

Generalizing Herbarium Specimen Label Understanding Under Limited Data Diversity

Weiwei Liu ^{1,2}, Stefano Martellos ²

¹ Department of Mathematics, Informatics and Geosciences, University of Trieste

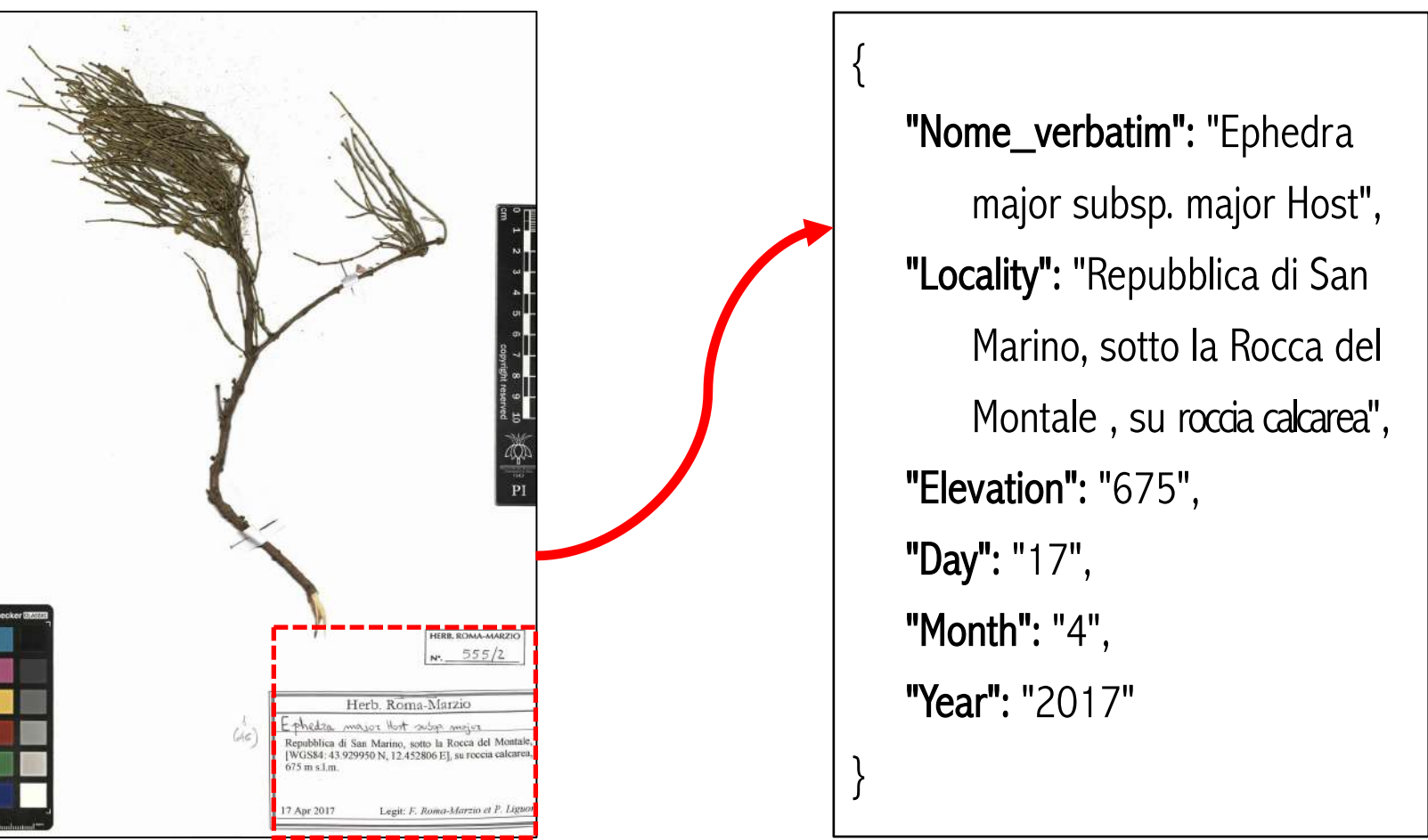
² Department of Life Sciences, University of Trieste



