Efficient adaptation of Foundation Models for Visual Grounding Remote Sensing

Ali J. Ghandour, Hasan Moughnieh, Mohammad Hasan Zahweh, Hasan Nasrallah, Mustafa Shukor, Cristiano Nattero, and Paolo Campanella

www.geogroup.ai

EGU General Assembly 2024

Overview and Motivation

Natural Image
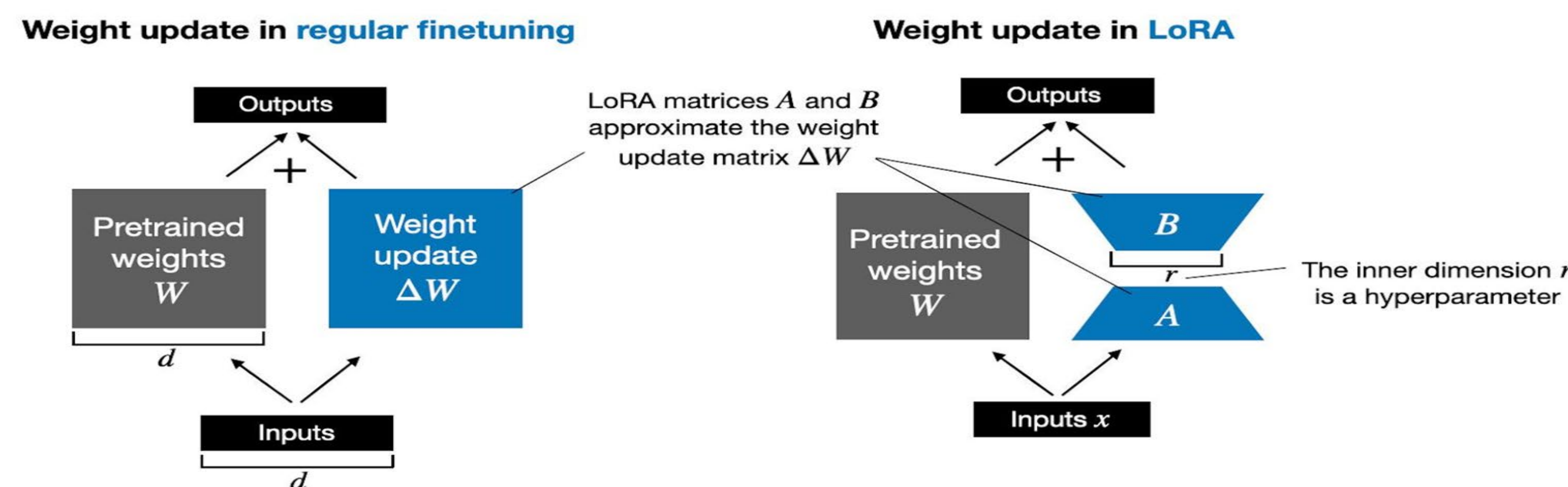**Expression:** lady right

Remote Sensing Image
Expression: Find a ground track field, located approximately 295 meters southeast of a baseball field. The baseball field is next to a road.

