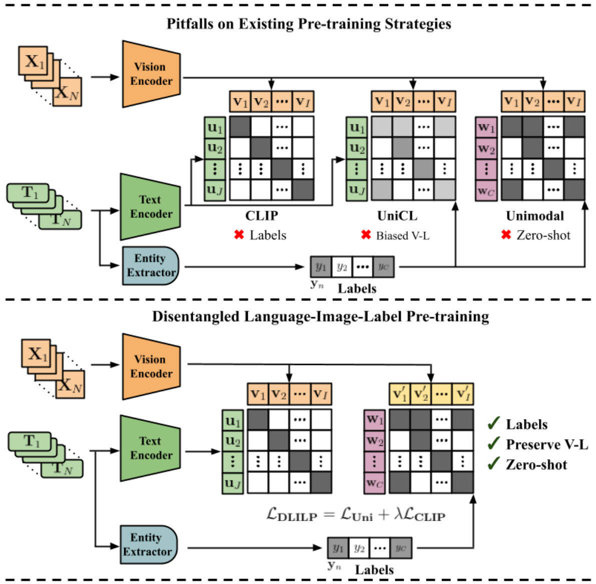
Montreal Medical Imaging Seminar
ÉTS Montreal, 1st May 2024

A Reality Check of Vision-Language Pre-training in Radiology: Have We Progressed Using Text?

