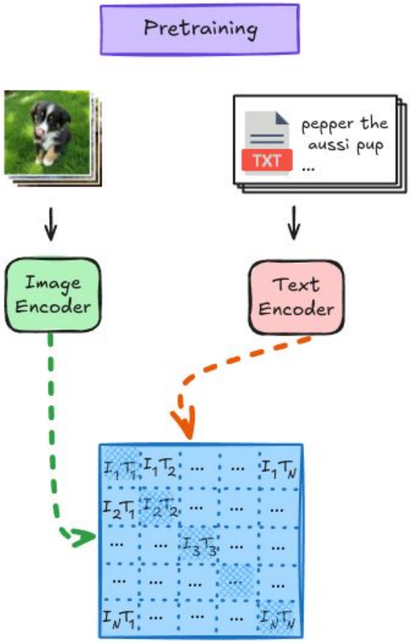
Results: generic pre-training

Dataset	Method	Model				
		CLIP	Quilt-B16	Quilt-B32	PLIP	CONCH
SICAP-MIL	Zero-shot	**29.85**	40.44	**35.04**	46.84	27.71
	Histo-TransCLIP	24.72	**58.49**	28.18	**53.23**	**32.58**
LC(Lung)	Zero-shot	**31.46**	43.00	76.24	84.96	84.81
	Histo-TransCLIP	25.62	**50.53**	**93.93**	**93.80**	**96.29**
SKINCANCER	Zero-shot	4.20	15.38	39.71	22.90	58.53
	Histo-TransCLIP	**11.46**	**33.33**	**48.80**	**36.72**	**66.22**
NCT-CRC	Zero-shot	25.39	29.61	53.73	63.17	66.27
	Histo-TransCLIP	**39.61**	**48.40**	**58.13**	**77.53**	**70.36**
Average	Zero-shot	22.73	32.1	51.18	54.47	59.33
	Histo-TransCLIP	**25.35**	**47.69**	**57.26**	**65.32**	**66.36**
	$\Delta_{\text{transductive}}$	**+2.62**	**+15.59**	**+6.08**	**+10.85**	**+7.03**

Results: histopathology pre-training

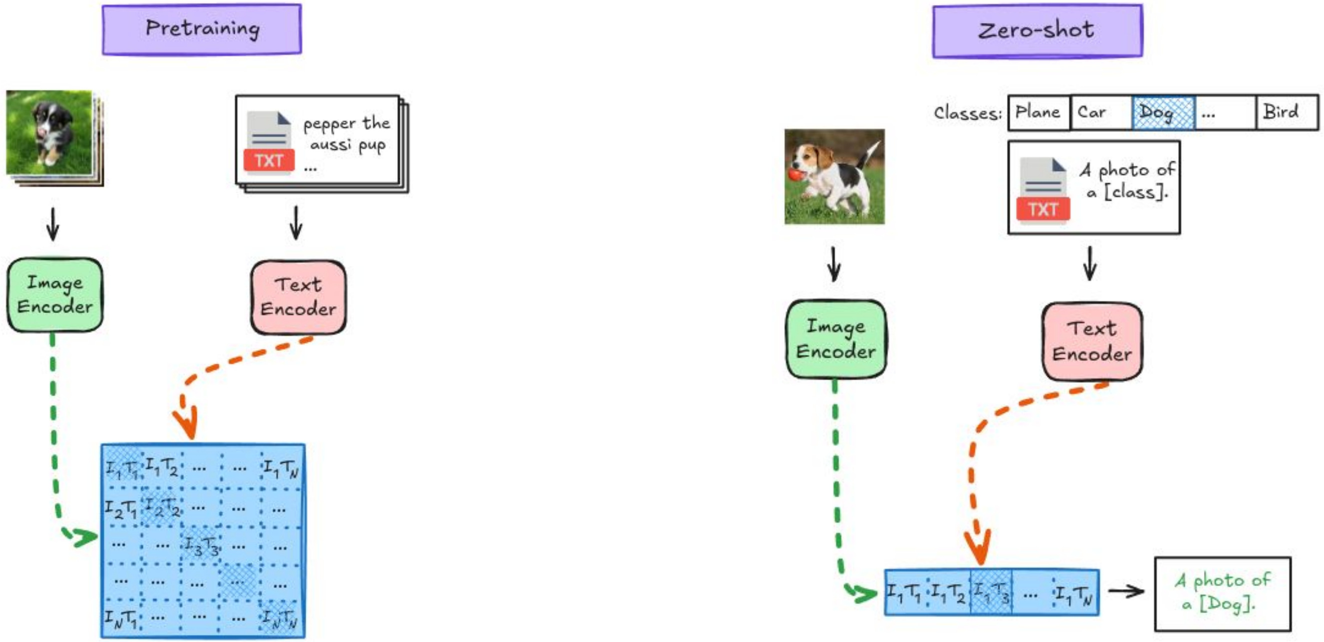
Dataset	Method	Model				
		CLIP	Quilt-B16	Quilt-B32	PLIP	CONCH
SICAP-MIL	Zero-shot	**29.85**	40.44	**35.04**	46.84	27.71
	Histo-TransCLIP	24.72	**58.49**	28.18	**53.23**	**32.58**
LC(Lung)	Zero-shot	**31.46**	43.00	76.24	84.96	84.81
	Histo-TransCLIP	25.62	**50.53**	**93.93**	**93.80**	**96.29**
SKINCANCER	Zero-shot	4.20	15.38	39.71	22.90	58.53
	Histo-TransCLIP	**11.46**	**33.33**	**48.80**	**36.72**	**66.22**
NCT-CRC	Zero-shot	25.39	29.61	53.73	63.17	66.27
	Histo-TransCLIP	**39.61**	**48.40**	**58.13**	**77.53**	**70.36**
Average	Zero-shot	22.73	32.1	51.18	54.47	59.33
	Histo-TransCLIP	**25.35**	**47.69**	**57.26**	**65.32**	**66.36**
	$\Delta_{\text{transductive}}$	**+2.62**	**+15.59**	**+6.08**	**+10.85**	**+7.03**

