NEURAL INFORMATION PROCESSING SYSTEMS

EPFL

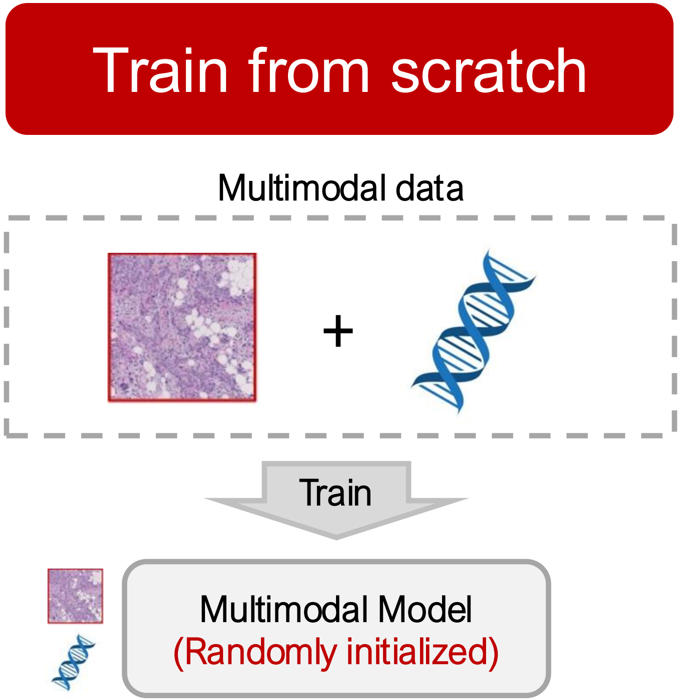
University of Basel

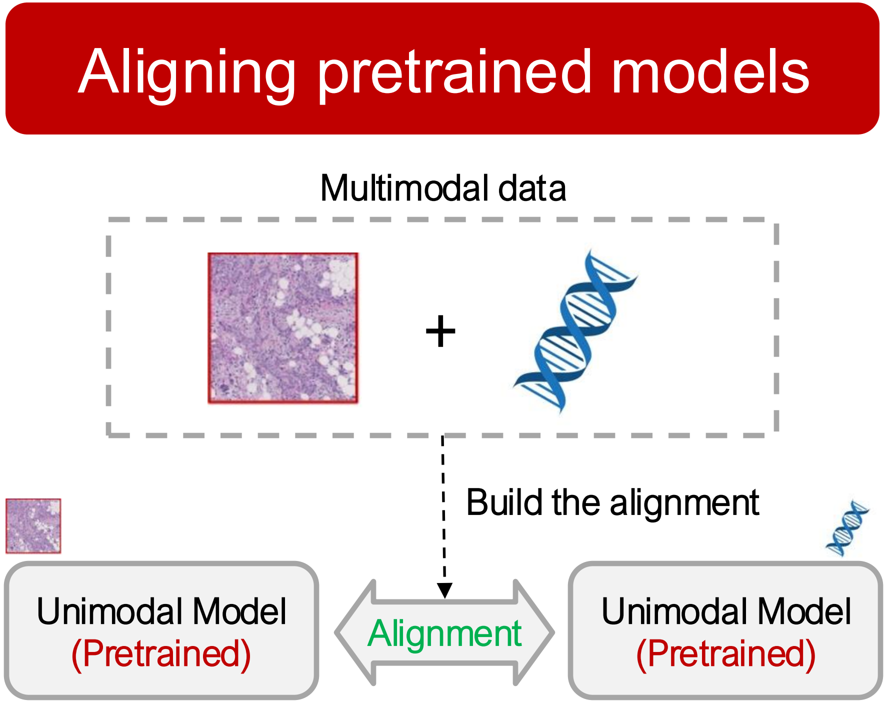
HSLU

With Limited Data for Multimodal Alignment, Let the **STRUCTURE** Guide You

Fabian Gröger*, Shuo Wen*, Huyen Le, Maria Brbić

Can We Create Multimodal Models by Aligning Pretrained Unimodal Models?