Julio Silva-Rodríguez, Jose Dolz and Ismail Ben Ayed

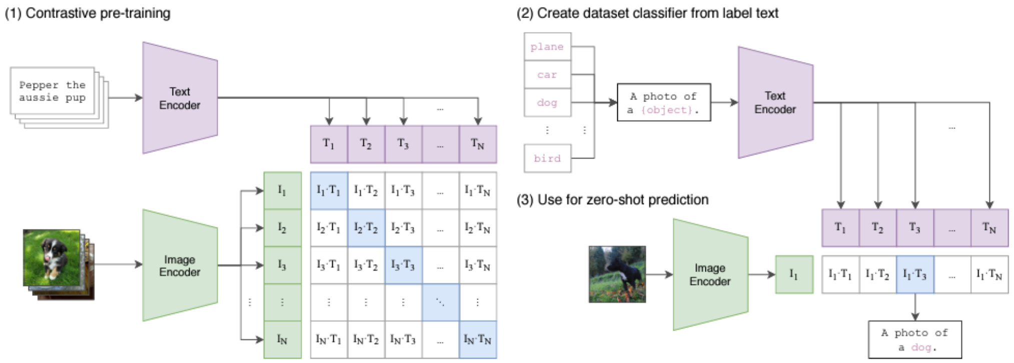
ÉTS Montréal



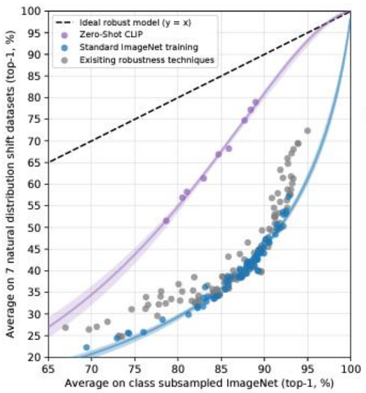
Vision-Language Foundation Models

Learning Transferable Visual Models From Natural Language Supervision

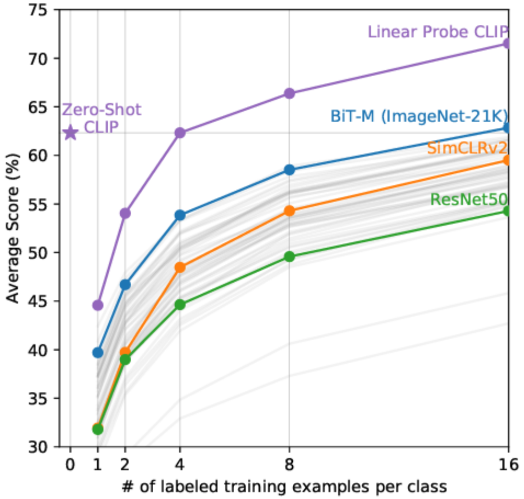
Alec Radford [*1] Jong Wook Kim [*1] Chris Hallacy [1] Aditya Ramesh [1] Gabriel Goh [1] Sandhini Agarwal [1]
Girish Sastry [1] Amanda Askell [1] Pamela Mishkin [1] Jack Clark [1] Gretchen Krueger [1] Ilya Sutskever [1]



400M image-text pairs



	ImageNet ResNet101	Zero-Shot CLIP	Δ Score
ImageNet	76.2	76.2	0%
ImageNetV2	64.3	70.1	+5.8%
ImageNet-R	37.7	88.9	+51.2%
ObjectNet	32.6	72.3	+39.7%
ImageNet Sketch	25.2	60.2	+35.0%
ImageNet-A	2.7	77.1	+74.4%



Vision-Language Pre-training in Radiology

- There is a large core of literature for **Chest X-ray (CXR)** image understanding driven by the text report, **driven by MIMIC dataset** + **labels in CheXpert and MIMIC extracted by NLP methods**.

Table 1: Frontal Chest X-ray datasets assembly. We compiled open-access datasets for training and evaluation. Green-colored categories indicate novel classes not explicitly used during CheXpert and MIMIC pre-training.

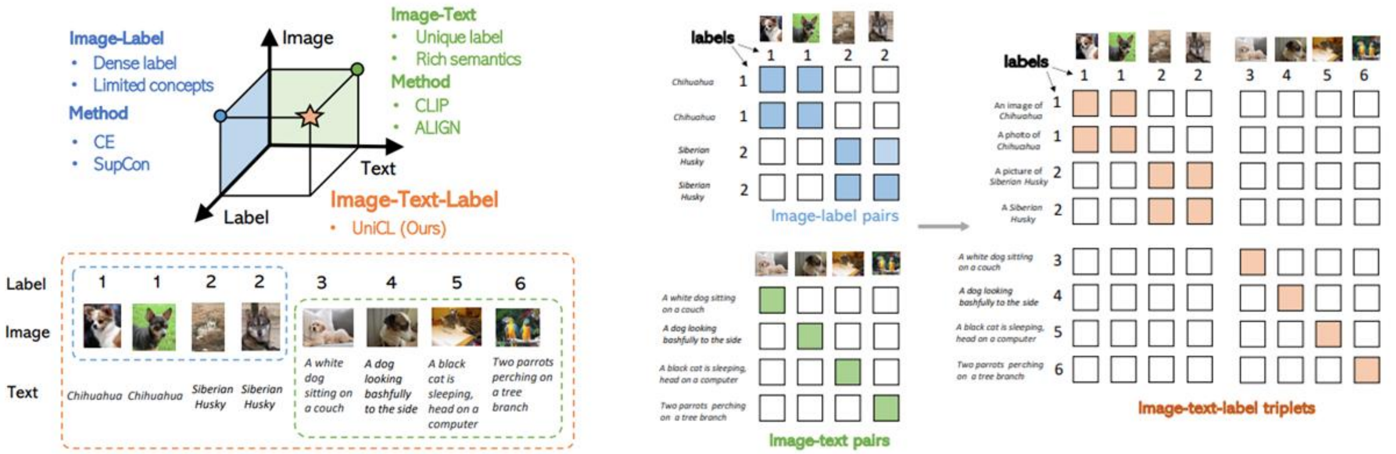
Pre-train	#Images	Reports	#Labels	Categories
CheXpert (C) [13]	191,026		14	[NoF, EnlCard, Card, LuLes, LuOp, Edema, Cons,
MIMIC (M) [16]	154,595	×	14	PnMo, Atel, PnThor, PleEffu, PleOth, Fract, Dev]
PadChest* (P) [3]	96,201		84	(see Supp. Materials)