Generalist Vision-language Models (CLIP)



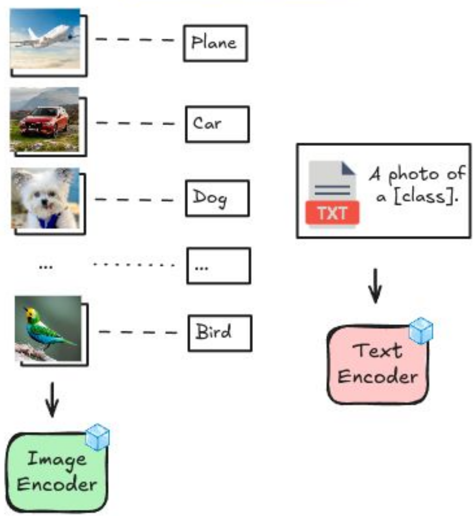
Radford, Alec, et al. "Learning transferable visual models from natural language supervision." PMLR, 2021.

Generalist Vision-language Models (CLIP)



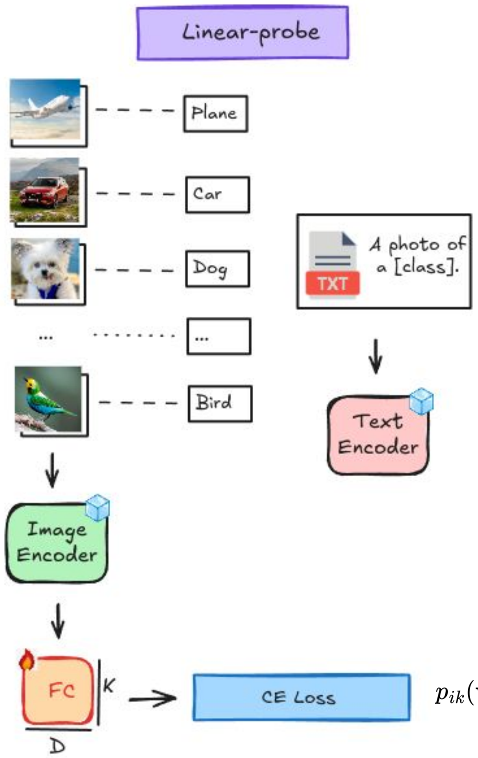
Radford, Alec, et al. "Learning transferable visual models from natural language supervision." PMLR, 2021.

