Medical Image Understanding and Analysis (MIUA), 2025

PolypDINO: Adapting DINOv2 for domain generalized polyp segmentation

Mansoor Ali[1], **Raneem Toman**[2], Gilberto Ochoa Ruiz[2], Sharib Ali[2]

Tecnologico de Monterrey, Mexico[1], University of Leeds, UK[2]

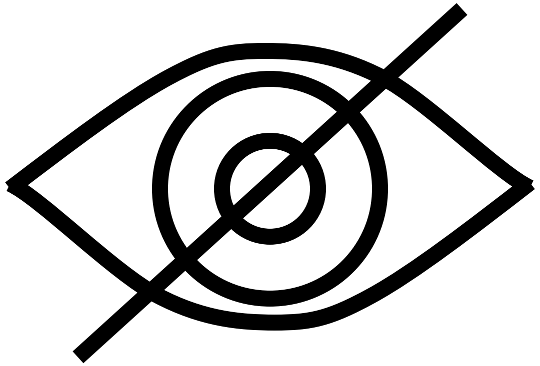
Tecnológico de Monterrey

CONAHCYT
CONSEJO NACIONAL DE HUMANIDADES CIENCIAS Y TECNOLOGÍAS

UNIVERSITY OF LEEDS

CONTENT WARNING!

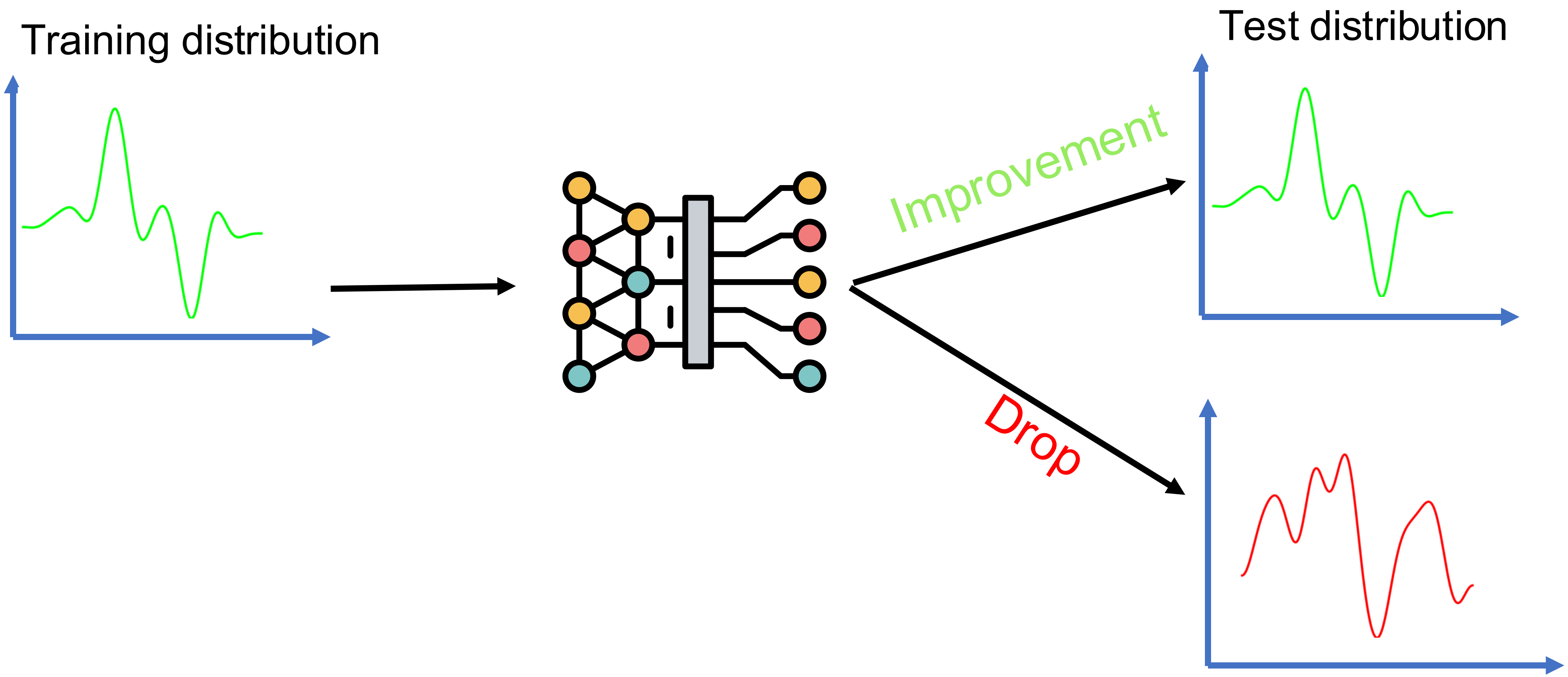
WARNING



The following presentation contains **sensitive content from real-life endoscopic images** which some people might find disturbing or traumatizing.

Risks of today's ML models!

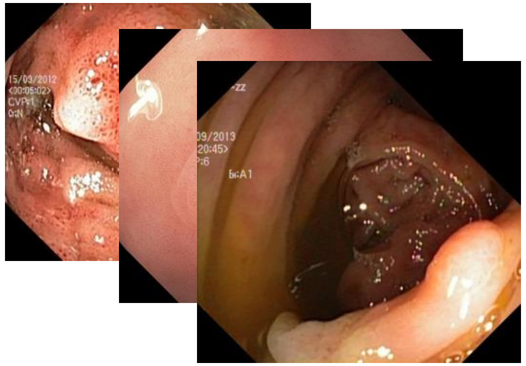
Most ML methods are developed under I.I.D hypothesis



Risks of today's ML models!

