EURECOM
Sophia Antipolis

# FOUNDATION MODELS IN HEALTHCARE

**PRESENTED BY:**
**Abdessamed Qchohi, Alessio Giuffrida**

26/06/2025

# Agenda

01 Foundation Models

02 DinoV2

03 Experiments

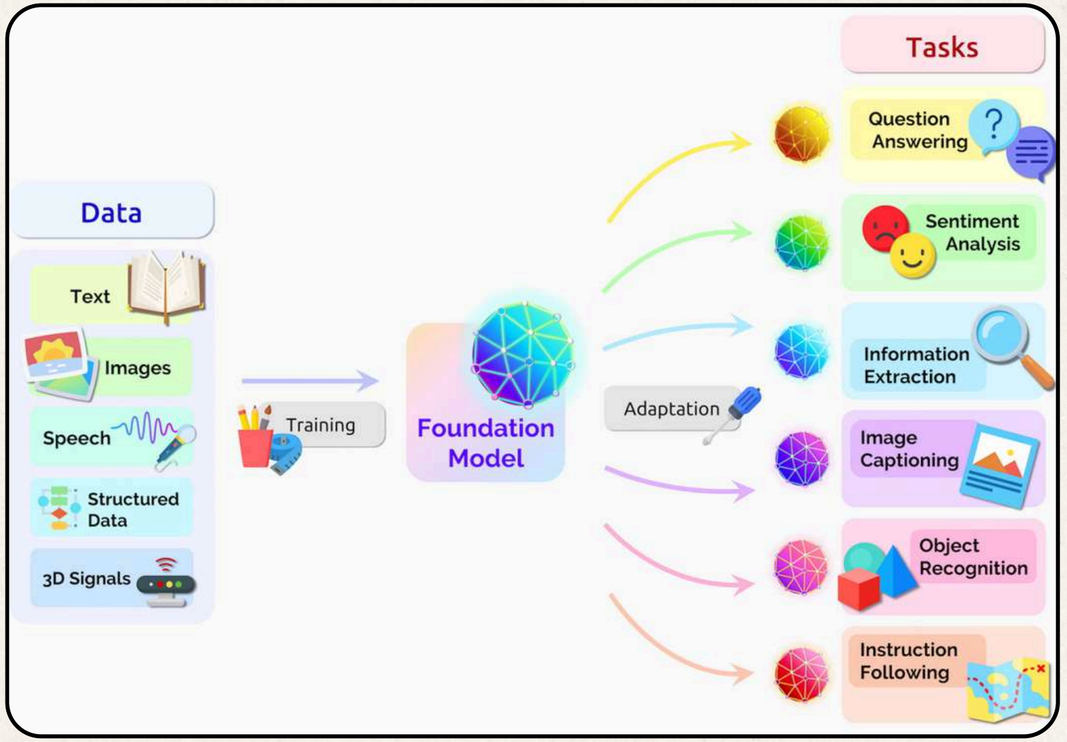
# Foundation Models

**Self-supervision**

A form of unsupervised learning where the model generates its own labels from the input data. Learning signals are created using patterns within the data.

**At Scale**

Training involves vast amounts of data and computational resources to achieve generalization and robustness.



# Goal

Start with a **shared foundation model** and **adapt it** for **specific applications**, unifying multiple tasks with shared pre-trained knowledge.

# How to use?

## Pre-Training

The model processes inputs and extracts features

## Linear Probing

A new model is trained on top of the extracted features without updating the foundation model's parameters (frozen features).

## Fine-Tuning

All model parameters are updated to better fit a specific task.  This requires more memory and it is likely to overfit.

# Model Robustness

## In-Distribution (ID)

- Inputs are drawn from the same distribution as the model's training data
- Fine-tuning typically performs better than linear probing when the dataset closely resembles the pretraining data.

## Out-Of-Distribution (OOD)

- Inputs are significantly different from the training data
- Linear probing often performs better. This happens when the pretrained features are high quality and the distribution shifts are large.

# Dinov2

**Learning Robust Visual Features without Supervision**

- **DINOv2** explores whether **self-supervised learning**, when scaled with large curated datasets, can produce **general-purpose visual features** that work across diverse tasks.

- Unlike **text-guided vision models (e.g., CLIP)**, which rely on captions, **DINOv2** learns features directly from images without supervision.