Sentences	Text Labels
1. Hazy widespread opacity which could be compatible with a coinciding pneumonia.	1. [Lung Opacity]
2. Pulmonary nodules in the left upper lobe are also not completely characterized on this study.	2. [Lung Lesion]
1. With exception of mild bibasilar atelectasis, the lungs are normally expanded without focal opacity to suggest pneumonia.	1. [Atelectasis]
2. Heart size is mildly enlarged.	2. [Cardiomelagy]
3. There is no pleural effusion or pneumothorax	3. [No Findings]

- CONVIRT (MLHC20)
- GlorIA (ICCV21)
- MedCLIP (EMNLP 22)
- CheXZerp (NatureBioEng22)
- BioVIL (ECCV22)
- MGCA (NeurIPS22)
- MedKLIP (ICCV23)
- CXR-CLIP (MICCAI23)
- KAD (NatureCom23)
- SAT (TMI23)

Image-Text-Label Alignment

UniCL: Unified Contrastive Learning in Image-Text-Label Space (CVPR22)



“A large, black husky”

“An small, white husky”

Label: Husky

CLIP Loss

$$\mathcal{L}^{\mathrm{i2t}}_{\mathrm{CLIP}}(\theta,\phi,\tau|\mathcal{B}) = -\sum_{i\in\mathcal{B}} \log\frac{\exp(\mathbf{v}_i^T\mathbf{u}_i/\tau)}{\sum_{j\in\mathcal{B}}\exp(\mathbf{v}_i^T\mathbf{u}_j/\tau))}$$

$$\mathcal{L}^{\mathrm{t2i}}_{\mathrm{CLIP}}(\theta,\phi,\tau|\mathcal{B}) = -\sum_{j\in\mathcal{B}} \log\frac{\exp(\mathbf{v}_j^T\mathbf{u}_j/\tau)}{\sum_{i\in\mathcal{B}}\exp(\mathbf{v}_i^T\mathbf{u}_j/\tau))}$$