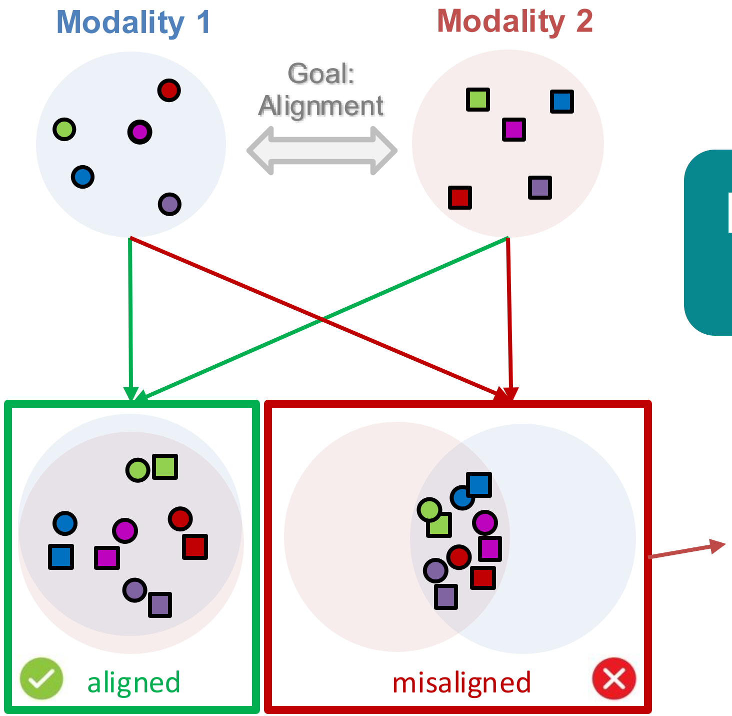


Current Alignment Still Require Large Amount of Paired Data

Millions of paired samples are **often unavailable** in many domains like healthcare and biology, where collecting high-quality multimodal data is expensive and labor-intensive.

Can We Align Pretrained Unimodal Models with Limited Data?

Can We Align Pretrained Unimodal Models with Limited Data?

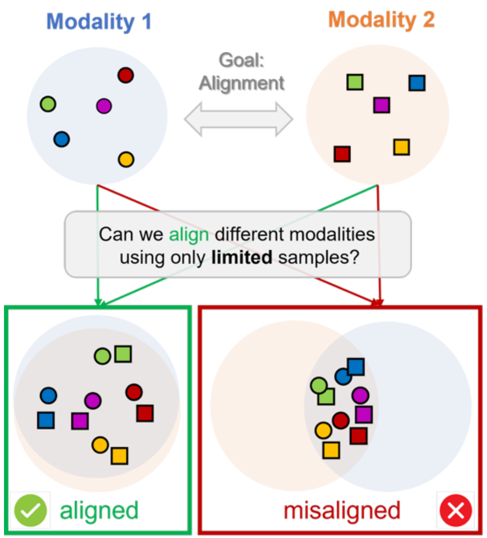


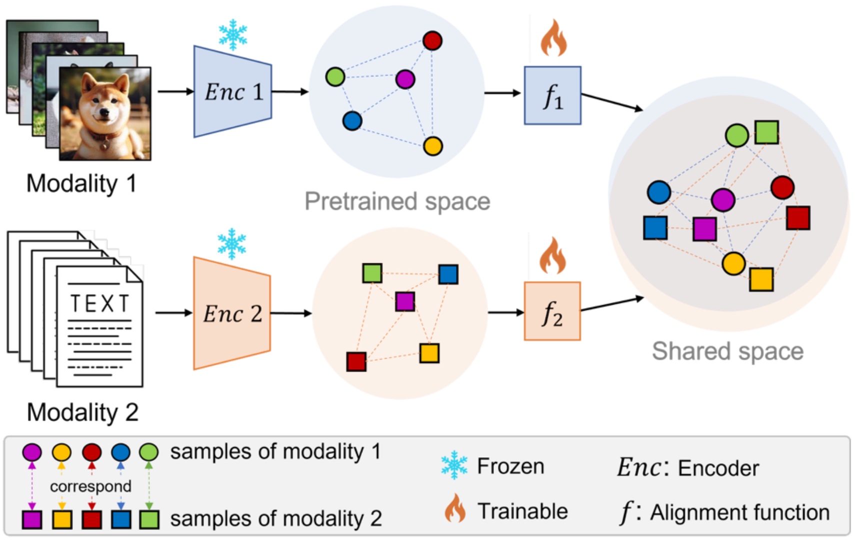
Main challenge: how to guiding the model toward a well-aligned solution?

The spaces can be misaligned due to overfitting.

Method: Overview

Key idea: Preserves the neighbourhood geometry of the latent space of the pretrained unimodal encoders.