Few-Shot Adaptation (linear probe)



Radford, Alec, et al. "Learning transferable visual models from natural language supervision." PMLR, 2021.

Few-Shot Adaptation (linear probe)

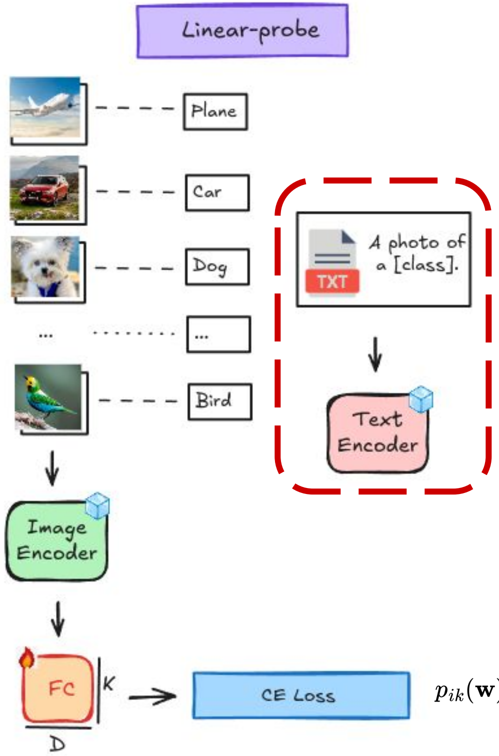


Features Learned FC weights

$$p_{ik}(\mathbf{w}) = \frac{\exp\left(\mathbf{f}_i^t \mathbf{w}_k\right)}{\sum_{j=1}^{K} \exp\left(\mathbf{f}_i^t \mathbf{w}_j\right)}$$

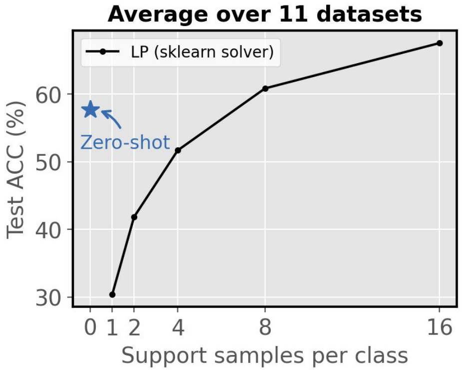
Radford, Alec, et al. "Learning transferable visual models from natural language supervision." PMLR, 2021.

Few-Shot Adaptation (linear probe)



Doesn't use text information during adaptation.

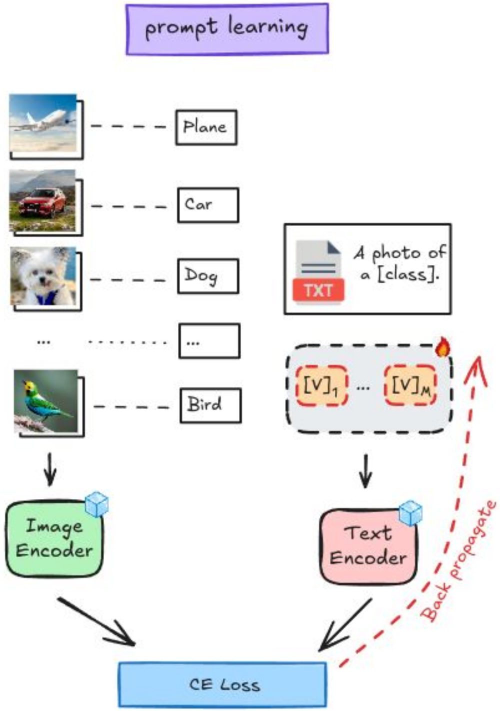
$$p_{ik}(\mathbf{w}) = \frac{\exp\left(\mathbf{f}_i^t \mathbf{w}_k\right)}{\sum_{j=1}^{K} \exp\left(\mathbf{f}_i^t \mathbf{w}_j\right)}$$



Linear Probe below Zero shot!!