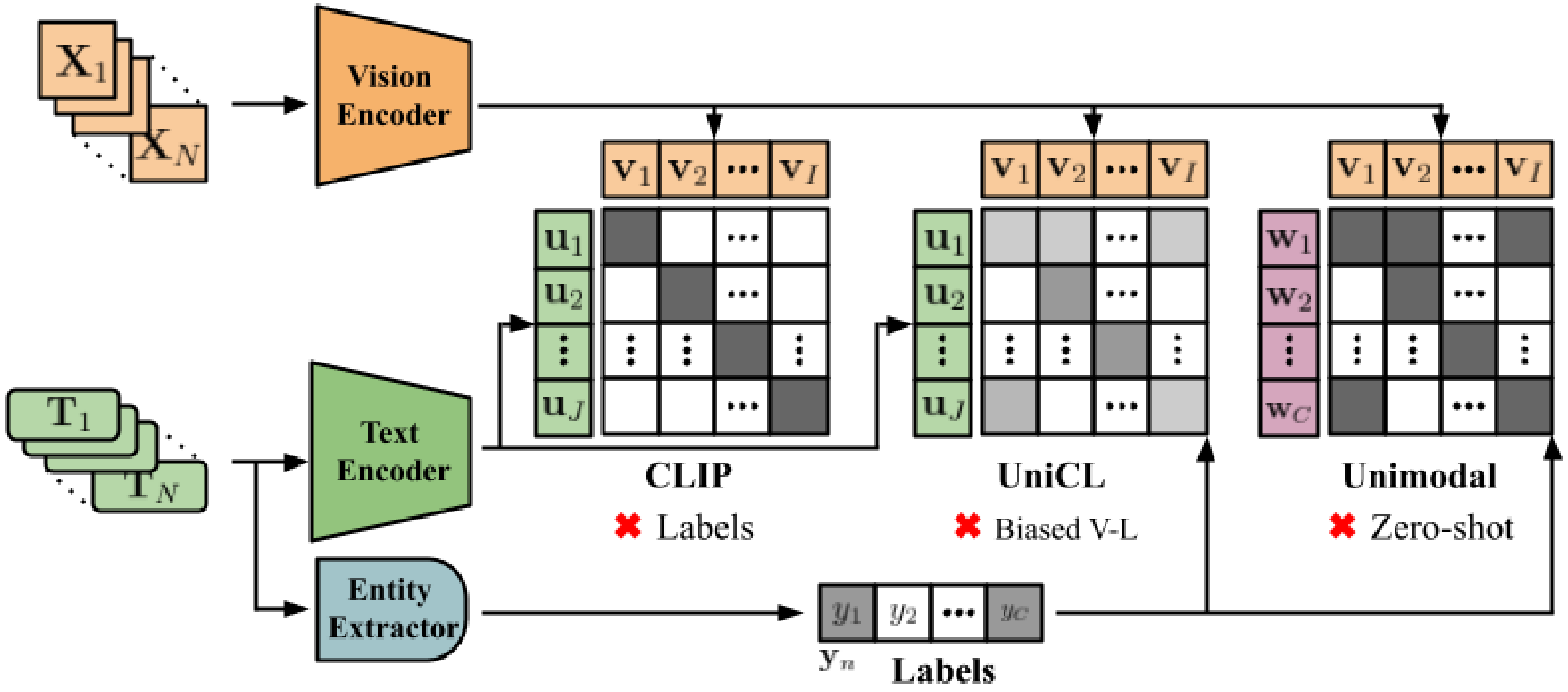
UniCL Loss

$$\mathcal{L}^{\mathrm{i2t}}_{\mathrm{UniCL}}(\theta,\phi,\tau|\mathcal{B}) = -\sum_{i\in\mathcal{B}} \frac{1}{|P_{\mathrm{i2t}}(i)|} \sum_{i'\in P_{\mathrm{i2t}}(i)} \log\frac{\exp(\mathbf{u}_i^T\mathbf{v}_{i'}/\tau)}{\sum_{j\in\mathcal{T}_{\mathcal{B}}}\exp(\mathbf{u}_i^T\mathbf{v}_j/\tau)}$$