# Framework

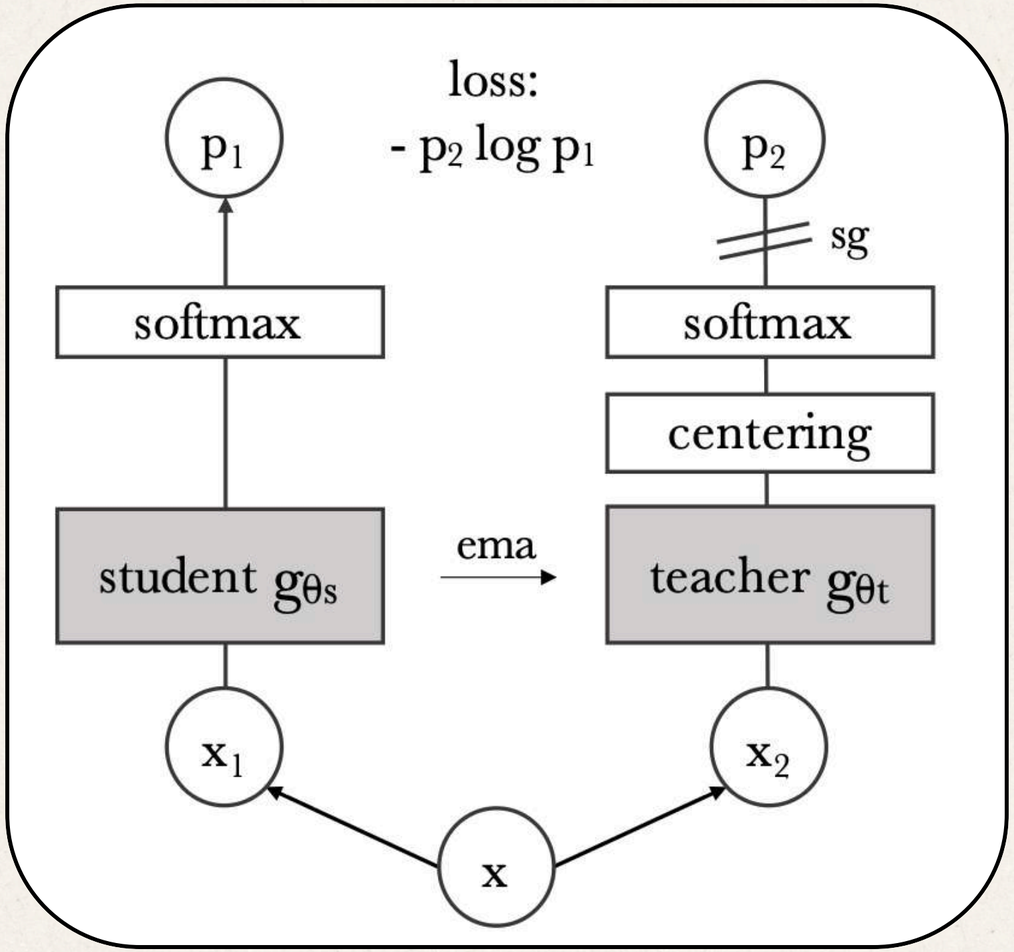
- **Core idea**: train a *student* network to match the output of a *teacher* network on various views (augmentations) of the same input image using a cross-entropy loss on softmax-normalized outputs.

- **Probability outputs** are computed by applying a temperature-scaled softmax over the model outputs. These encourage either smoother or sharper distributions depending on the temperature.

- **Training process**:
  - Multiple augmented views of an image are created—two global crops and several smaller local ones.
  - All views go through the student, whereas only the global ones go through the teacher.
  - The student is trained to align with the teacher's output across these local-global view pairs.

- **Loss**: cross-entropy between student and teacher outputs for all view combinations (excluding identical ones).

# Framework


loss:
- p2 log p1
p1
p2
sg
softmax
softmax
centering
student gθs
ema
teacher gθt
x1
x2
x

# Framework

- **Teacher update:**
  - The teacher is not fixed, it's updated as an exponential moving average (EMA) of the student (momentum encoder). This ensures that the teacher always provides a more stable target.

- **Architecture:**
  - Both student and teacher use the same network (e.g., ViT or ResNet) with different weights.