- **Visual Grounding:** predicting the image region corresponding to a text.
- **Objective:** Parameter-efficient adaptation of existing foundation models for remote sensing visual grounding.
- **Motivation**: Limited data and large-scale models for remote sensing applications. Fine Tuning large models is costly.
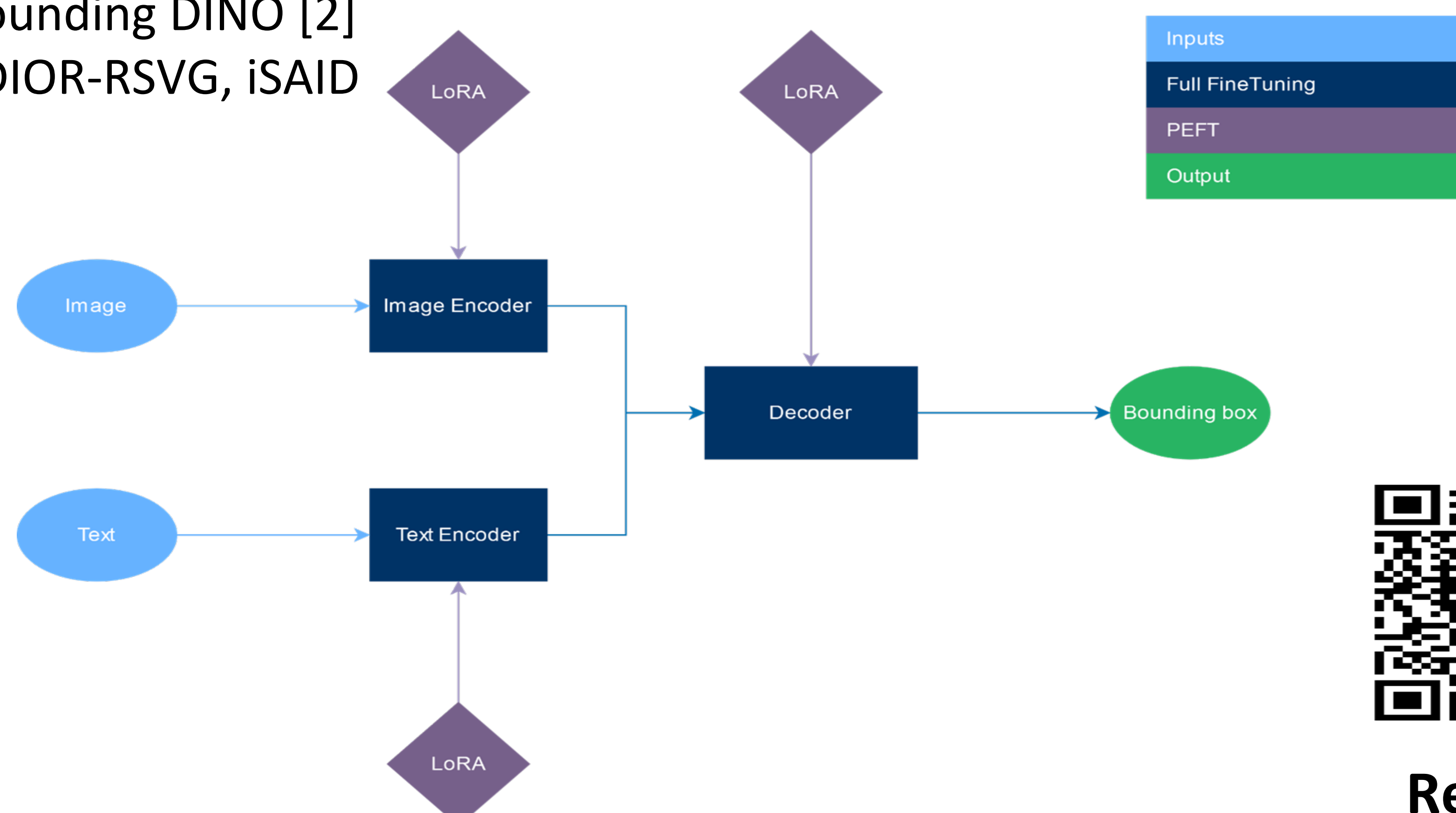
Efficient Adaptation For Remote Sensing