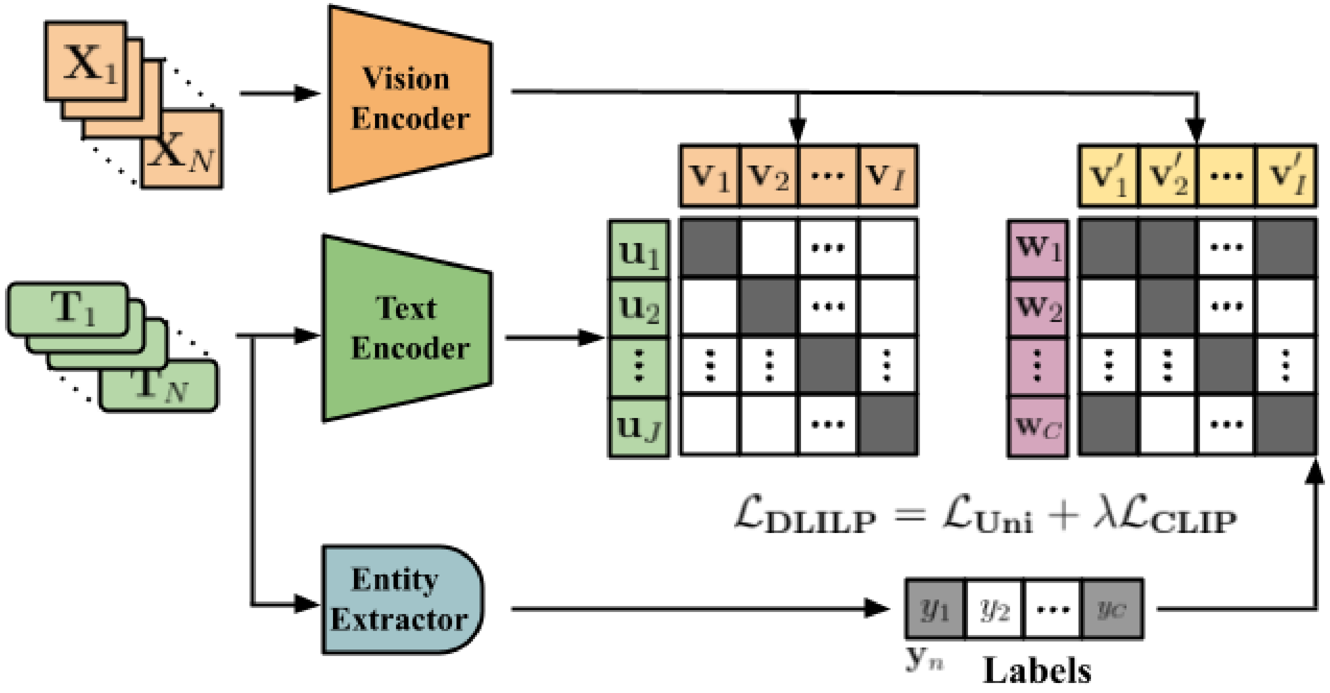
$$\mathcal{L}^{\mathrm{t2i}}_{\mathrm{UniCL}}(\theta,\phi,\tau|\mathcal{B}) = -\sum_{j\in\mathcal{B}} \frac{1}{|P_{\mathrm{t2i}}(j)|} \sum_{j'\in P_{\mathrm{t2i}}(j)} \log\frac{\exp(\mathbf{u}_{j'}^T\mathbf{v}_j/\tau)}{\sum_{i\in\mathcal{X}_{\mathcal{B}}}\exp(\mathbf{u}_i^T\mathbf{v}_j/\tau)}$$

Samples with same labels

Pitfalls on Existing Pre-training Strategies



DLILP: Disentangled Language-Image-Label Pre-training



✓ Labels
✓ Preserve V-L
✓ Zero-shot

$$\mathcal{L}_{\text{DLILP}} = \mathcal{L}_{\text{Uni}}(\{\theta_f, \theta_p^{\text{I-L}}\}, \tau^{\text{I-L}}, \mathbf{W}|\mathcal{B}) + \lambda \cdot \mathcal{L}_{\text{CLIP}}(\{\theta_f, \theta_p^{\text{I-T}}\}, \phi, \tau^{\text{I-T}}|\mathcal{B})$$

Experimental Setting

Table 1: Frontal Chest X-ray datasets assembly. We compiled open-access datasets for training and evaluation. Green-colored categories indicate novel classes not explicitly used during CheXpert and MIMIC pre-training.

Pre-train	**#Images**	**Reports**	**#Labels**	**Categories**
CheXpert (C) [13]	191,026		14	[NoF, EnlCard, Card, LuLes, LuOp, Edema, Cons,
MIMIC (M) [16]	154,595	×	14	PnMo, Atel, PnThor, PleEffu, PleOth, Fract, Dev]
PadChest* (P) [3]	96,201		84	(see Supp. Materials)
Evaluation	**#Train**	**#Test**	**#Labels**	**Categories**
$\text{CheXpert}_{5\times200}$	1,000	1,000	5	[Atel, Card, Cons, Edema, PleEffu]
$\text{MIMIC}_{5\times200}$	1,000	1,000	5	[Atel, Card, Cons, Edema, PleEffu]
RSNA [33]	8,400	3,600	2	[NoF, PnMo]
SSIM[3]	4800	1200	2	[NoF, PnThor]
COVID [4, 30]	1,200	4,000	4	[Normal, Covid, PnMo, LuOp]
NIH-LT [10, 40]	920	920	20	[Atel, Card, PleEffu, Inf, mass, Nod, PnMo, noF, PnThor, Cons, Edema, Emph, Fib, PleThi, PnPer, PnMed, SubEm, TorAor, CalAor]
VinDr [26]	2,000	2,000	5	[NoF, Bro, BrPn, BrLi, PnMo]

*PadChest is used for additional scalability experiments, only when specified.