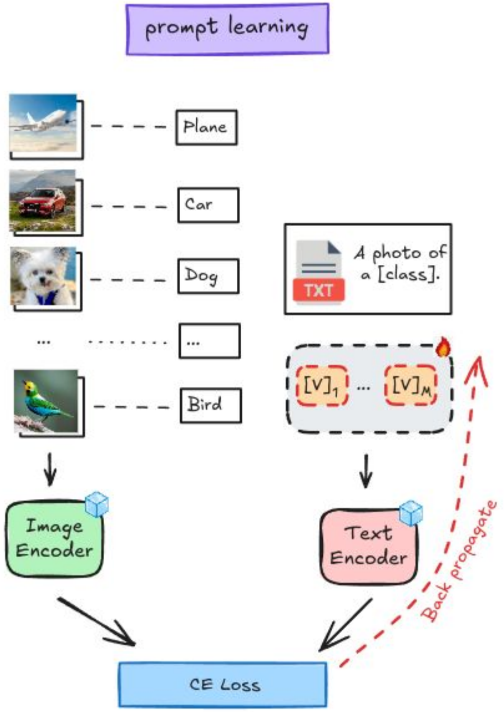
Radford, Alec, et al. "Learning transferable visual models from natural language supervision." PMLR, 2021.

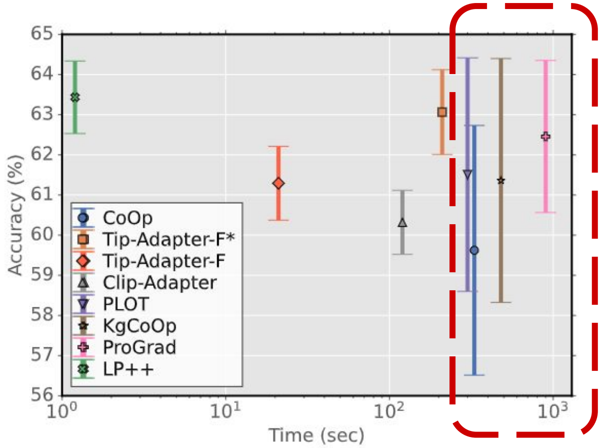
Few-Shot Adaptation (prompt learning)



Prompt learning: Zhou, Kaiyang, et al. "Learning to prompt for vision-language models." *IJCV* (2022).

Few-Shot Adaptation (prompt learning)



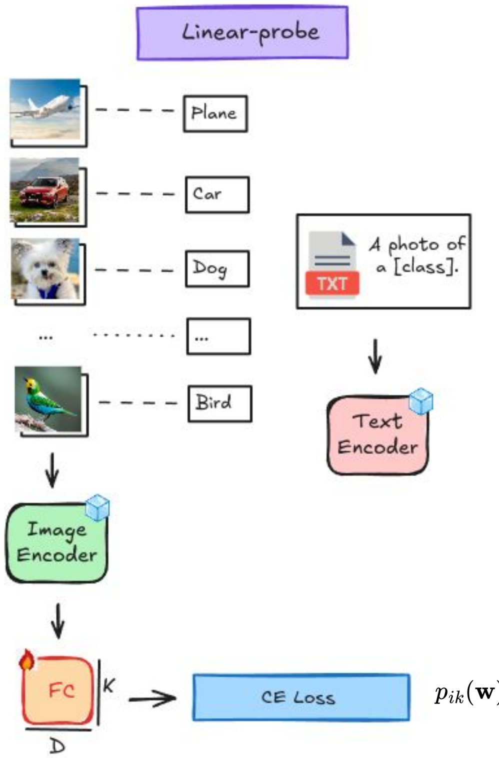


Heavy computational & memory load

Not valid in black-box scénario

Prompt learning: Zhou, Kaiyang, et al. "Learning to prompt for vision-language models." *IJCV* (2022).
Figure from: Huang, Y, et al. . “LP++: A Surprisingly Strong Linear Probe for Few-Shot CLIP” . CVPR (2024).