UNIVERSITÀ
DEGLI STUDI
DI TRIESTE

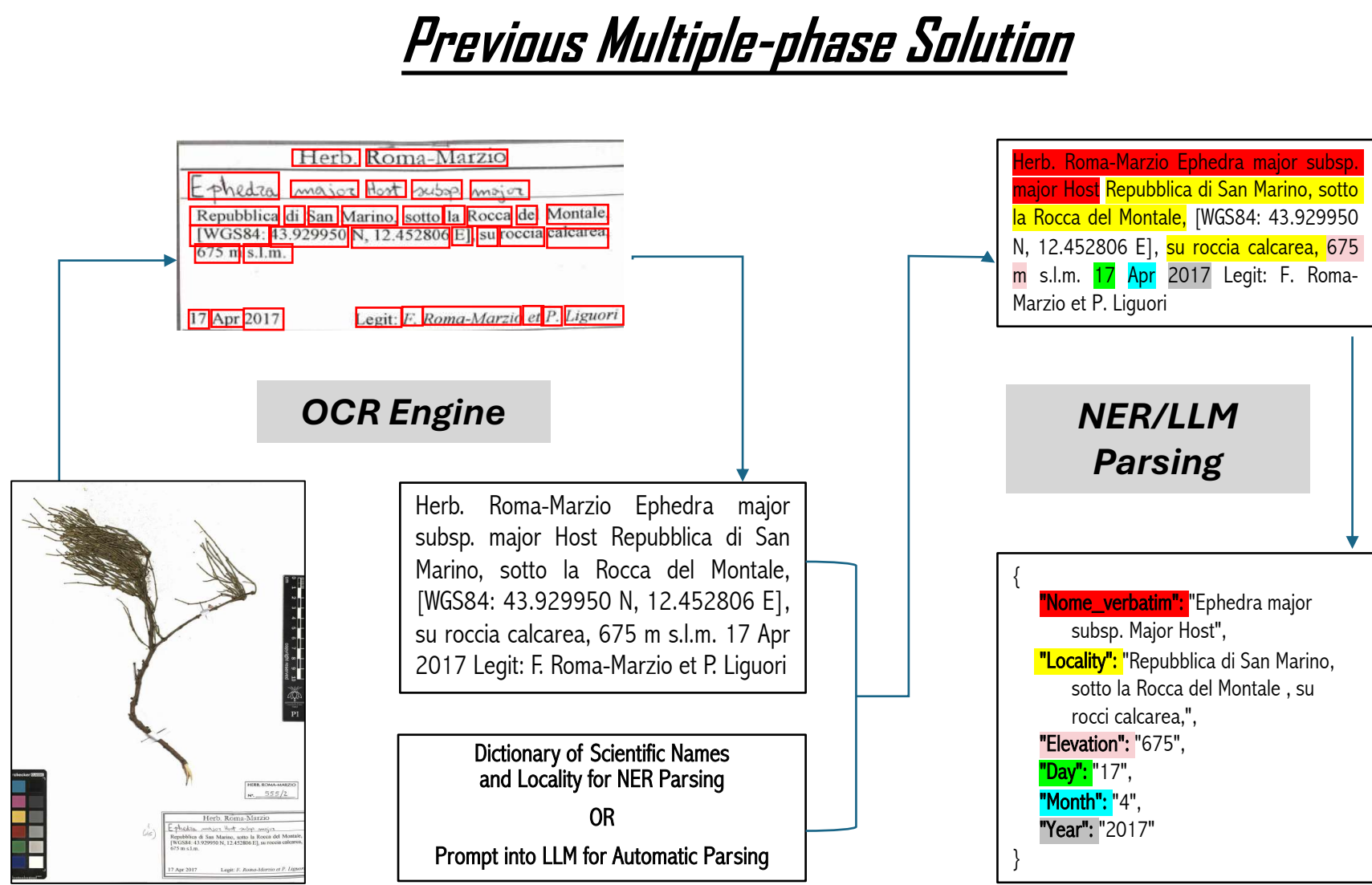


Work in Progress – ECCV 2026 CVNH Non-Archival Poster



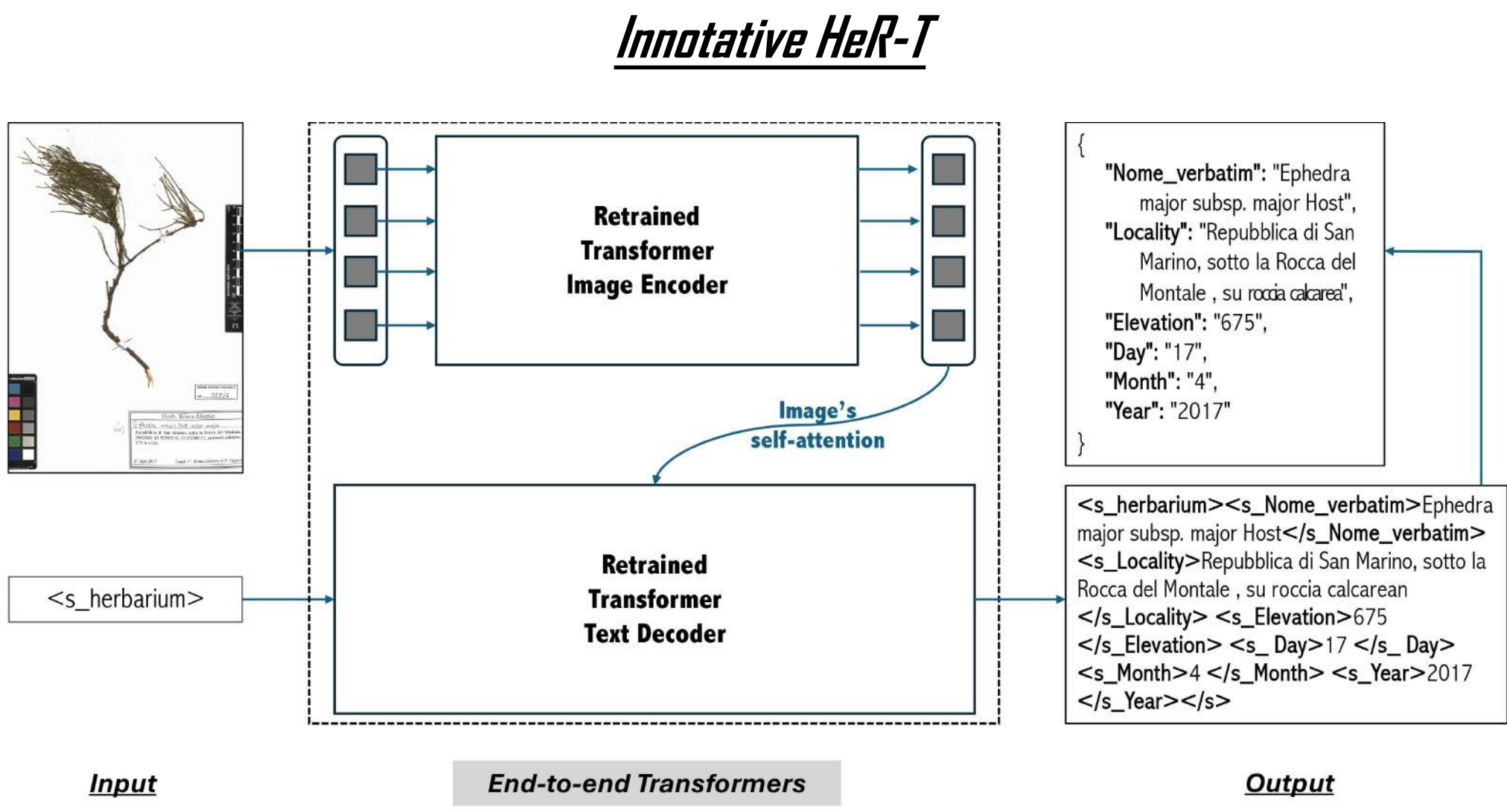
Task Definition

- Multilingual recognition of prints and handwritings
- Parsing of the unstructured data
- End-to-end solutions integrating open sources
- Generalized transcription across different herbarium