Risks are more critical in high-stake scenarios, such as endoscopic imaging

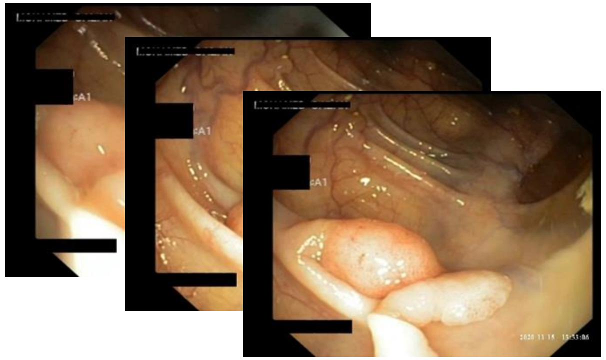
Hospital



Training set

Domain shift

Hospital



Test set

Drop in performance -14%

Kvasir-SEG

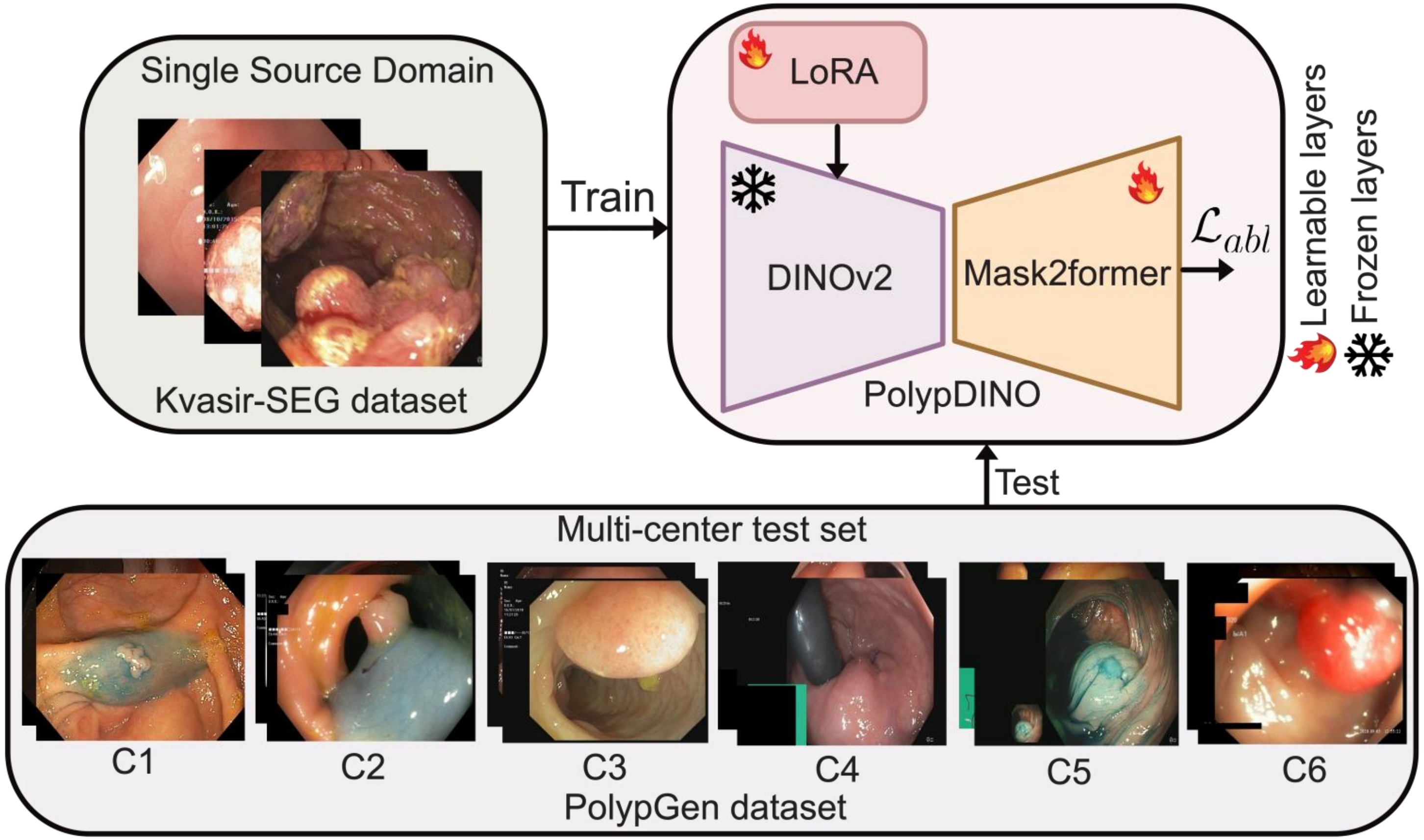
DeepLabv3+ model

PolypGen-C1

Objective

Goal- Learning a generalizable representation from a single source domain to perform better on multiple unseen target domains.

Approach



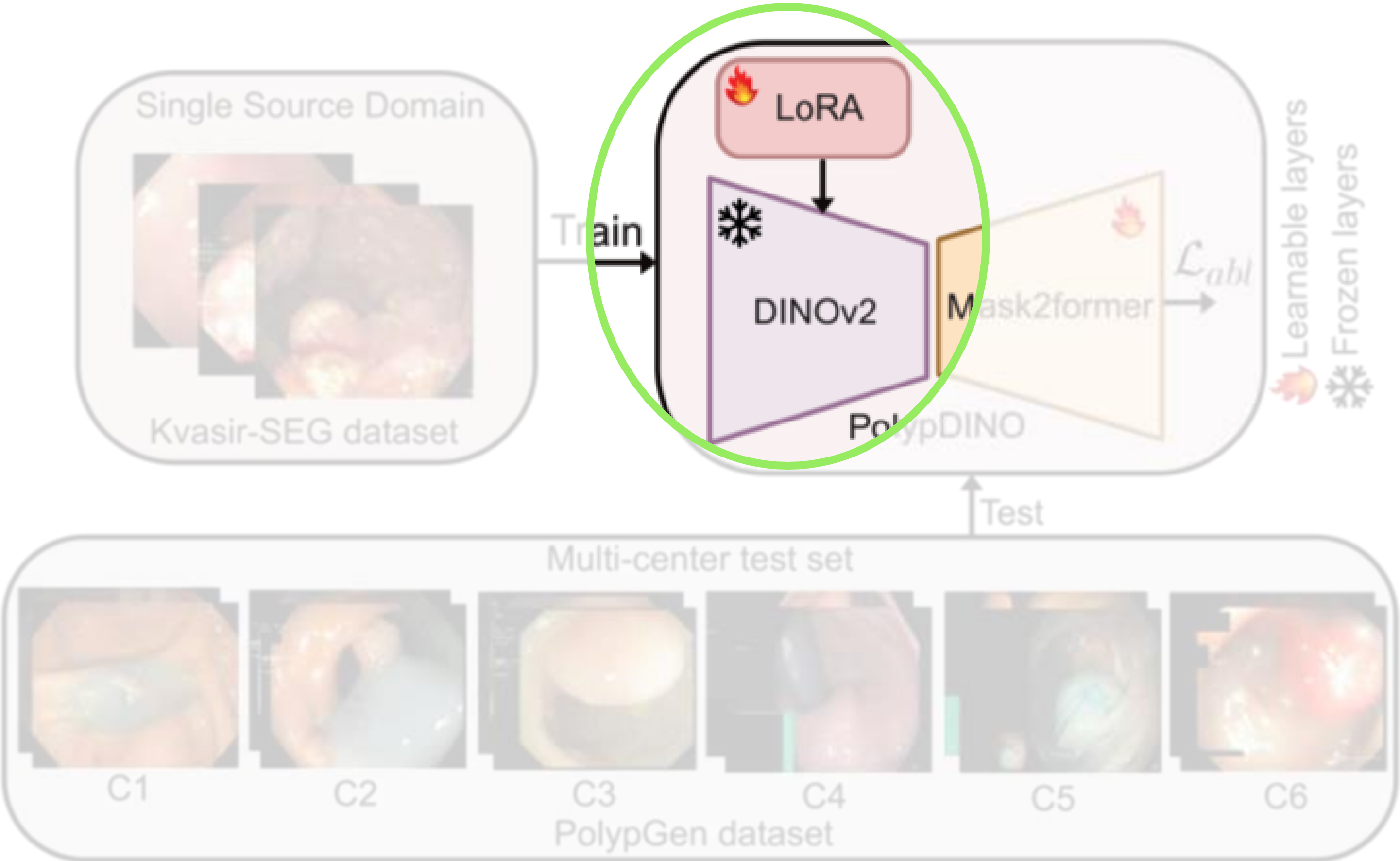
Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., ... & Bojanowski, P. (2023). Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Cheng, B., Misra, I., Schwing, A. G., Kirillov, A., & Girdhar, R. (2022). Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* .

Objective

Goal- Learning a generalizable representation from a single source domain to perform better on multiple unseen target domains.