- **Avoiding collapse:**
  - DINO avoids the common SSL issue of collapse not by contrastive loss, but by:
    - **Centering** the teacher’s output → subtracting the batch mean
    - **Sharpening** → lowering the teacher softmax temperature
  - These balance each other: centering avoids dominant dimensions; sharpening avoids flat outputs.

# Loss Functions

- **DINO loss:** cross-entropy between student and teacher network

$$\mathcal{L}_{DINO} = -\sum p_t \log p_s$$

- **iBOT loss:** patch-based learning with masking, again a cross-entropy loss between the two networks

$$\mathcal{L}_{iBOT} = -\sum_i p_{ti} \log p_{si}$$

# Semantic segmentation

**Approach used:**

- Dinov2 is a **feature extractor**
- The weights are frozen and a **segmentation head** is added on top
- The head can be either **pre-trained** or **not**

# Experiments

| | |
|---|---|
| 01 | FoodSeg |
| 02 | BraTS - 3 modalities |
| 03 | BraTS - 1 modality |
| 04 | Fine-Tuning Rein-LoRa |

# FoodSeg

**First attempt at using DINOv2 for semantic segmentation**

**Approach used:**

- Dataset with 104 classes
- DINOv2 as backbone + linear classifier on top implemented as a 1x1 convolutional filter
- Trained only for 10 epochs to test the effectiveness of the model

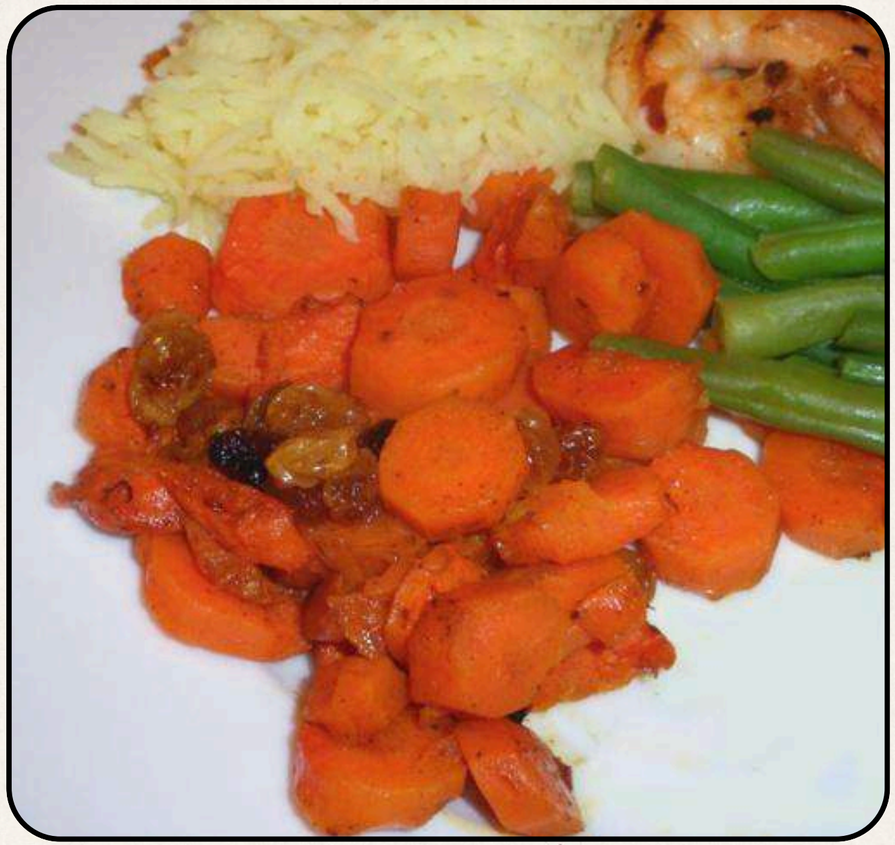
**Results achieved:**

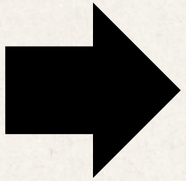
- Cross-Entropy Loss: 0.9028
- Mean IoU = 0.5120