Few-Shot Adaptation (linear probe)

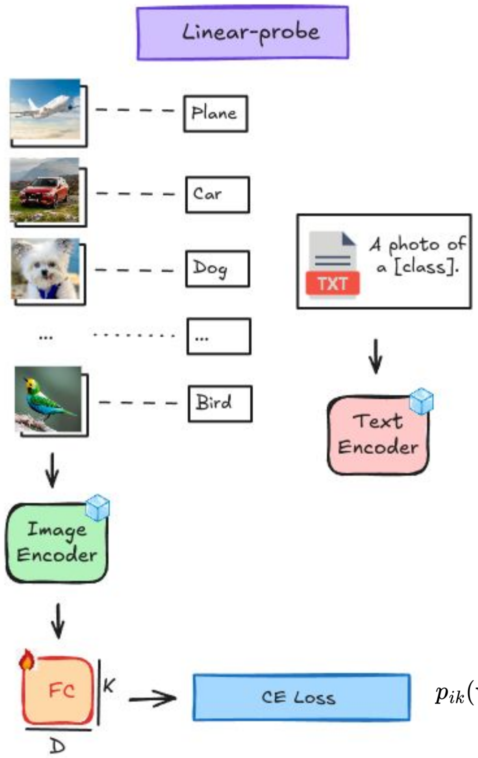


Features   Learned FC weights

$$p_{ik}(\mathbf{w}) = \frac{\exp\left(\mathbf{f}_i^t \mathbf{w}_k\right)}{\sum_{j=1}^{K} \exp\left(\mathbf{f}_i^t \mathbf{w}_j\right)}$$

Radford, Alec, et al. "Learning transferable visual models from natural language supervision." PMLR, 2021.

Few-Shot Adaptation (linear probe)



Features

Learned FC weights

$$p_{ik}(\mathbf{w}) = \frac{\exp\left(\mathbf{f}_i^t \mathbf{w}_k\right)}{\sum_{j=1}^{K} \exp\left(\mathbf{f}_i^t \mathbf{w}_j\right)}$$

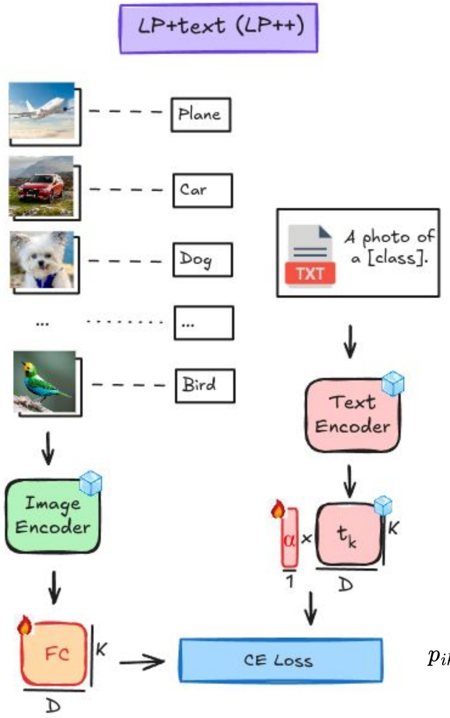
LP+text (LP++)

learnable image-text blending parameters

Text Prototype

$$p_{ik}(\mathbf{w}, \alpha) = \frac{\exp\left(\mathbf{f}_i^t(\mathbf{w}_k + \alpha_k \mathbf{t}_k)\right)}{\sum_{j=1}^{K} \exp\left(\mathbf{f}_i^t(\mathbf{w}_j + \alpha_j \mathbf{t}_j)\right)}$$

Few-Shot Adaptation (LP++)



W: visual class prototypes

t_k: text class prototypes

α: blending parameter

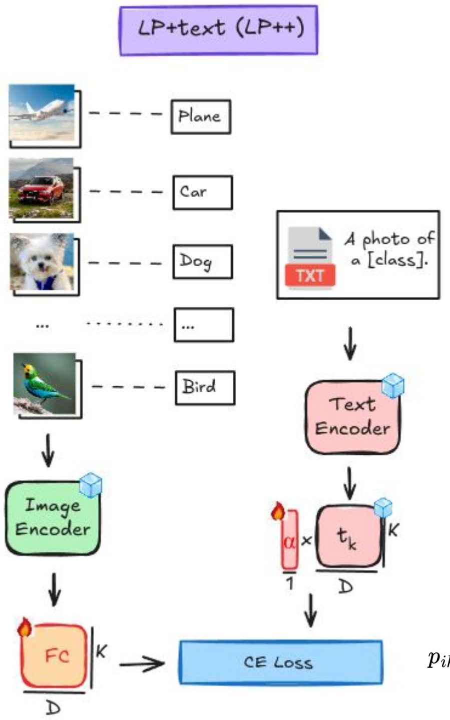
K: number of classes

D: featuremap dimension

$$p_{ik}(\mathbf{w}, \alpha) = \frac{\exp\left(\mathbf{f}_i^t(\mathbf{w}_k + \alpha_k \mathbf{t}_k)\right)}{\sum_{j=1}^{K} \exp\left(\mathbf{f}_i^t(\mathbf{w}_j + \alpha_j \mathbf{t}_j)\right)}$$