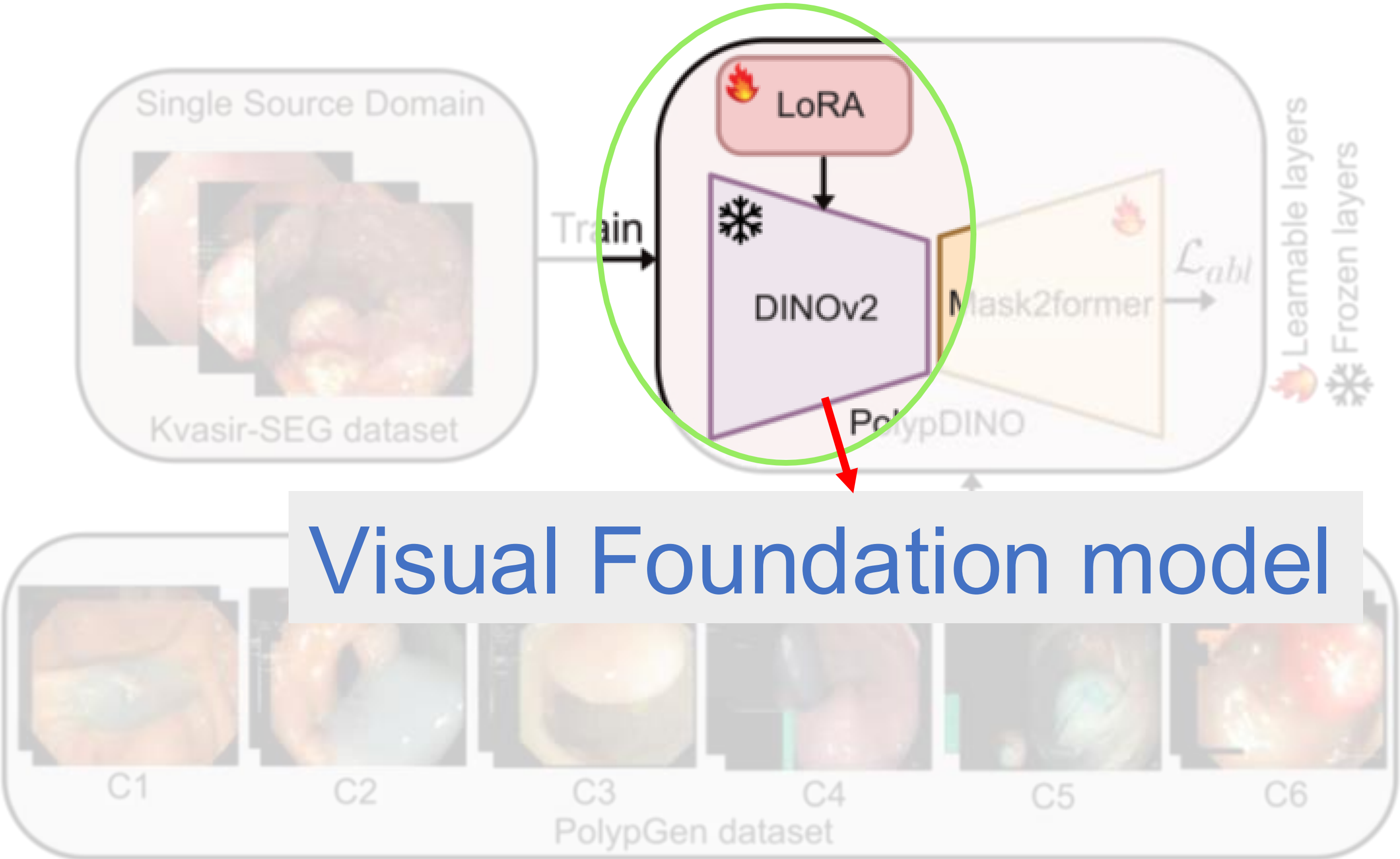
Approach



Objective

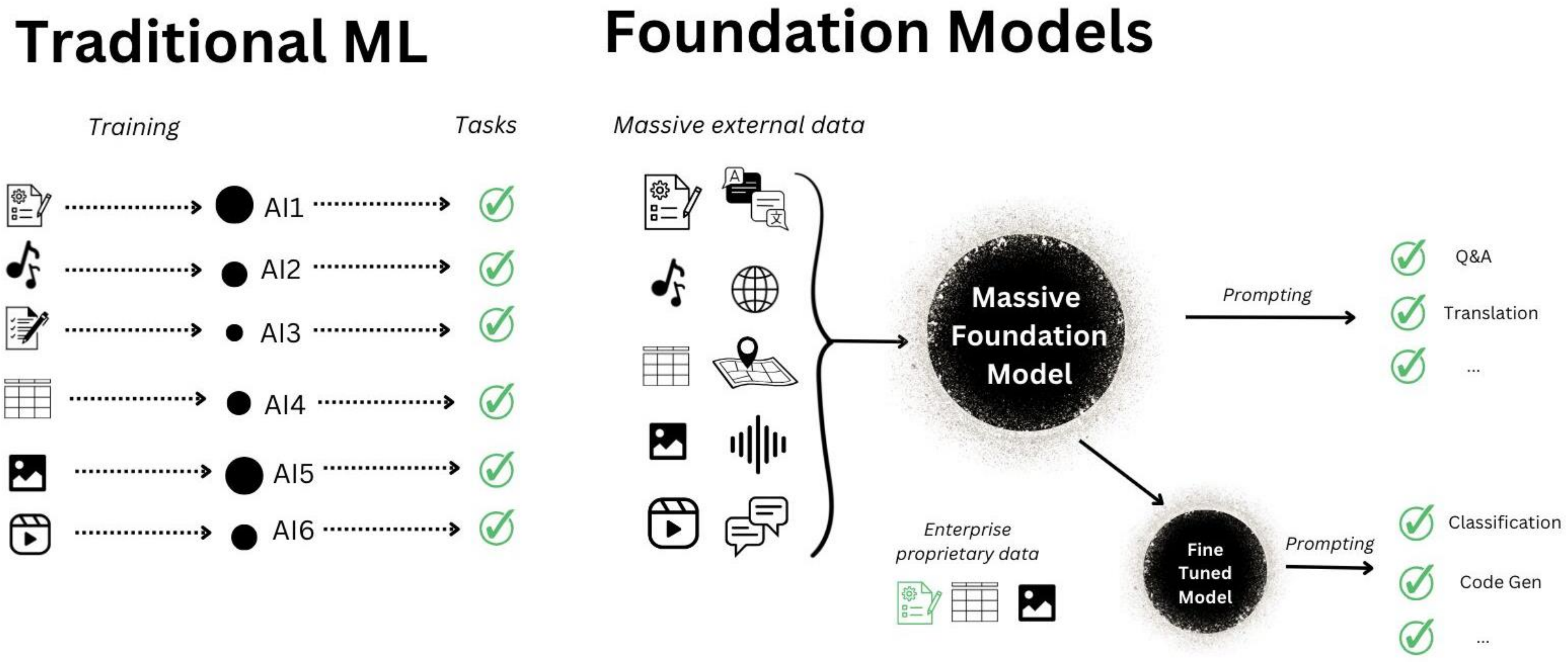
How generalizable are the foundation models when applied to different demographics, cameras source settings?