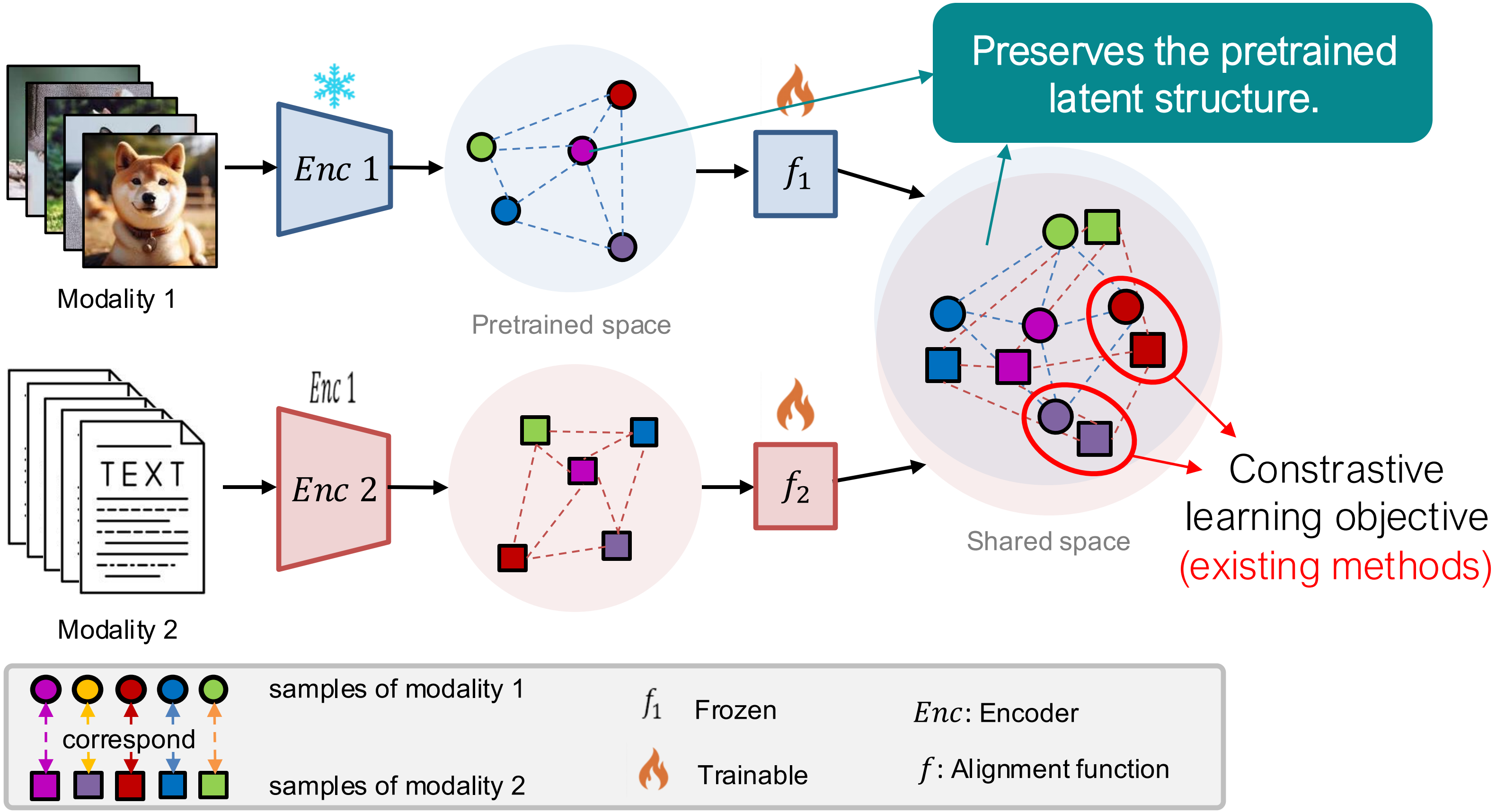


Fabian Gröger

Shuo Wen

Huyen Le

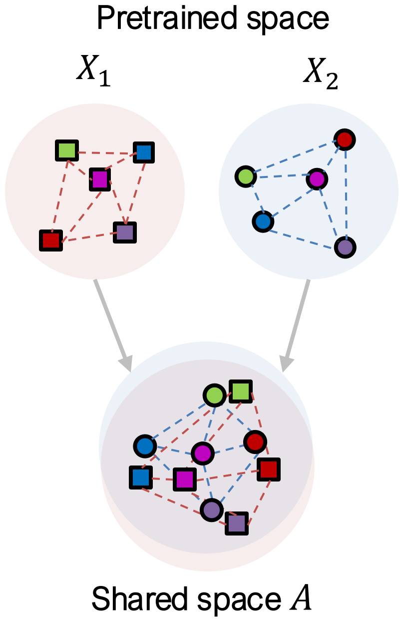
Method: STRUCTURE Regularization



Method: STRUCTURE Regularization

Key idea: Preserve the structure of pretrained space!

Data points that are similar to each other in the pretrained space should remain similar in the aligned space.