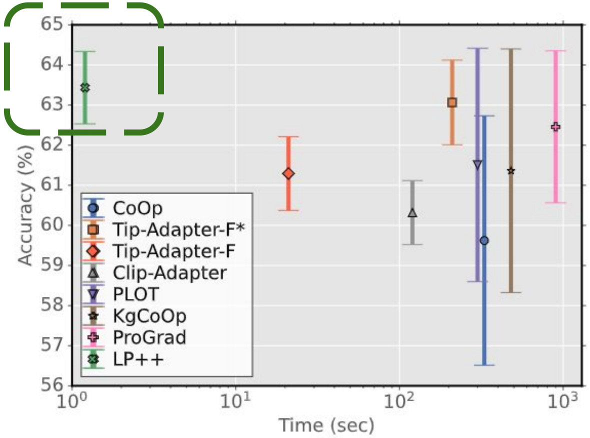
Huang, Y, et al. . "LP++: A Surprisingly Strong Linear Probe for Few-Shot CLIP" .CVPR 2024.

Few-Shot Adaptation (LP++)



Orders-of-magnitude faster

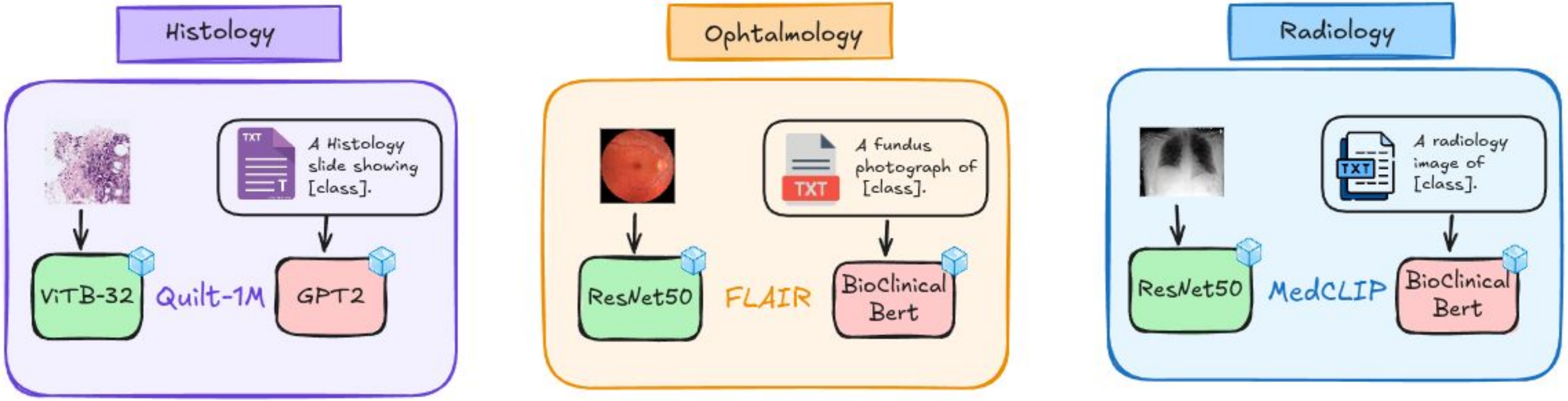
Préserve black-box



$$p_{ik}(\mathbf{w}, \alpha) = \frac{\exp\left(\mathbf{f}_i^t(\mathbf{w}_k + \alpha_k \mathbf{t}_k)\right)}{\sum_{j=1}^{K} \exp\left(\mathbf{f}_i^t(\mathbf{w}_j + \alpha_j \mathbf{t}_j)\right)}$$