Approach



Visual Foundation model

Objective

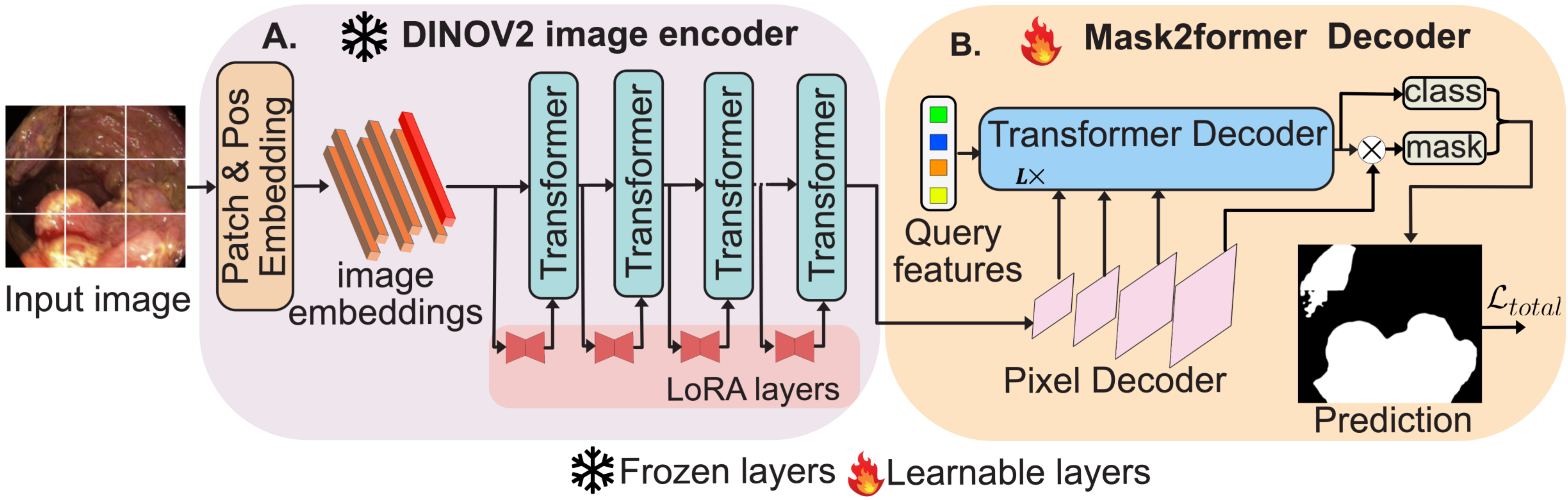


Individual models for tasks

Massive multi-tasking model

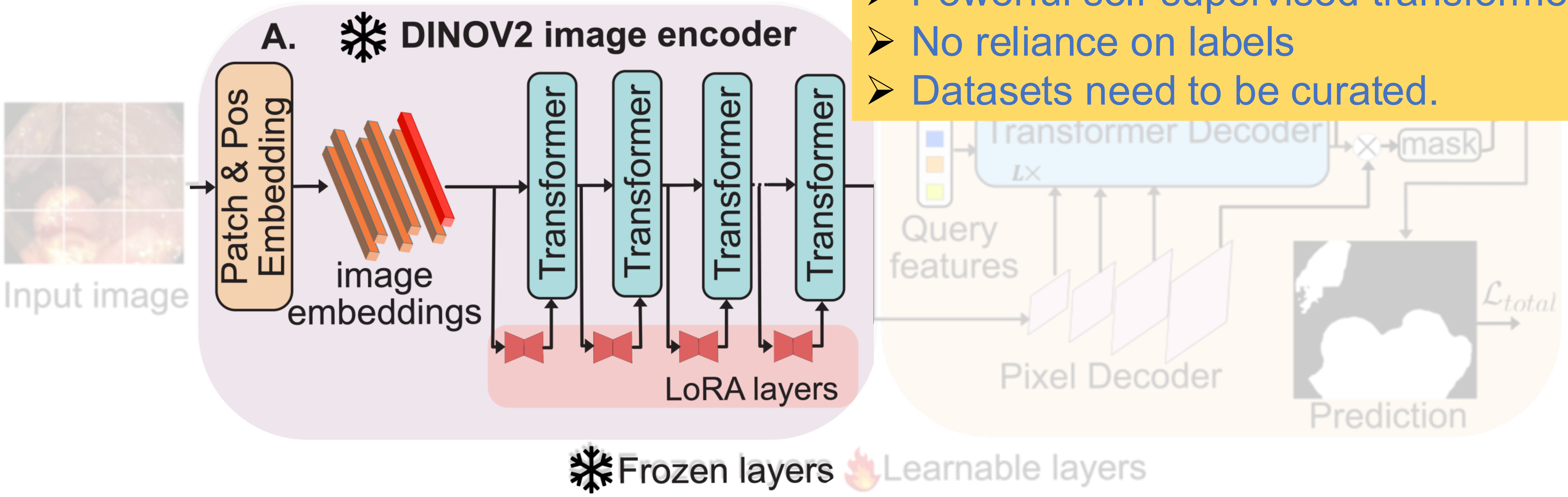
https://humanloop.com/blog/foundation-models

Methodology

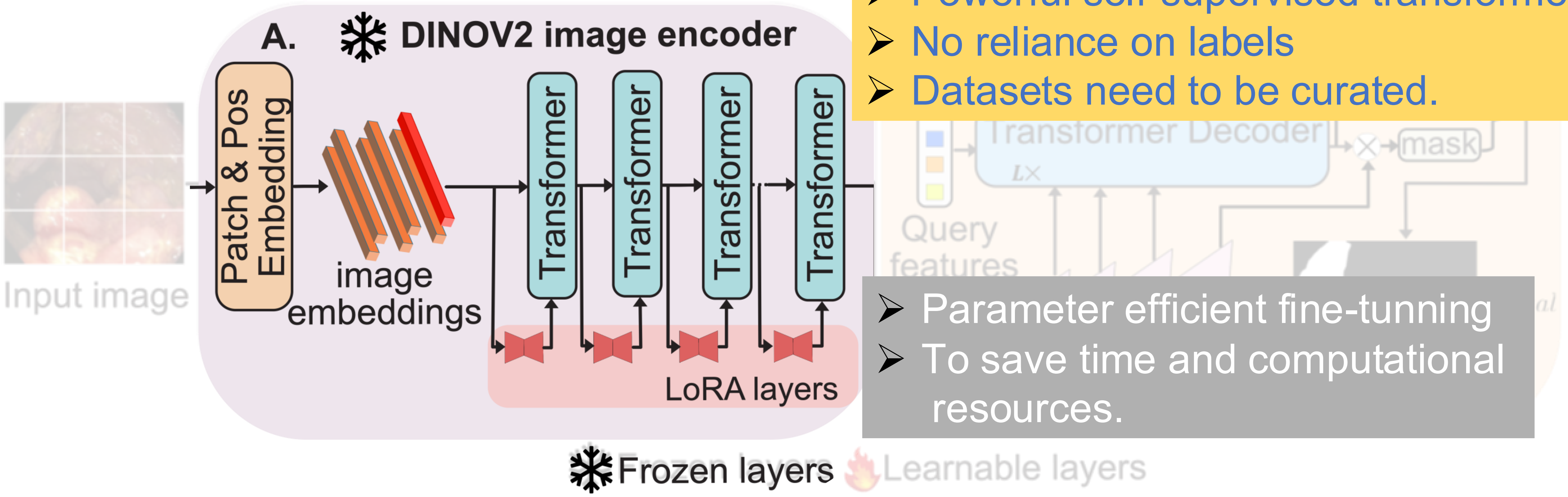


Methodology

- ➢ Powerful self-supervised transformer
- ➢ No reliance on labels
- ➢ Datasets need to be curated.