**Problems:**

- Struggles to segment the background

# FoodSeg



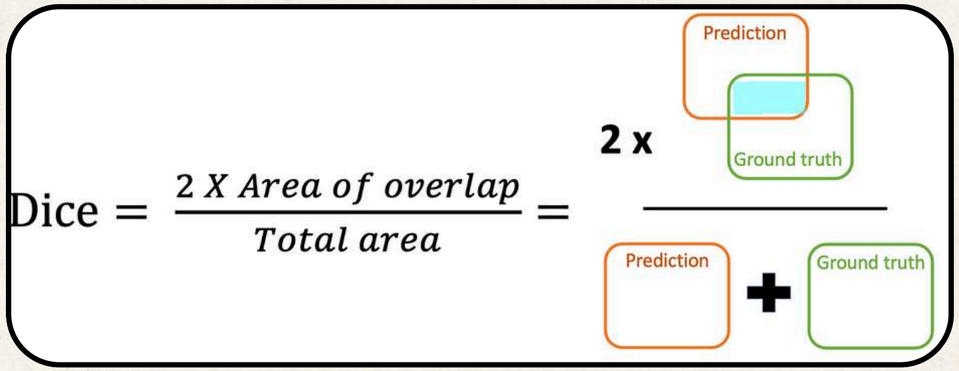


MONAI

**PyTorch-based, open-source framework for deep learning in healthcare imaging**

- Flexible pre-processing for multi-dimensional medical imaging data
- Domain-specific implementations for networks, losses and evaluation metrics

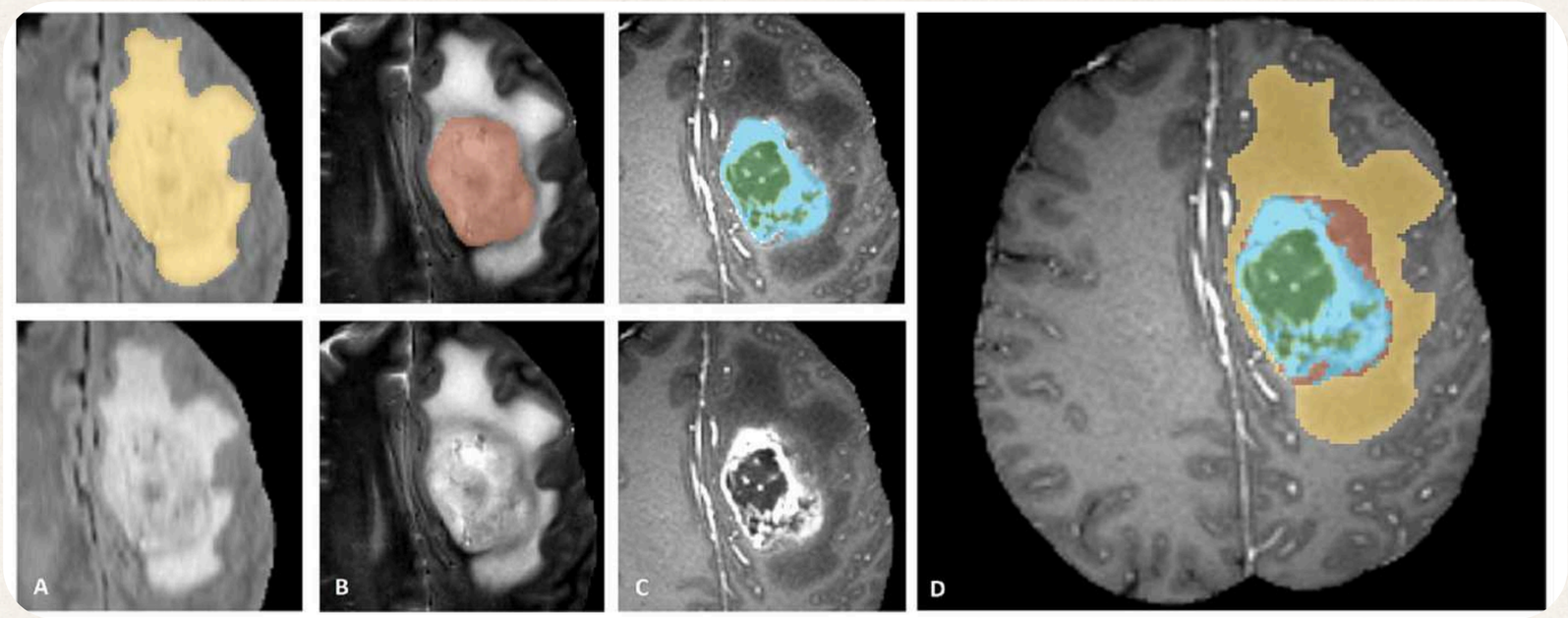
**Dice Loss is complementary to the Dice score**



# BraTS

**Brain Tumor Segmentation**

- Modality: Multimodal multisite MRI data (FLAIR, T1w, T1gd,T2w)
- Size: 750 4D volumes (484 Training, 110 Validation,  156 Test)
- Source: BRATS 2016 and 2017 datasets.