Not explicitly labeled on pre-training dataset

1. Pre-training using image-text-label datasets.

A. Image-Label.

"A chest x-ray of ***[CLS]*** *" / There is* ***[CLS]****"*

B. Image-Text.

"…" → NLP → [0 0 0 0 0 0 1 0 0]

CheXpert-Labeler

2. Transferability

A. Zero-shot classification.
B. Few-shot Linear Probing.
C. Base/New disentanglement.

Transferability Performance

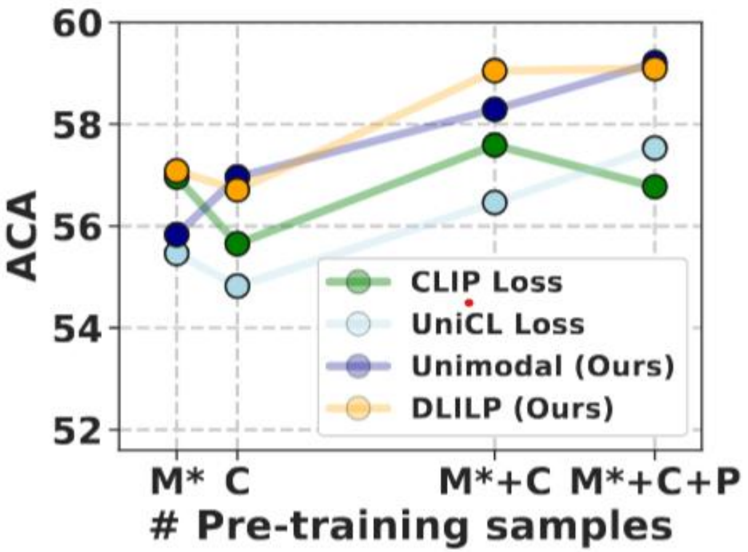


Fig. 2: Scalability of pre-training methods. Linear Probing results using $K=16$. M: MIMIC; C: CheXpert; P: PadChest.

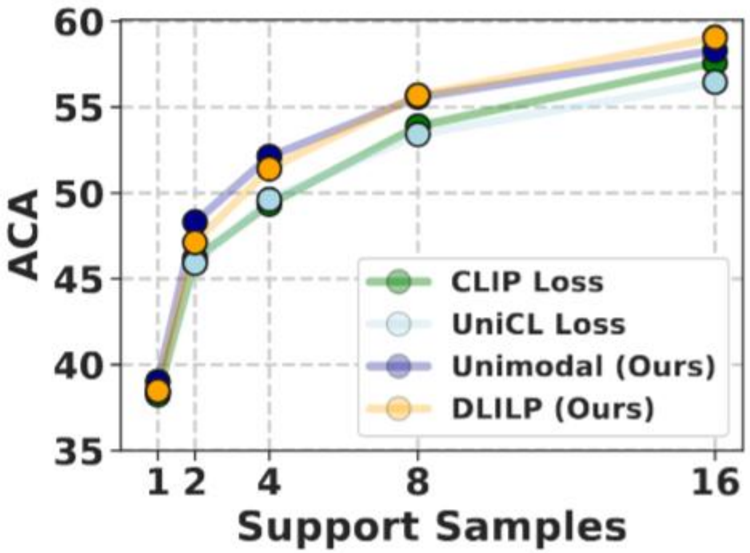


Fig. 3: Few-shot transferability results using Linear Probing for baseline and proposed pre-training strategies.