Methodology

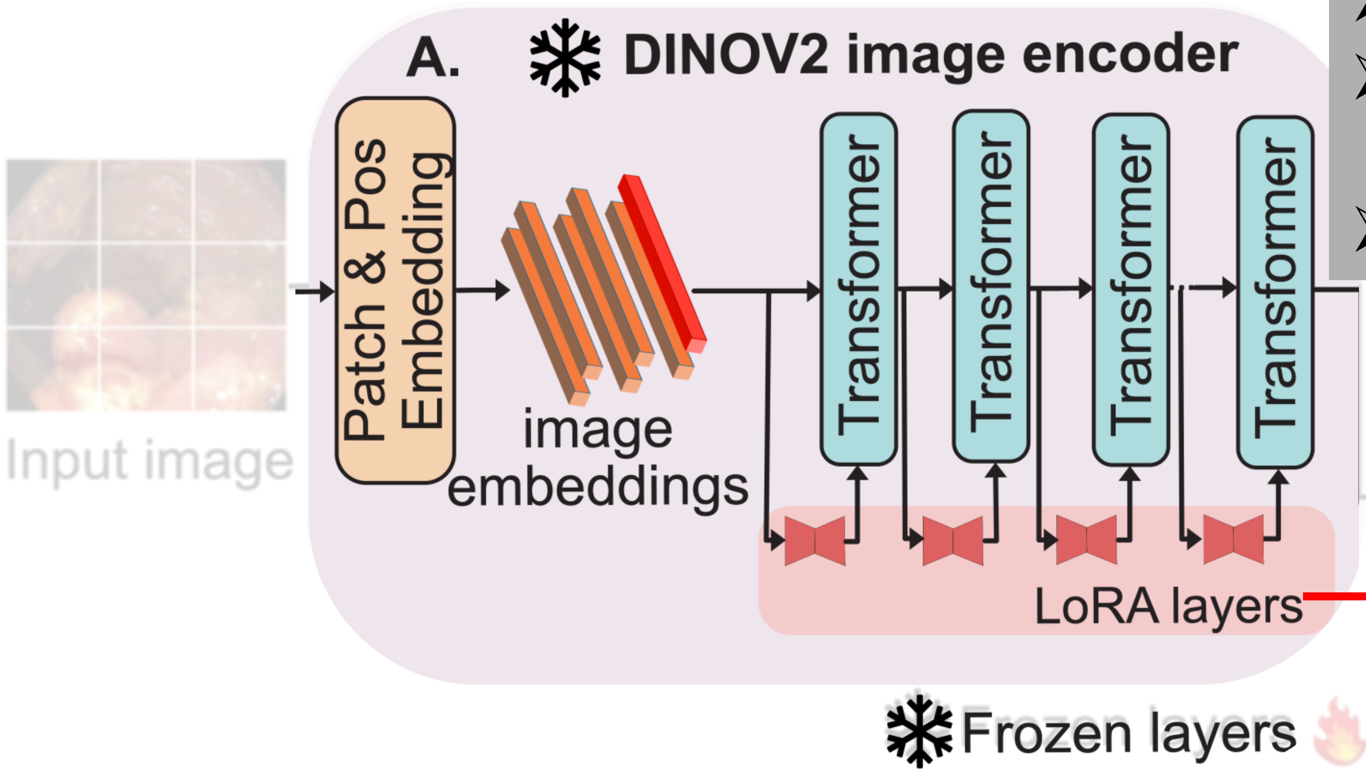


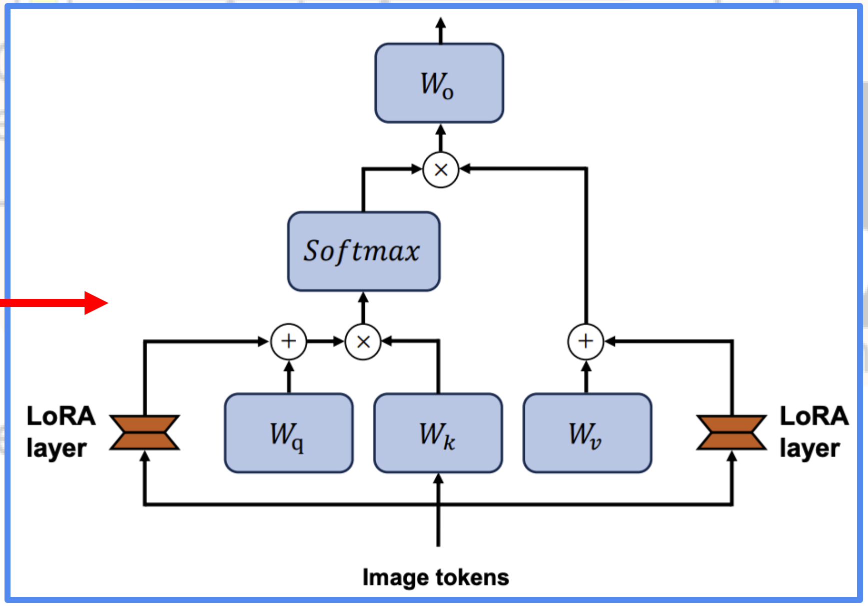
- Powerful self-supervised transformer
- No reliance on labels
- Datasets need to be curated.

- Parameter efficient fine-tunning
- To save time and computational resources.

Methodology

- Parameter efficient fine-tunning
- To save time and computational resources.
- We use Low-rank adaptation (LoRA)