# BraTS

## Approach with three modalities

- BraTS dataset provides four MRI modalities: FLAIR, T1w, T1gd, and T2w.
- The task is to segment Tumor Core (TC), Whole Tumor (WT) and Enhancing Tumor (ET).
- DINOv2 requires 3-channel RGB input: only three modalities are selected to match this constraint.
- The T1w modality was excluded since there are two T1 variants and T1gd is better at distinguishing enhancing tumor.
- The remaining modalities (FLAIR, T1gd, T2w) were mapped to the three input channels and fed to the backbone as a pseudo-RGB image.
- Linear classifier on top for segmentation

# BraTS

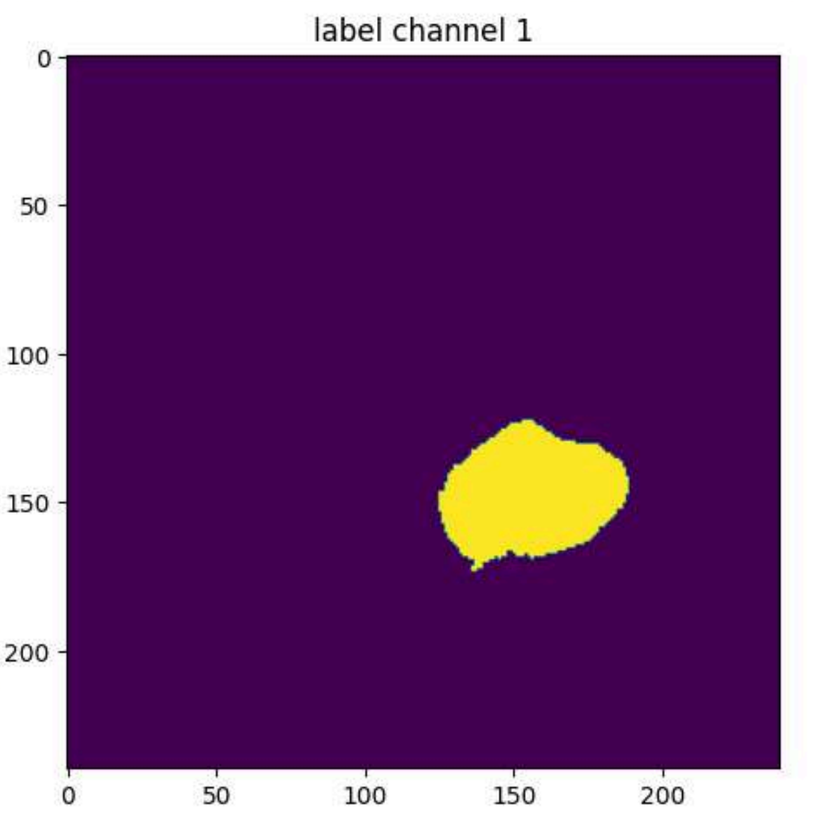
**Data preprocessing & transformation**

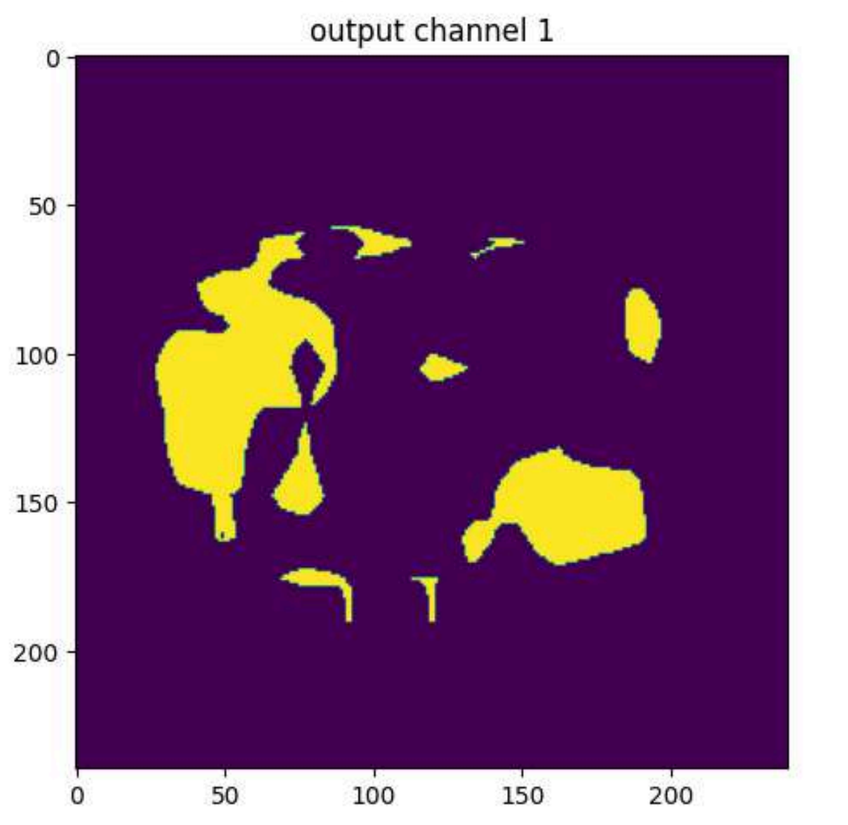
- Images and labels were preprocessed to ensure consistent orientation, resolution, and spatial size.
- Data augmentation techniques such as random flipping and intensity adjustments were applied to improve model robustness.