From Multiple-phase Pipeline To End-to-end One-pass Extraction

- Problem 1: Multiple passes propagate errors from OCR results
- Problem 2: OCR Engine relies on extra preprocessing, including segmentation, cropping, etc.
- Problem 3: LLM-based parsing provides limited interpretability of the generated outputs
- Achievement: Published the 1st end-to-end one-pass extraction solution for herbarium specimen labels: Herbarium specimen Recognition Transformer (HeR-T)

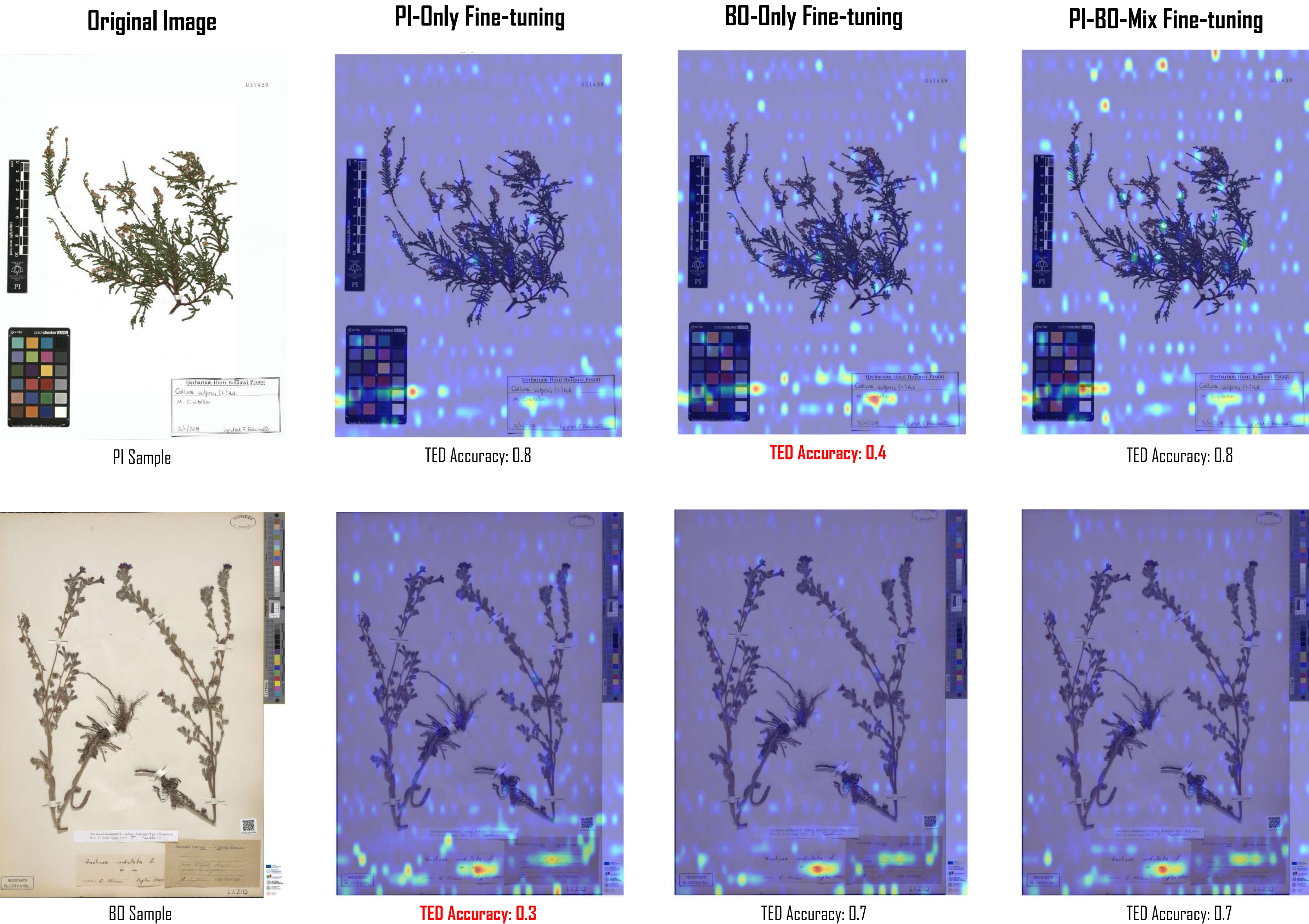


Number of Training Images			Average Testing Tree Editing Distance (TED) Accuracy		
PI	BO	Total	PI	BO	Mean
32.165	0	32.165	0.788	0.481	0.635
0	32.166	32.166	0.589	0.864	0.726
16.084	16.083	32.167	0.754	0.712	0.733
32.165	32.166	64.331	0.765	0.802	0.784