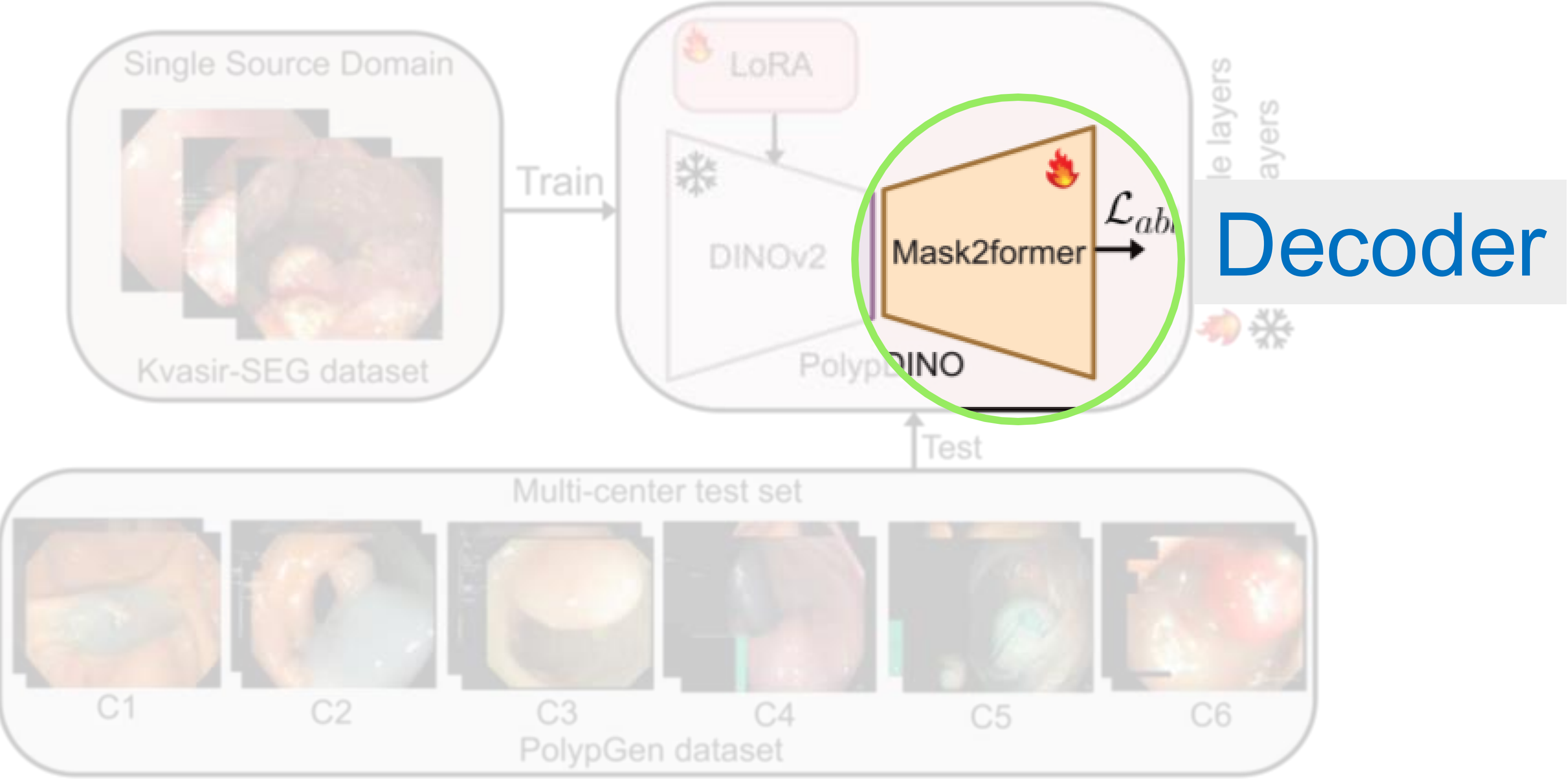




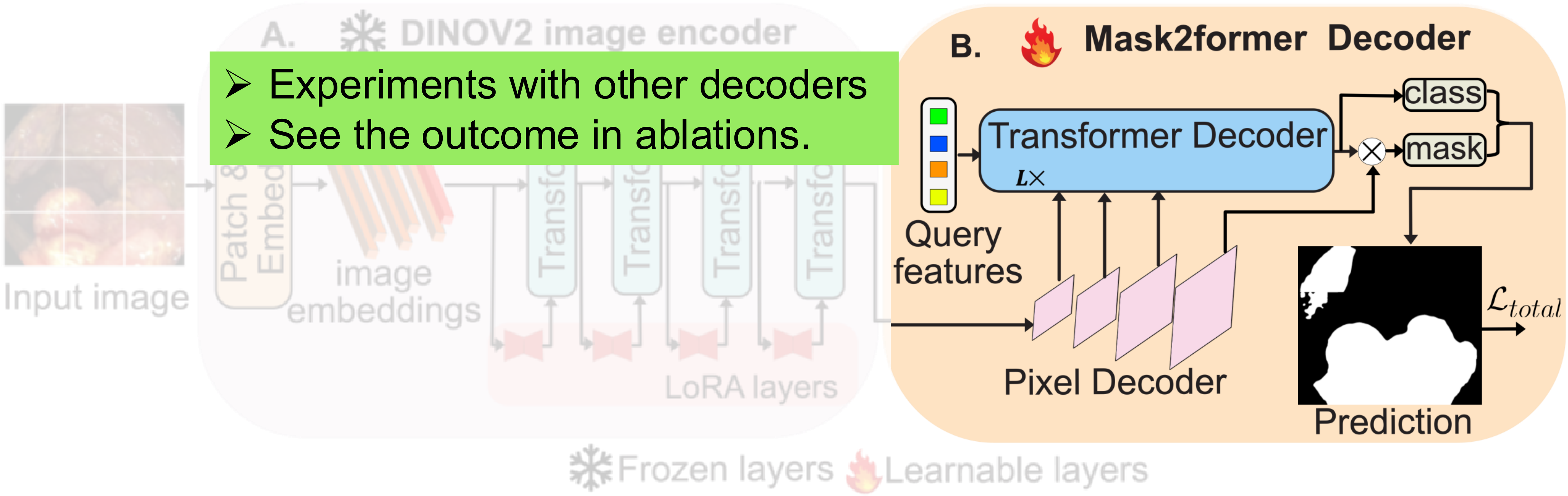
Low-rank adaptation

Methodology

Approach

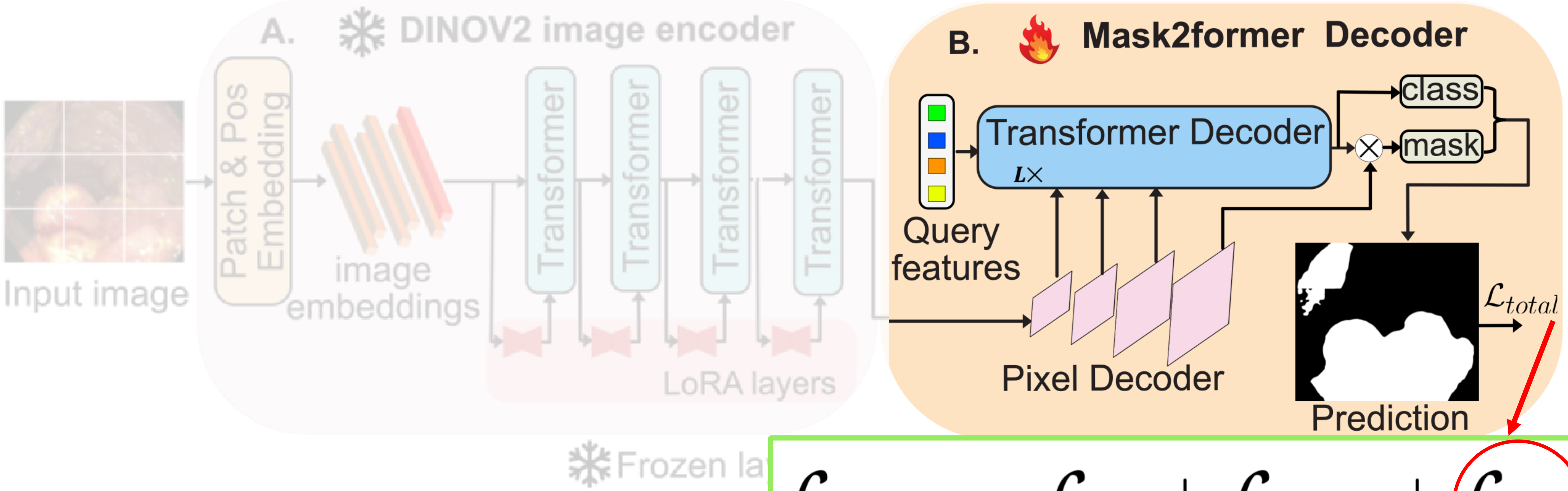


Methodology



- Experiments with other decoders
- See the outcome in ablations.

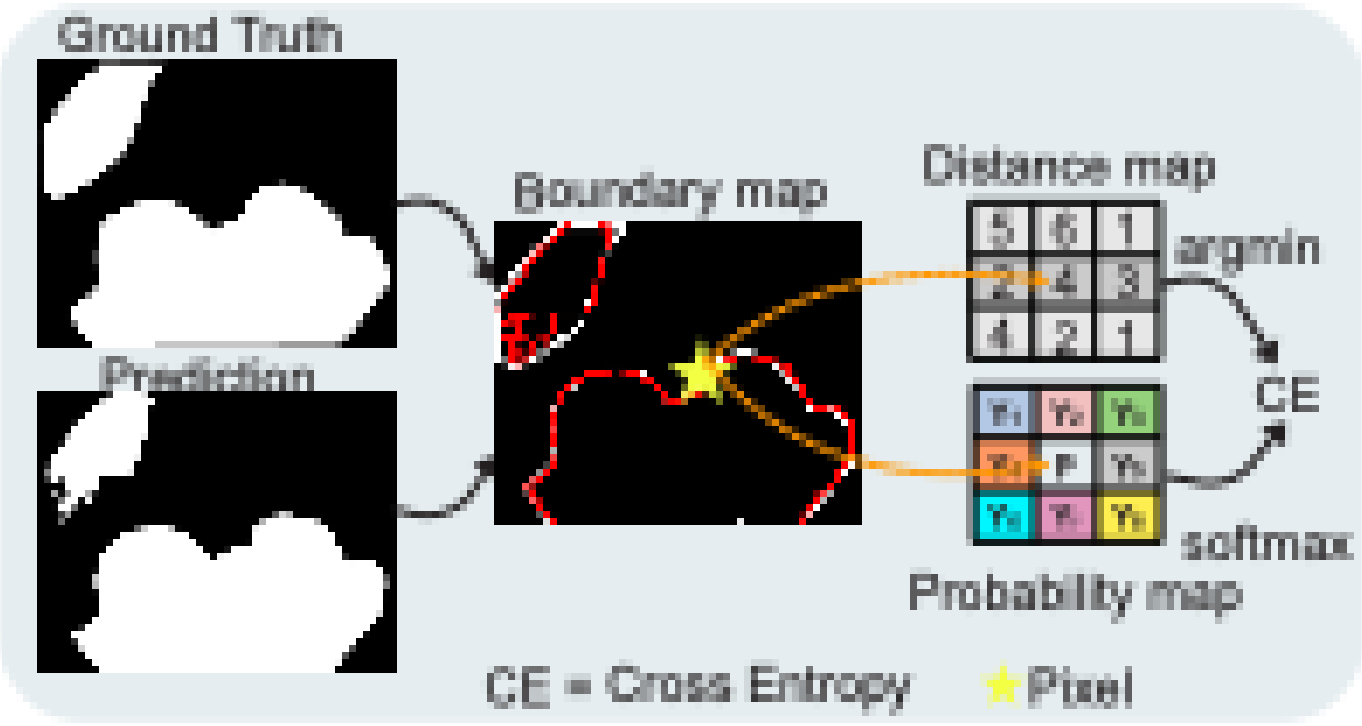
Methodology



$$\mathcal{L}_{total} = \mathcal{L}_{ce} + \mathcal{L}_{dice} + \mathcal{L}_{abl}$$

Methodology

$$\mathcal{L}_{total} = \mathcal{L}_{ce} + \mathcal{L}_{dice} + \mathcal{L}_{abl}$$



Pred, gt boundary vectors

$$\mathcal{L}_{abl} = \frac{1}{N_p} \sum_{i}^{N_p} w(M_i) \mathcal{L}_{ce}(D_i^p, D_i^g)$$