Similarity matrices

$$S_X = \frac{\tilde{X}\tilde{X}^T}{\tau}$$

(X can be either X_1 or X_2)

$$S_A = \frac{\tilde{A}\tilde{A}^T}{\tau}$$

Similarity distributions

$$P_X = softmax(S_X)$$

$$P_A = softmax(S_A)$$

Jensen-Shannon divergence

$$JS(P_X^{(l)}, P_A^{(l)})$$

Minimize to preserve the structure!

Method: STRUCTURE Regularization

Regularizer:

To capture relationships reachable by exactly l hops on the similarity graph

$$\mathcal{R}_{\mathrm{S}}^{(L)}(X, A) = \frac{1}{L} \sum_{l=1}^{L} \frac{\mathrm{JS}(P_X^{(l)}, P_A^{\boxed{(l)}})}{l}$$

Loss function:

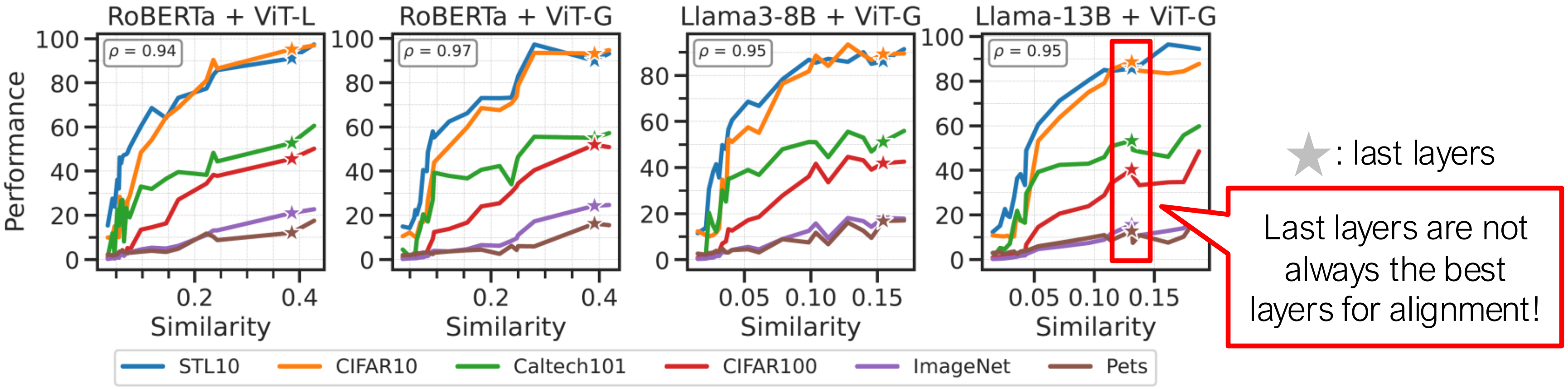
$$\mathcal{L} = \underline{\mathcal{L}_{\mathrm{A}}} + \lambda\Big(\underbrace{\mathcal{R}_{\mathrm{S}}^{(L)}(X_1, f_1(X_1))}_{\text{Reg. for Modality 1}} + \underbrace{\mathcal{R}_{\mathrm{S}}^{(L)}(X_2, f_2(X_2))}_{\text{Reg. for Modality 2}}\Big)$$

Existing alignment objective function, e.g., CLIP loss.

STRUCTURE Regularization can be easily Incorporated into existing alignment methods!

Method: Layer selection

Finding: The alignment quality (model performance) is highly correlated with the layer similarity (measured by mutual KNN)!

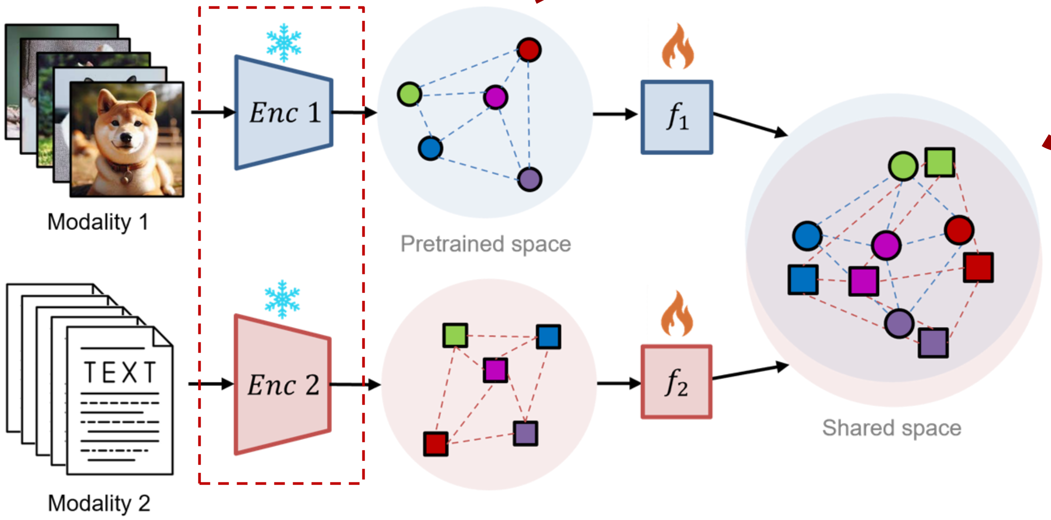


Key idea: Align most similar layers (which is not necessarily the last ones)!