- For training and validation, the 3D images were transformed into 2D slices resembling RGB images, combining the selected modalities and normalizing them.
- Ground truth labels were converted into a multi-channel format that highlights the main tumor regions (tumor core, whole tumor, enhancing tumor).

# Results

**Best result on a single image from the validation set during 10 epochs training:**

- Dice score: 0.1280



# BraTS

## Single-Modality Approach (FLAIR only)

- A second approach was explored by using only the FLAIR modality, inspired by the methodology presented in the paper “Do Vision Foundation Models Enhance Domain Generalization in Medical Image Segmentation?”.

- The same preprocessing pipeline was applied as in the previous approach, but the input was reduced to a single-channel (FLAIR) image, which was repeated across the three RGB channels to match the model’s input format.

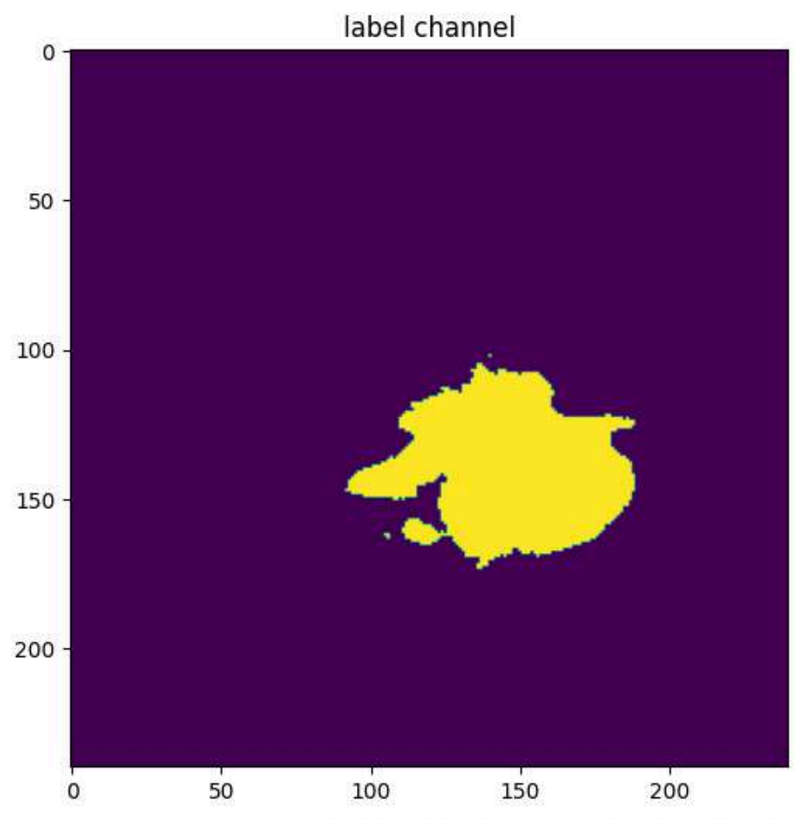
- To further simplify the task, the label space was reduced to two classes: background and lesion, instead of the original three tumor regions.