Boundary distance transform

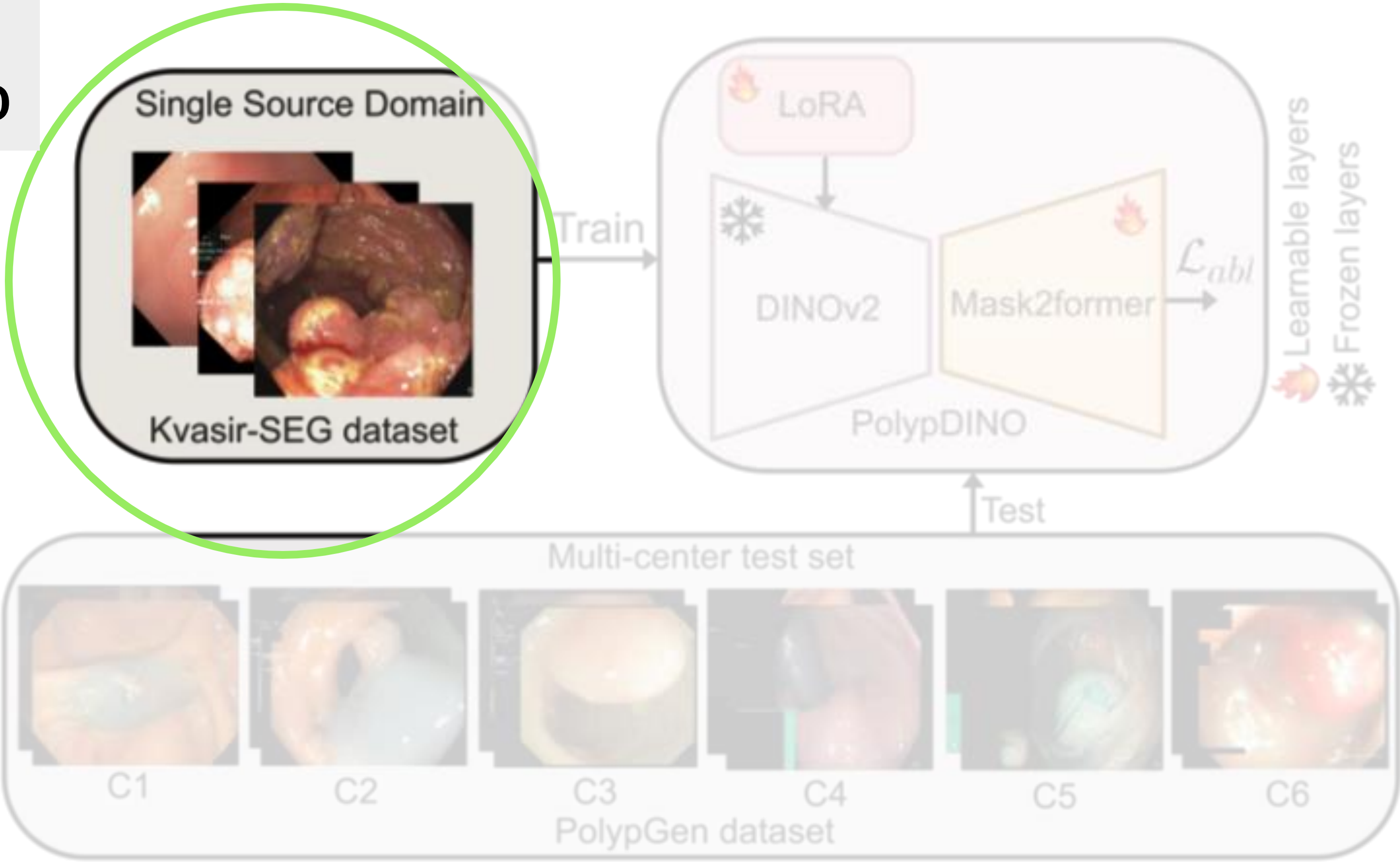
weight

Wang, C., Zhang, Y., Cui, M., Ren, P., Yang, Y., Xie, X., ... & Xu, W. (2022, June). Active boundary loss for semantic segmentation. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 36, No. 2, pp. 2397-2405).

Datasets and experiments

Training- 80%
Validation – 20%

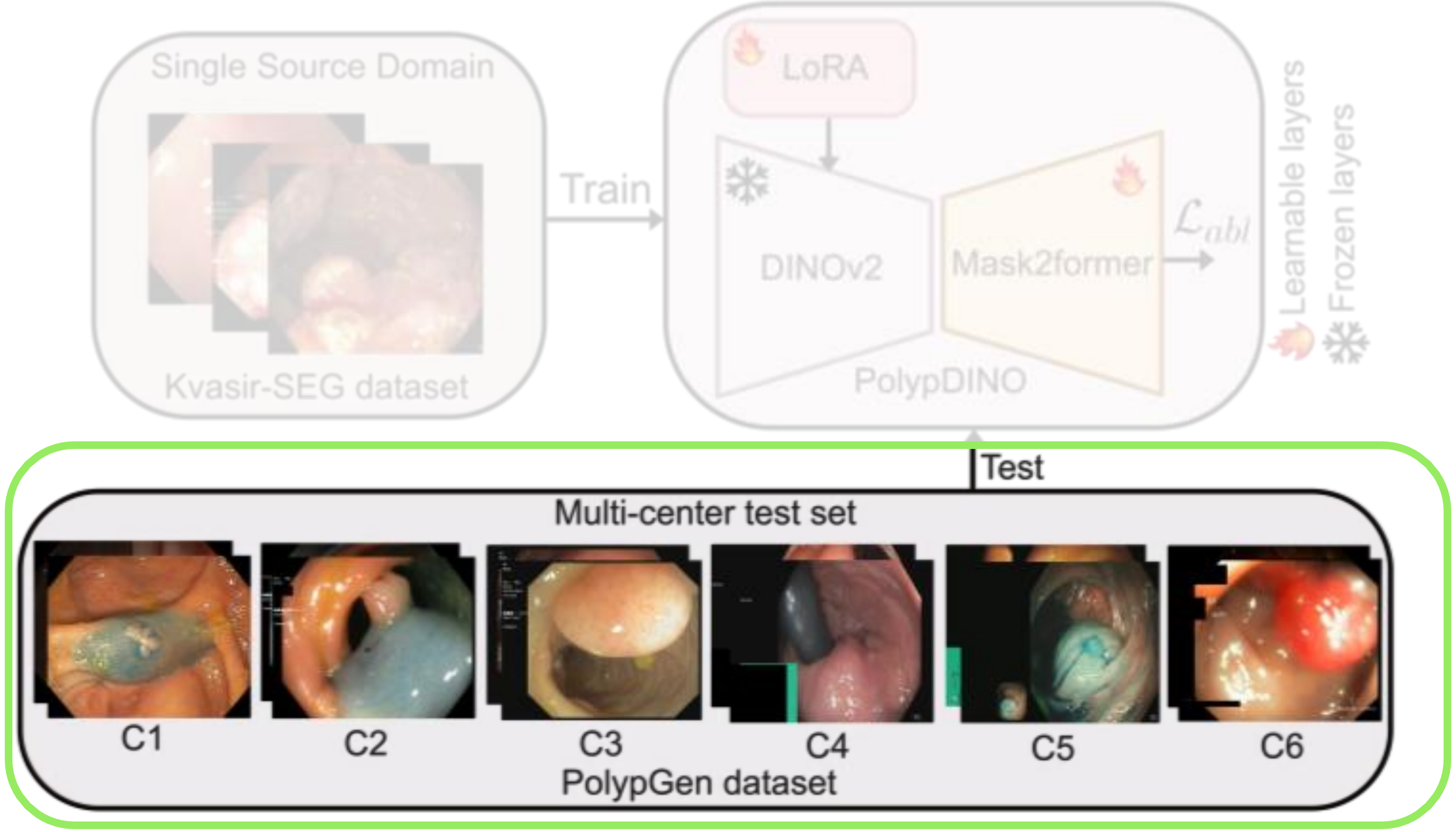
Approach



Datasets and experiments

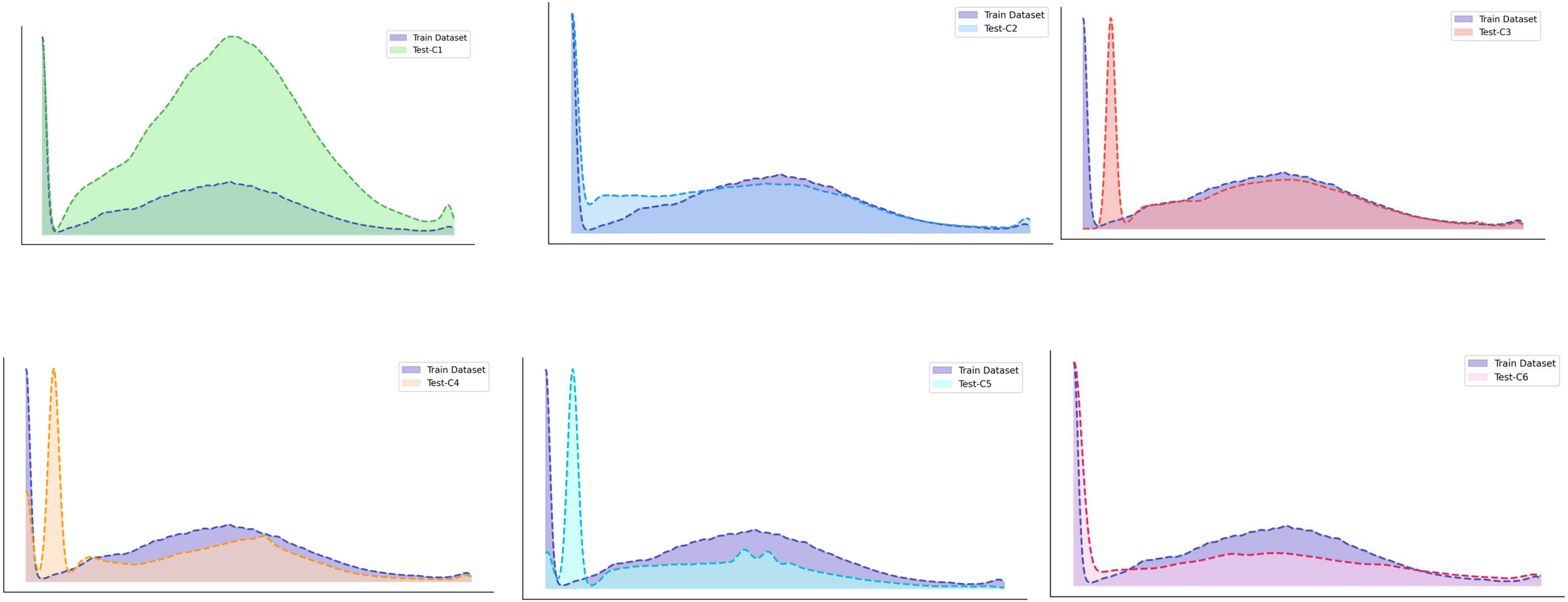
Approach

Test dataset



Multi-center PolypGen dataset

Datasets and experiments



Comparison of normalized intensity histograms between training and out-of-distribution test sets