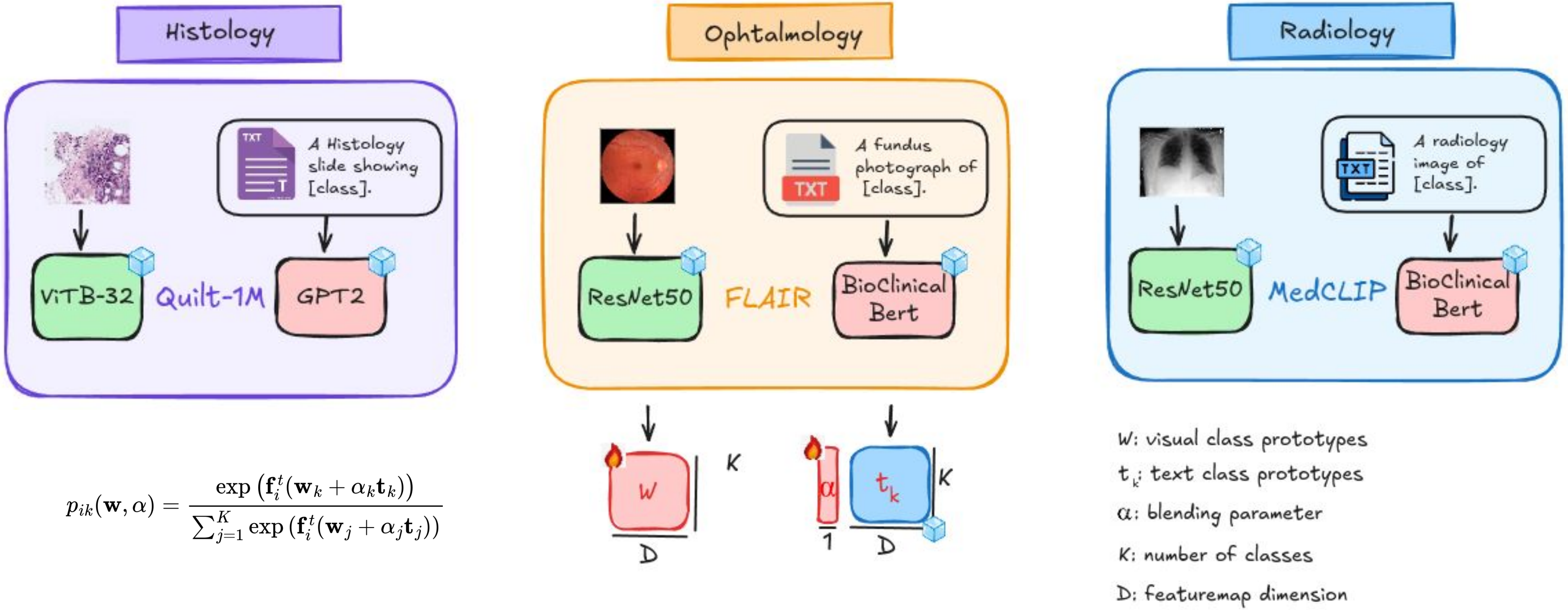
Huang, Y, et al. . "LP++: A Surprisingly Strong Linear Probe for Few-Shot CLIP" .CVPR 2024.

Medical VLMs



FLAIR: Silva-Rodríguez et al., A Foundation Language-Image Model of the Retina, MedIA 2024
Quilt-1M: Ikezogwo et al., One Million Image-Text Pairs for Histopathology, NeurIPS 2023
MedCLIP: Wang et al., Contrastive Learning from Unpaired medical images and text, EMNLP 2022