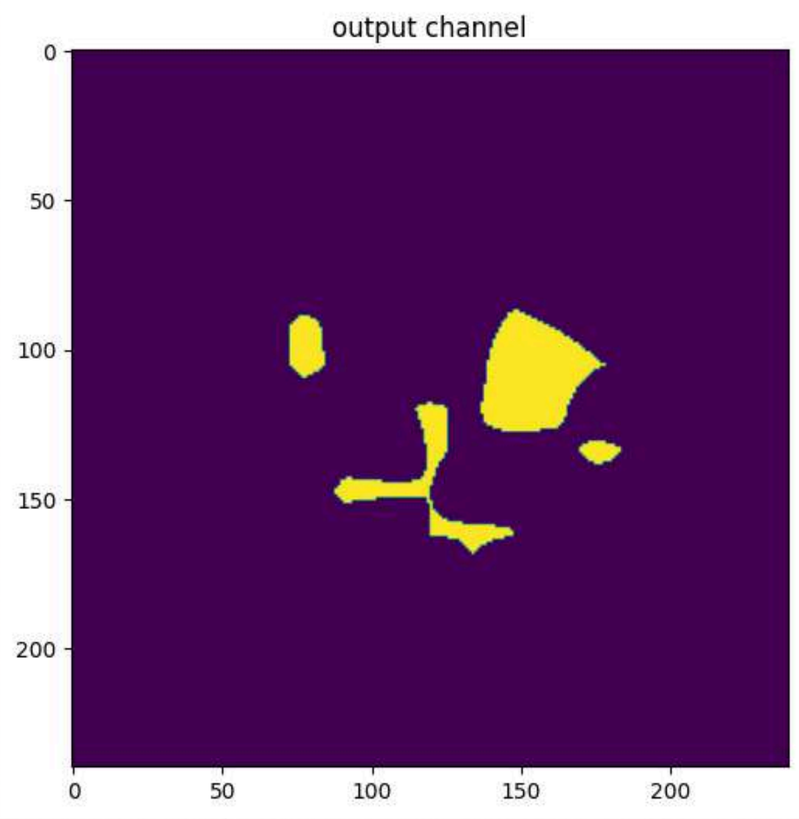
- This simplification was motivated by the difficulty DINOv2 showed when handling the full 3-modality setting, aiming to make the segmentation task easier and potentially improve model performance.

# Results

**Best result on a single image from the validation set during 10 epochs training:**

- Dice score: 0.2392



# Key Challenges

- Long training times
- Large domain gap
- Adapting the inputs to the 3-channel format required by Dinov2 implies loss of information
- Operating on 2D slices leads to loss of 3D context
- Dice score is very sensitive to small structures (like enhancing tumor core), so a small prediction error can cause a large drop in Dice score

# Fine-Tuning

**Rein-LoRA method**

- Parameter-Efficient Fine-Tuning (PEFT) method designed to improve model generalization while keeping the number of trainable parameters low.
- It combines LoRA (Low-Rank Adaptation) with re-injection of learned representations into the model, allowing better adaptation to downstream tasks.
- The method injects task-specific features at multiple layers, enhancing the model's capacity to specialize without full fine-tuning.
- This leads to a balance between model adaptability and parameter efficiency, making it suitable for large vision foundation models.

# Fine-Tuning

## Rein-LoRA Fine-Tuning Attempt

- Based on the results presented in the paper “Do Vision Foundation Models Enhance Domain Generalization in Medical Image Segmentation?”, where Rein-LoRA shows competitive fine-tuning performance, this method was also attempted on our model to potentially improve segmentation accuracy.
- However, several challenges were encountered when using the official repository.
- The environment setup was complex due to the large number of library dependencies and version conflicts, making reproducibility difficult.
- Moreover, the data loading and preprocessing required by the repository significantly differed from our pipeline, causing further integration issues.
- These obstacles limited the feasibility of fully applying Rein-LoRA fine-tuning in our experiments.

# Final considerations

- DINOv2 shows promise as a feature extractor
- However, adapting it to medical imaging  presents integration and data representation challenges
- The loss of 3D context and having to work on 2D slices may be too limiting for multi-class segmentation tasks

# Thank you !