Method: Summary

1. Select and align the most similar layers using mutual KNN

2. Adding STRUCTURE regularization R_S to the objective function



L_A: Alignment function from existing methods

Both components can be easily Incorporated into existing alignment methods!

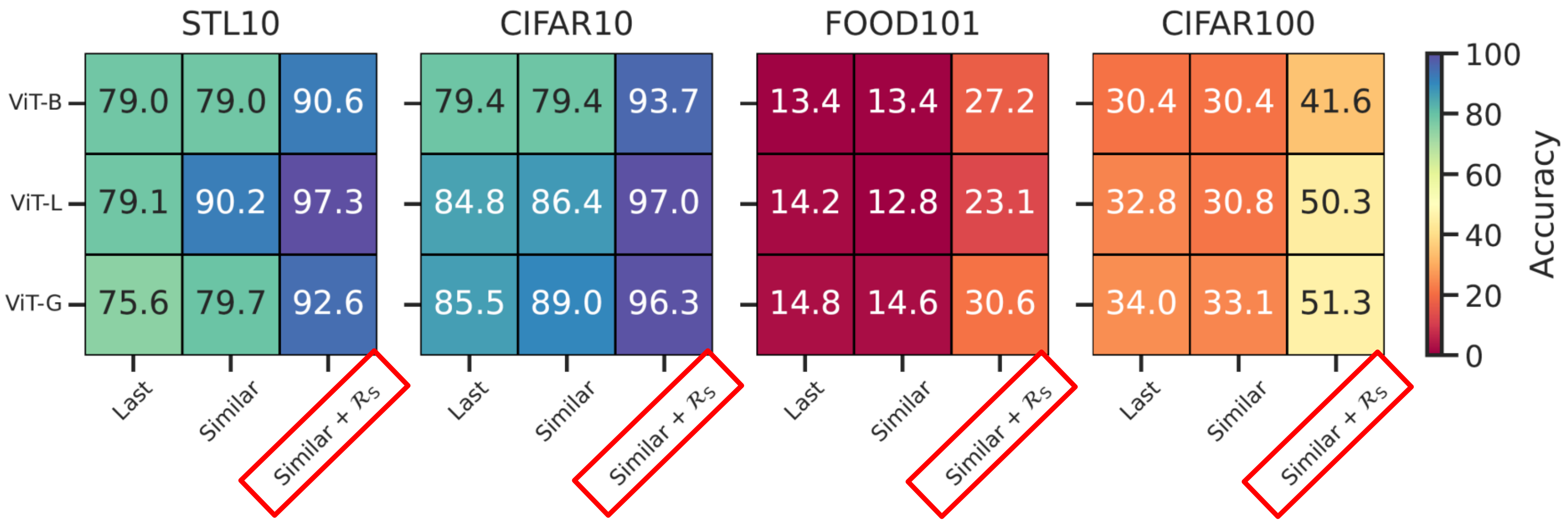
Main Results

Existing method benefit a lot from incorporating with our STRUCTURE regularization (R_s) and layer selection strategy.

		Zero-shot Classification (Accuracy)							Retrieval (R@1)	
	Method	STL10	CIFAR10	Caltech101	Food101	CIFAR100	ImageNet	Pets	Flickr I2T	Flickr T2I
	Linear + Last [10]	75.6	85.5	37.9	14.8	34.0	9.9	7.0	32.5	22.1
2.5% improv.	*Linear + Similar*	79.7	89.0	39.5	14.6	33.1	10.5	4.9	35.3	24.0
68.4% improv.	*Linear + Similar +* $\mathcal{R}_S$	92.6	96.3	56.0	**30.6**	51.3	24.7	13.2	65.8	53.7
	MLP + Last [9]	76.6	79.2	38.2	15.6	35.3	10.6	5.3	31.6	20.3
4.8% improv.	*MLP + Similar*	84.0	81.5	38.8	17.1	34.5	11.4	6.1	36.4	25.0
74.0% improv.	*MLP + Similar +* $\mathcal{R}_S$	**92.7**	96.3	56.0	30.5	52.1	25.1	13.2	**65.9**	**53.8**
	CSA + Last [24]	77.9	78.5	31.4	29.3	47.4	23.2	14.4	47.0	38.3
2.0% improv.	*CSA + Similar*	80.0	80.8	33.6	28.0	47.4	23.3	14.9	48.6	39.0
26.8% improv.	*CSA + Similar +* $\mathcal{R}_S$	91.7	**97.2**	**61.5**	28.6	**56.4**	**26.8**	**17.0**	56.1	43.1