- **Parameter-Efficient Fine-tuning(PEFT):** adapting models by training only few number of parameters.
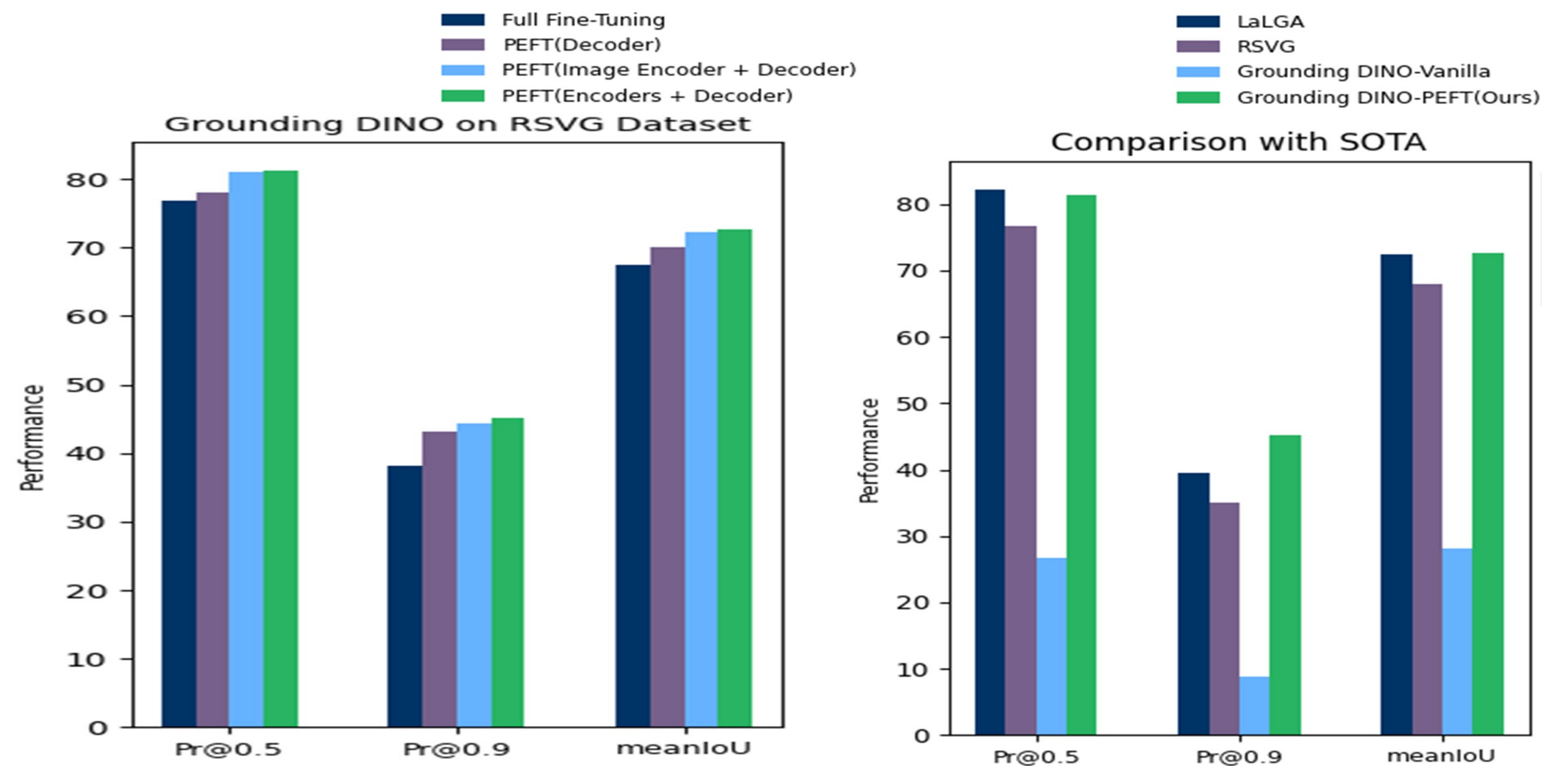- **PEFT Technique:** Low Rank Adaptation(LoRA) [1]

Weight update in **regular finetuning**

Weight update in **LoRA**

LoRA matrices A and B approximate the weight update matrix ΔW

The inner dimension r is a hyperparameter

Experiments

- **Model**: Grounding DINO [2]
- **Datasets**: DIOR-RSVG, iSAID

References

❖ Experimental Results [4][5]

Legend: Full Fine-Tuning, PEFT(Decoder), PEFT(Image Encoder + Decoder), PEFT(Encoders + Decoder)

Grounding DINO on RSVG Dataset

Legend: LaLGA, RSVG, Grounding DINO-Vanilla, Grounding DINO-PEFT(Ours)

Comparison with SOTA

❖ Qualitative Results

DIOR-RSVG test set [6]

A large storage tank

A dam in the middle

The basketball court in the middle

A dam on the left

ISAID Dataset [7]

Airplane at the bottom right

Tennis court on the upper left