PI: Herbarium in University of Pisa
BO: Herbarium in University of Bologna

Generalization Failure

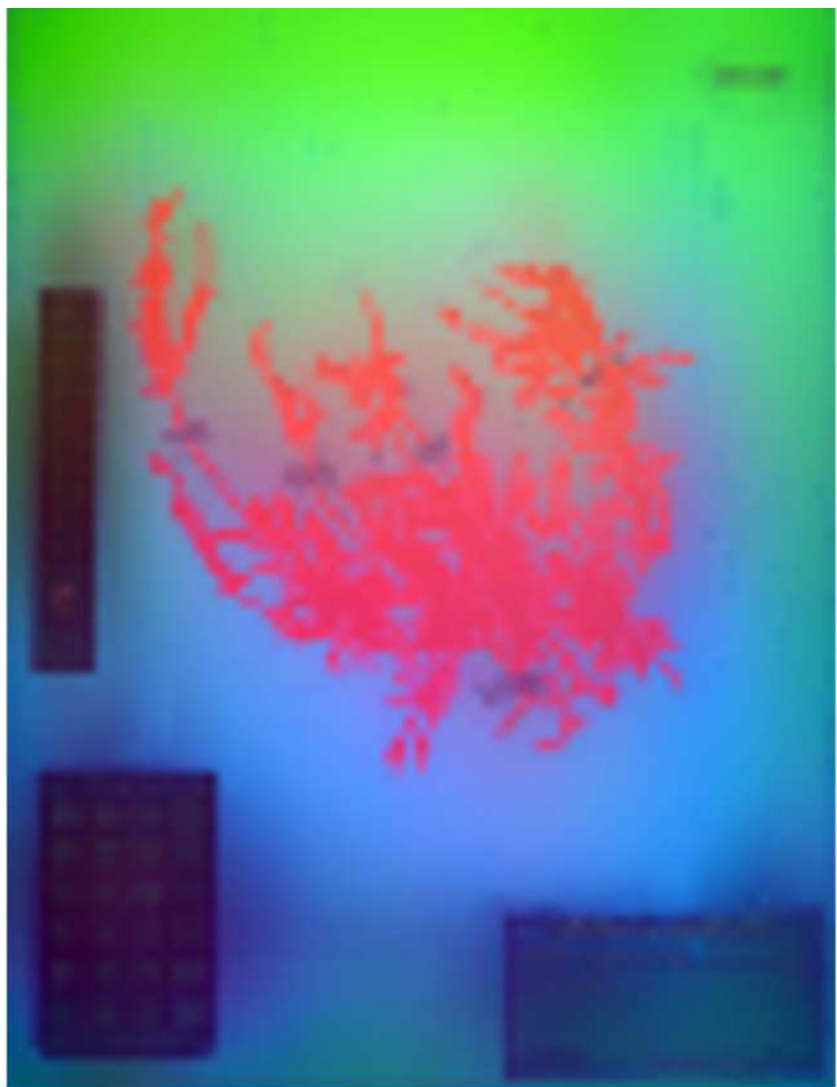
- Performance in one herbarium does not transfer reliably to another
- Mixing data to increase data diversity can improve the generalization of HeR-T
- Increasing mixed-domain training data narrows the gap toward the best within-domain performance
- Only 2 herbarium with the same language;
What about the other 3500 herbarium using more languages?



Heatmap of Cross-attentions

- Cross-attention patterns remain broadly consistent across the three HeR-T variants
- Minor additional noise can be found in the worst cases
- Insights: Do generalizing visual representations or reducing noise improve transcription across different herbaria?