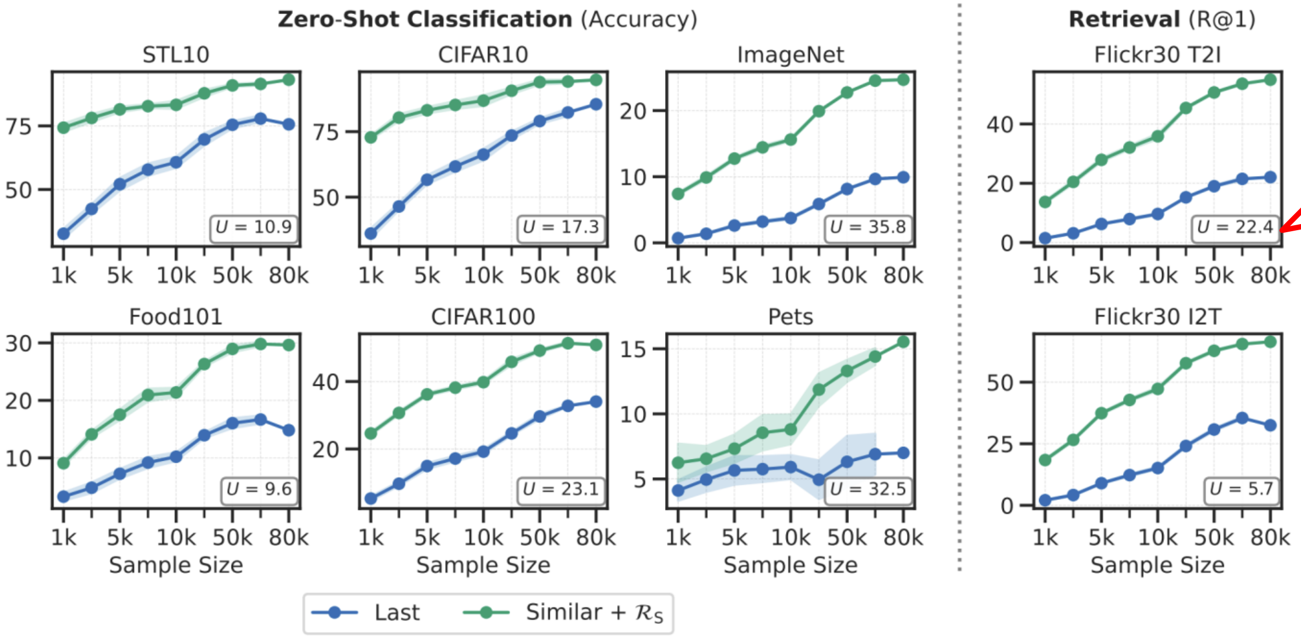
Across Different Model Combinations.

The same improvement exists across different model combinations.



Scaling Down the Training Data

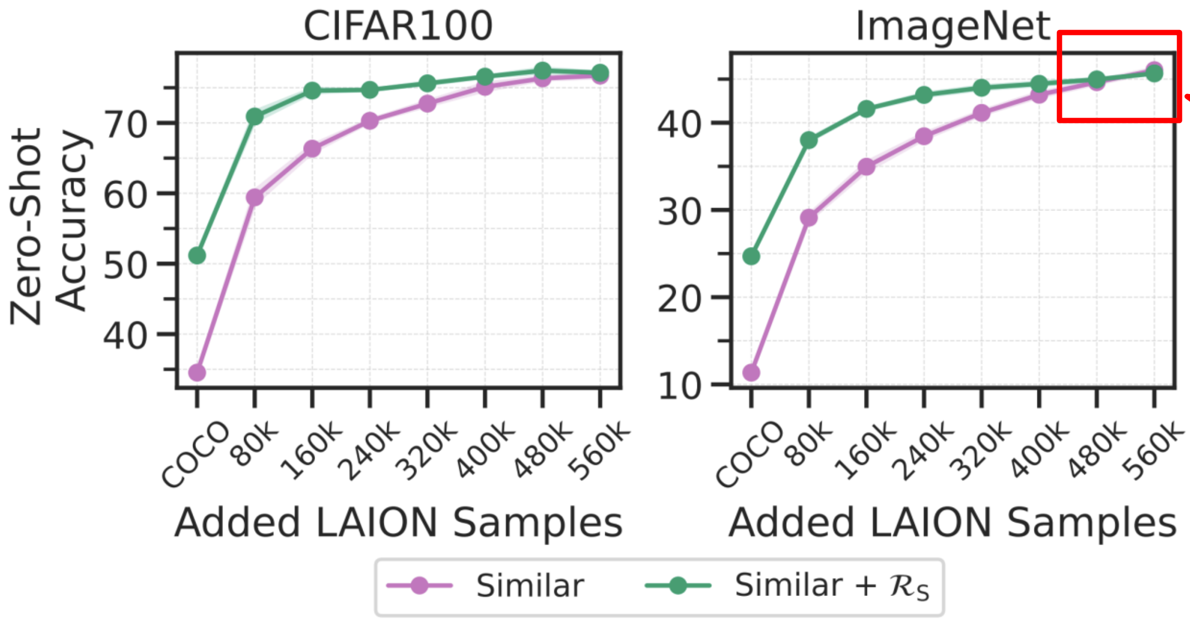
Proposed approach works well even with less data.