Base/New Categories Evaluation

Table 2: **Generalization/Transferability results**. Performance of different pre-training strategies disentangling known ($\mathcal{B}$) and new findings ($\mathcal{N}$).

Pre-training	CheXp	MIMIC	SSIM	RNSA	NIH_{LT}		VinDR		**Avg.**		
	$\mathcal{B}$	$\mathcal{B}$	$\mathcal{B}$	$\mathcal{B}$	$\mathcal{B}$	$\mathcal{N}$	$\mathcal{B}$	$\mathcal{N}$	$\mathcal{B}$	$\mathcal{N}$	Avg.
(a) Zero-shot generalization											
CLIP Loss [29]	51.50	49.70	77.80	63.04	40.98	29.10	68.66	32.20	58.61	**30.65**	44.63
UniCL Loss [46]	45.40	46.60	75.30	90.86	57.66	9.10	73.16	42.20	64.83	25.65	45.24
Unimodal	42.80	47.40	77.20	94.60	61.70	-	65.80	-	**64.92**	-	-
DLILP	49.50	48.60	77.90	93.50	60.80	29.10	54.20	31.10	64.08	30.10	**47.09**
(b) Linear probing transferability ($K = 16$)											
CLIP Loss [29]	54.50	49.60	69.10	93.20	46.52	32.50	71.68	38.20	64.10	35.35	49.73
UniCL Loss [46]	53.10	50.90	65.58	93.78	46.50	27.52	71.32	37.54	63.53	32.53	48.03
Unimodal	54.20	53.70	67.68	94.36	47.16	33.20	75.34	37.44	65.41	35.32	50.37
DLILP	55.60	54.50	72.74	93.82	50.66	32.24	71.36	40.76	**66.45**	**36.50**	**51.48**

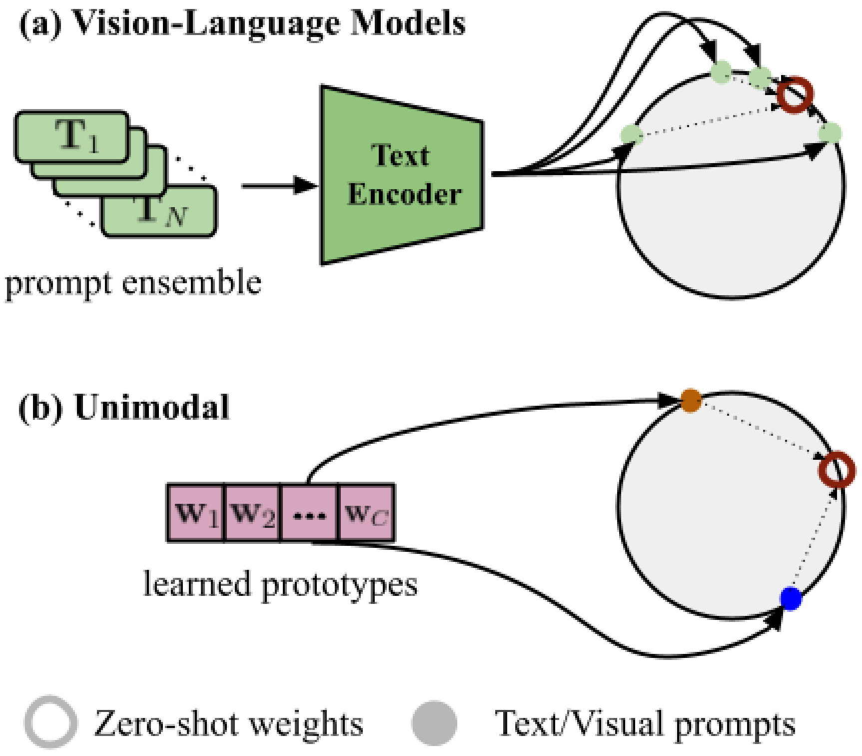
No Zero-Shot?