Few-Shot Adaptation in Medical VLMs

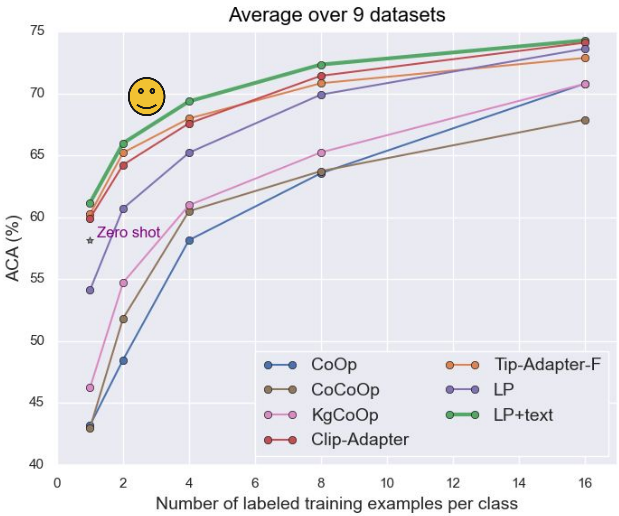


$$p_{ik}(\mathbf{w},\alpha) = \frac{\exp\left(\mathbf{f}_i^t(\mathbf{w}_k + \alpha_k \mathbf{t}_k)\right)}{\sum_{j=1}^{K} \exp\left(\mathbf{f}_i^t(\mathbf{w}_j + \alpha_j \mathbf{t}_j)\right)}$$

Shakeri, F, et al. "Few-shot adaptation of medical vision-language models." *MICCAI*, 2024.

Few-Shot Adaptation in Medical VLMs

3 modalities / 9 datasets

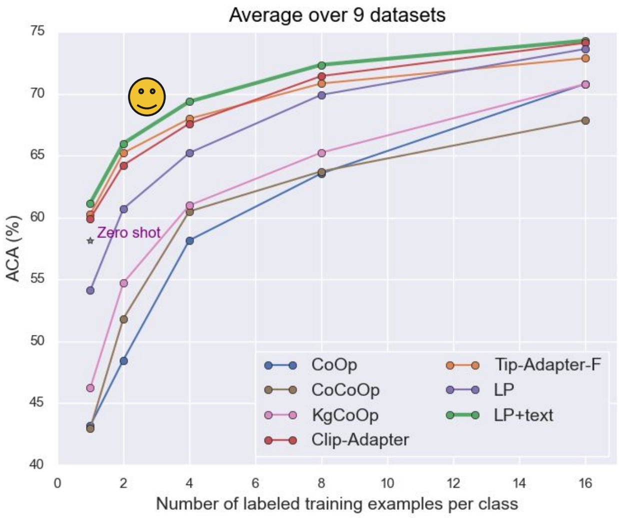


LP+text is competitive!

Shakeri, F, et al. "Few-shot adaptation of medical vision-language models." *MICCAI*, 2024.

Few-Shot Adaptation in Medical VLMs

3 modalities / 9 datasets



LP+text is <u>competitive</u>!

LP+text is <u>extremely efficient</u>!

- Adaptation in a **matter of seconds**
- Trainable on **commodity GPUs**
- **Black-box adaptation**

Methods	Category	Training Time	Black-box	#Parameters
Zero-shot [21]		n/a	✓	n/a
CoOp [35]	*Prompt-Learning*	3min	✗	$K \times n_{ctx1} \times D$
CoCoOp [34]		12min	✗	$n_{ctx2} \times D + C$
KgCoOp [31]		3min	✗	$K \times n_{ctx1} \times D$
Clip-Adapter [5]	*CLIP-based Adapters*	2min	✓	$2(D_1 \times D_2)$
Tip-adapter-F [32]		2min	✓	$K \times S \times D$
LP	*Linear probe*	43s	✓	$K \times D$
LP+text [7]		4s	✓	$K(D+1)$

Shakeri, F, et al. "Few-shot adaptation of medical vision-language models." *MICCAI*, 2024.

References

- Few-shot Adaptation of Medical Vision Language Models [Spotlight]
  F Shakeri, Y Huang, JR Silva, H Bahig, A Tang, J Dolz, IB Ayed
  IMedical Image Computing and Computer Assisted Intervention (MICCAI), 2024

- LP++: A Surprisingly Strong Linear Probe for Few-Shot CLIP
  Y Huang, F Shakeri, J Dolz, M Boudiaf, H Bahig, IB Ayed
  IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024

- Boosting Vision-Language Models for Histopathology Classification: Predict all at once [Best Paper Award]
  M Zanella, F Shakeri, Y Huang, H Bahig, IB Ayed
  MICCAI Workshop on Foundation Models for General Medical AI (MedAGI), 2024

Questions? :)

Thank you for your attention!