Results

Training – Kvasir-SEG	Test set- PolypGen-C1			Test set- PolypGen-C1		
Model	mIoU	mDSC	F_2	mIoU	mDSC	F_2
DeepLabv3+	74.75	85.55	84.45	50.44	67.06	72.59
IBN-Net	74.64	85.48	84.73	63.12	77.39	80.47
RobustNet	75.99	86.36	86.5	64.56	78.46	83.12
SAN-SAW	73	84.39	83.98	57.89	73.33	78
TransNetR	65.38	72.04	72.69	66.08	72.32	73.66
SNR	78.26	87.8	87.23	68.77	81.5	81
MLMI'24	79.54	88.61	87.59	69.73	82.17	81
SAMed	78.28	87.82	87.42	64.5	78.42	76
EVA-02	77.47	87.31	85.77	56.22	71.98	79.33
MaskDINO	67.08	80.31	73.74	37.5	54.55	44.71
PolypDINO (Ours)	79.5	88.58	91.07	66.5	79.88	76.63
PolypDINO (w/ ABL)	83.48	91	91.9	65.59	79.22	81.68

CNN backbone

VFM backbone

C1 – 4% IoU

Results

	Training – Kvasir-SEG	Test set- PolypGen-C3			Test set- PolypGen-C4		
	Model	mIoU	mDSC	F_2	mIoU	mDSC	F_2
CNN backbone	DeepLabv3+	78.5	87.95	87.34	30.84	47.14	37.32
	IBN-Net	75.92	86.31	86.13	32.48	49.04	39.5
	RobustNet	77.08	87.06	87.86	38.45	55.54	48.05
	SAN-SAW	74.26	85.23	85.3	41.07	58.23	66.3
	TransNetR	72.17	78.74	78.63	46.01	50.42	50.96
VFM backbone	SNR	79.16	88.37	89.43	36.44	53.42	45.34
	MLMI'24	80.1	88.95	89.28	41.48	58.64	50.78
	SAMed	11.11	20	20.15	40.06	57.21	49.89
	EVA-02	78.15	87.74	86.19	46.6	63.58	71.46
	MaskDINO	11.17	20.08	18.17	22.11	36.2	26.83
	PolypDINO (Ours)	78.91	88.21	89.1	45.19	62.25	65.13
	PolypDINO (w/ ABL)	82.39	90.34	92.65	48.76	65.56	71.9

C5 – 5% IoU

C5 – 2% IoU

Results

	Training – Kvasir-SEG	Test set- PolypGen-C5			Test set- PolypGen-C6		
	Model	mIoU	mDSC	F_2	mIoU	mDSC	F_2
CNN backbone	DeepLabv3+	60.93	75.72	73.42	50.55	90.98	58.03
	IBN-Net	56.78	72.44	68.2	58.58	73.88	67.39
	RobustNet	58.73	74	72.54	65.63	79.25	78.03
	SAN-SAW	48.03	64.9	68.21	65.46	78.28	80.26
	TransNetR	35.97	42.14	42.32	63.35	69.17	68.03
VFM backbone	SNR	64.5	78.42	80.36	67	80.24	79.94
	MLMI'24	60.73	75.57	74.01	67.11	80.32	79.19
	SAMed	48.4	65.23	61.23	65.49	79.15	82.34
	EVA-02	62.94	77.26	79.43	56.04	71.83	78.41
	MaskDINO	38.06	55.12	46.98	53.35	69.61	60
	PolypDINO (Ours)	66.52	79.89	84.88	66.14	79.62	79.08
	PolypDINO (w/ ABL)	69.55	82.04	82.31	69.78	82.2	88.79

C5 – 5% IoU

C5 – 2% IoU

Results

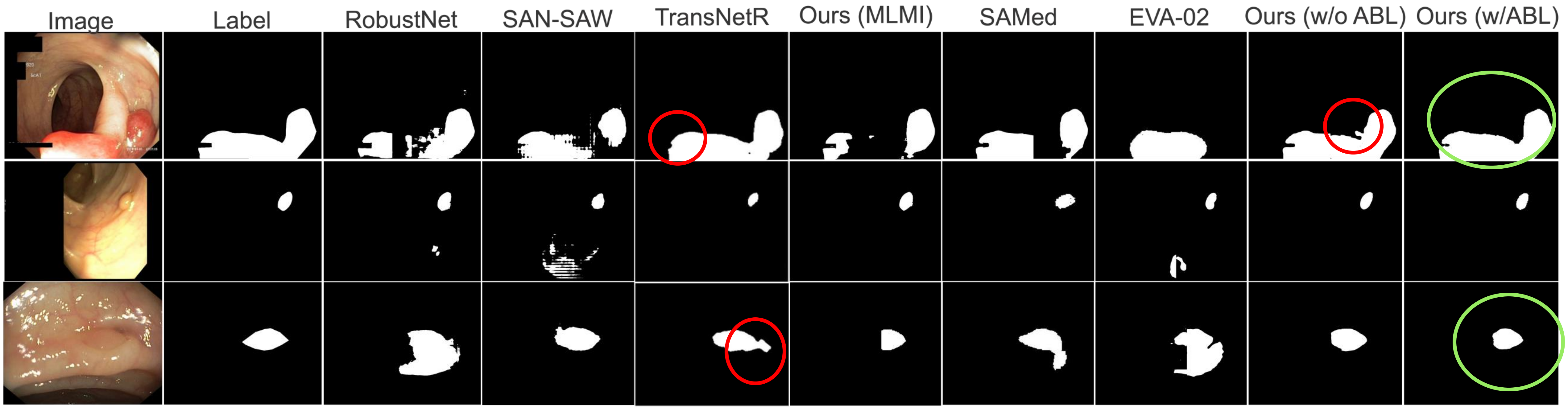


Image
Label
RobustNet
SAN-SAW
TransNetR
Ours (MLMI)
SAMed
EVA-02
Ours (w/o ABL)
Ours (w/ABL)

Ablations

Ablation-1 Impact of decoder type

Backbone	Decoder	mIoU	mDSC	Recall	Precision	F_2
DINOv2	PSP	78.11	87.71	84.09	91.66	85.5
DINOv2	Deeplabv3+	82.72	90.54	91.42	89.69	91.06
DINOv2	FCN	79.22	88.12	86.72	92.34	86.94
DINOv2	Mask2former	85.43	91.3	94.8	93.25	94.66

Ablation-2 Impact of LoRA rank size

Rank Size	mIoU	mDSC	Recall	Precision	F_2
1	84.93	90.75	93.68	92.58	93.46
4	85.37	91.22	94.59	93.37	94.52
8	85.43	91.3	94.81	93.25	94.66
16	85.25	91.04	94.62	93.03	94.21

Conclusion

Goal- The main objective of the work was to leverage visual foundation model to perform generalized polyp segmentation.

Approach – Used Dinov2 and performed efficient fine-tuning with LoRA. Also incorporated active boundary loss to enhance the polyp segmentation boundaries.

Result- We comprehensively performaed comparisons with CNN and VFM backbone methods and tested the model on PolypGen, a six center Polyp dataset. Results indictate consistent improvement in generaliability performance.

Code will be uploaded soon here.

Any detailed questions, or suggestions can be directed to – ali.mansoor2024@gmail.com