- MedCLIP (EMNLP22) and MedKLIP (ICCV23) defend the effectivenes of visión-language models to **generalize to novel categories using the COVID disease prediction**.

Name: [**Description**]
COVID: [“*the presence of patchy or confluent band like ground glass opacity or consolidation*”]

Table 3: **Zero-shot on COVID.**

Pre-training	2-class Name	2-class Desc.	4-class Name	4-class Desc.
MedCLIP [43]	74.1	78.8	40.5	42.9
MedKLIP [29]	51.8	82.9	20.2	32.5
CLIP Loss [29]	69.6	74.2	32.7	48.8
UniCL Loss [46]	80.5	83.7	45.5	44.8
Unimodal	-	**85.1**	-	**51.6**
DLILP	77.0	81.6	36.6	50.0



SoTA Comparison

Table 4: **Available vision-language models transferability.** Linear probing results ($K = 16$) for SoTA pre-trained models.

Method	Data	CheXp	MIMIC	SSIM	RNSA	NIH	VinDR	COVID	**Avg.**
$\text{MedKLIP}_{\text{ICCV'23}}$[44]	M	34.30	32.60	64.82	88.18	14.04	26.34	68.04	46.90
$\text{KED}_{\text{Nat.Com.'23}}$[49]	M	42.50	40.20	66.04	92.12	19.40	26.18	73.24	51.38
$\text{BioVIL}_{\text{ECCV'22}}$[3]	M+Pub	46.70	43.80	73.68	94.08	21.22	26.20	62.46	52.59
Unimodal	M	51.80	51.30	68.04	93.42	21.20	27.68	77.40	55.83
DLILP	M	53.30	52.90	69.80	93.78	25.34	26.84	77.62	**57.08**
$\text{GlorIA}_{\text{ICCV'21}}$[12]	C	46.00	41.60	66.30	91.16	18.78	23.02	72.92	51.40
Unimodal	C	52.30	48.20	71.52	93.88	24.20	29.14	79.48	**56.96**
$\text{MedCLIP}_{\text{EMNLP'22}}$[43]	M+C	54.40	50.50	69.48	94.20	20.98	27.80	72.30	55.67
$\text{CXR-CLIP}_{\text{MICCAI'23}}$[47]	M+C	52.20	46.10	69.34	92.00	25.90	26.26	76.82	55.52
Unimodal	M+C	54.20	53.70	67.68	94.36	26.20	30.26	81.62	58.29
DLILP	M+C	55.60	54.50	72.74	93.82	26.72	28.98	81.02	**59.05**
Unimodal	M+C+P	56.00	55.20	73.84	94.00	26.12	28.48	80.86	**59.21**
DLILP	M+C+P	56.30	53.00	73.56	94.08	25.72	29.28	81.66	59.09

M: MIMIC; C: CheXpert; P: PadChest; PubMed: PubMed text abstracts.

Take-home messages

- **Vision-language pre-training** is a powerfull tool to leverage **large datasets with text supervision**.
- **Nevertheless, it does not do magic**. The **zero-shot performance is highly correlated with the class proportion present in the pre-training dataset***. **No "Zero-Shot" Without Exponential Data: Pretraining Concept Frequency Determines Multimodal Model Performance (ICLR24), DPFM Workshop.**
- **In the medical domain, several challenges limit its usability**.
    - Challenging expert, fine-grained concepts.
    - Limited data with text supervision: label information is predominant.
- Simple **Supervised pre-training is largely competitive**.
- We still need **better methods to combine fine-grained labels and weak text supervisión.**
    - Let's not cheat ourselves : Base/New evaluation.

Le génie pour l'industrie

A Reality Check of Vision-Language Pre-training in Radiology: Have We Progressed Using Text?

Julio Silva-Rodríguez, Jose Dolz and Ismail Ben Ayed

ÉTS Montreal