Requires less data to achieve the same performance.

Scaling Up the Training Data

Proposed approach brings the most benefit with limited data.



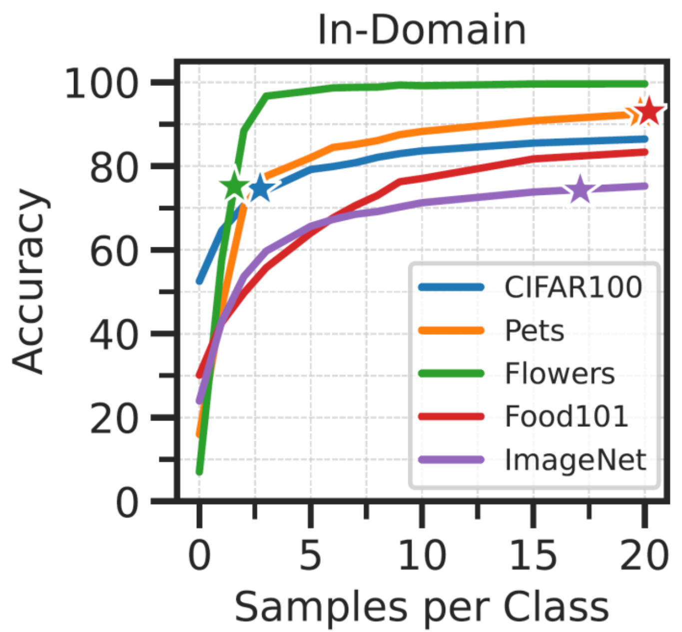
Impact of STRUCTURE regularizer diminishes as more data is available.

Train-test Data Distribution Shift

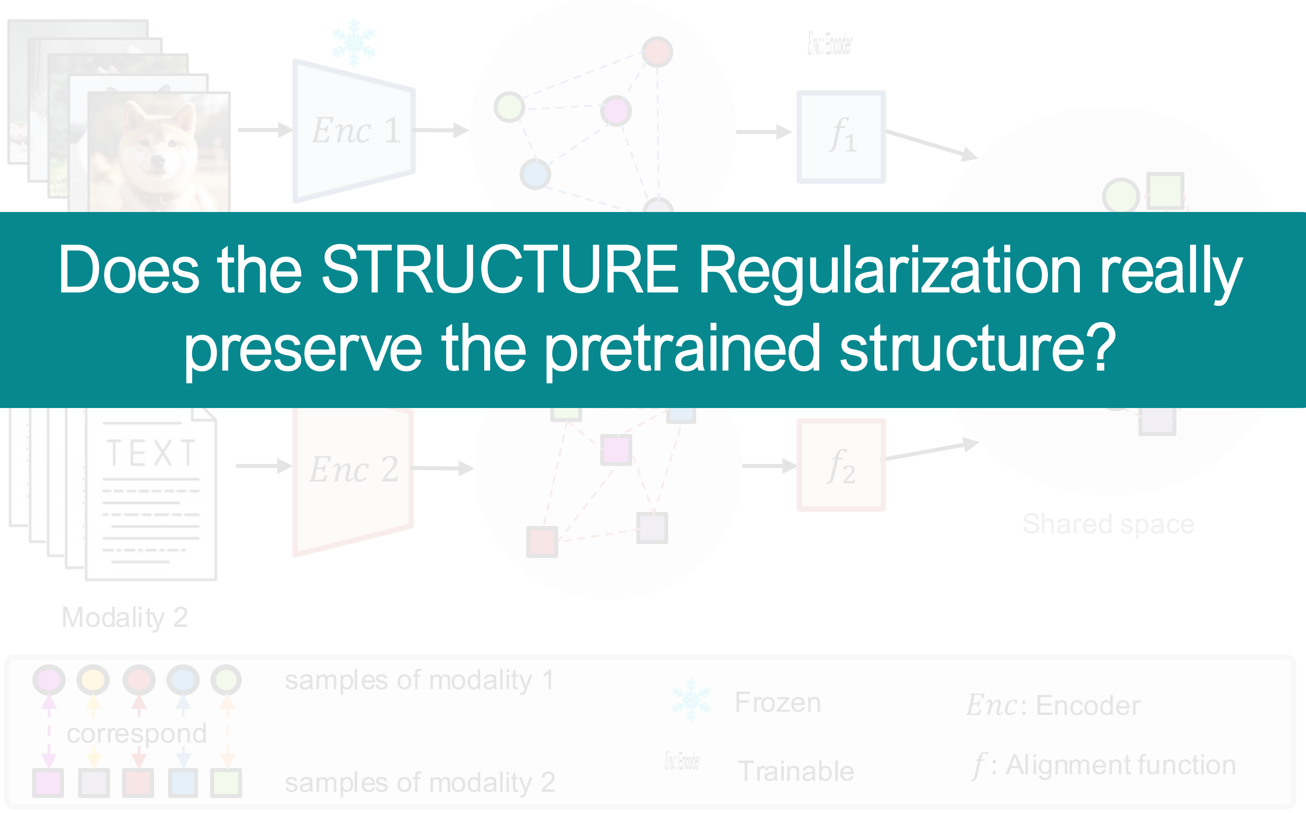
Despite the advantages of the proposed alignment approach, in low-data regimes, performance remains low on certain datasets:

Method	Food101	CIFAR100	ImageNet	Pets
Linear + Last [10]	14.8	34.0	9.9	7.0
Linear + Similar	14.6	33.1	10.5	4.9
Linear + Similar + $\mathcal{R}_S$	**30.6**	51.3	24.7	13.2

Include a small number of in-domain samples into the training set can significantly improve the performance!



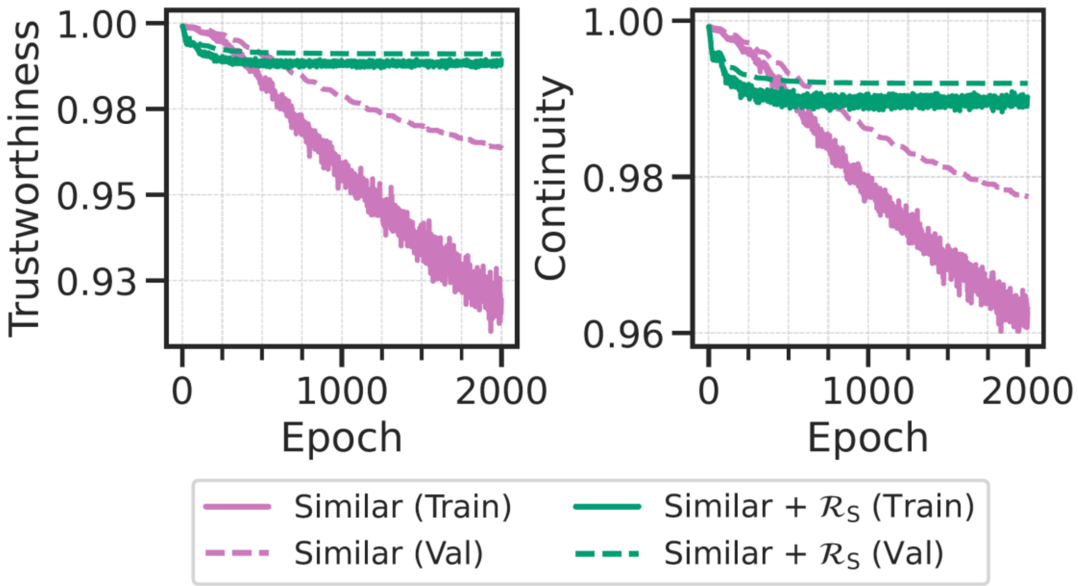
Recap: STRUCTURE Regularization



Does the STRUCTURE Regularization really preserve the pretrained structure?

Neighborhood preservation

STRUCTURE Regularization preserves the pretrained structure!



NEURAL INFORMATION PROCESSING SYSTEMS

EPFL

University of Basel

HSLU

With Limited Data for Multimodal Alignment, Let the **STRUCTURE** Guide You

Fabian Gröger*, Shuo Wen*, Huyen Le, Maria Brbić