PCR Analysis of DINOv3 Visual Representation



- Swap the Swin-T encoder in the HeR-T with DINOv3
- Add an intermediate layer as a bridge between DINOv3 and the BART decoder in HeR-T
- Failure in transcription:
 - Freeze the encoder and decoder, and train the bridge from scratch
 - Freeze the encoder, and train the bridge and decoder from scratch
 - Unfreeze the last layers of the encoder, and repeat the training of a. and b.
- Discriminative features of the contents on labels are missing
- Global structures are distinct

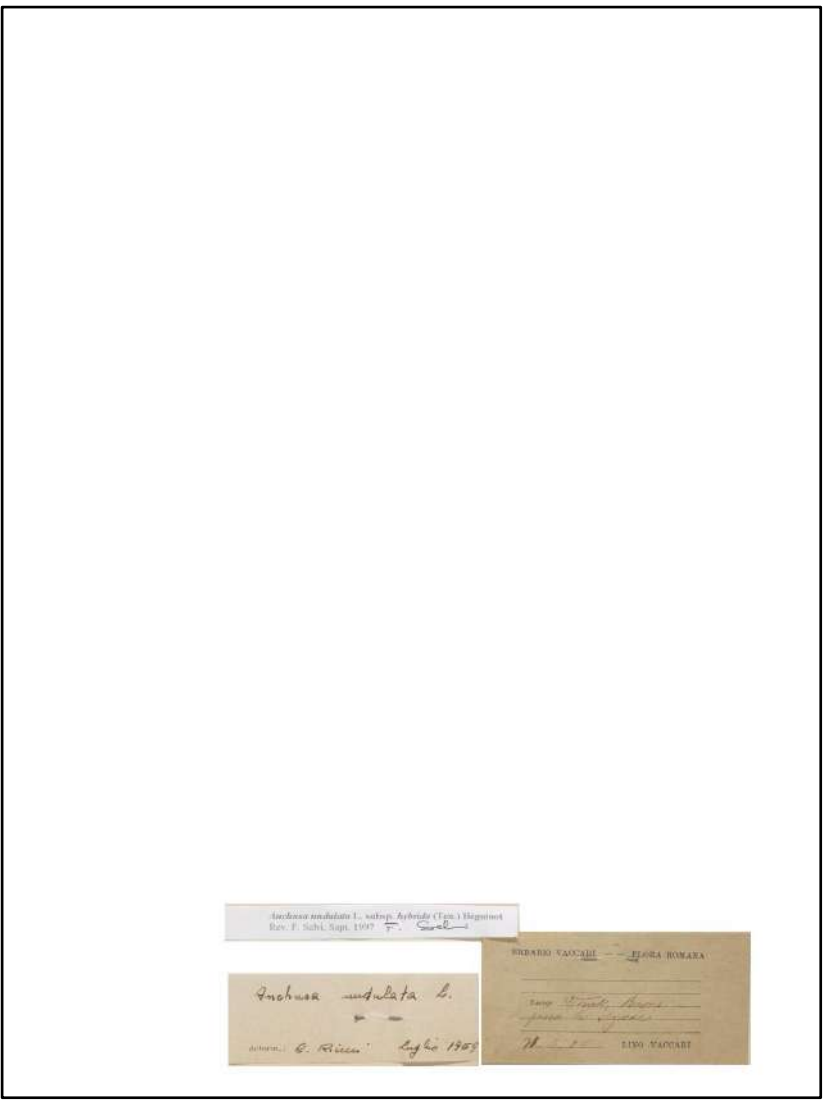
General-purpose visual encoder failed the task but provided a new insight:

Generalized masking of non-label areas in specimens



WE ARE WORKING ON:
Curated Auxiliary Supervision by Masking Background and All Areas Other Than Labels

Visual embeddings from the DINOv3 encoder can be helpful for generalized masking of non-label areas in specimens



BUT

Does masking these areas improve the generalization of the specimen-label transcription across different herbaria?

WE ARE ON THE WAY!

General-Purpose Visual Representation by Swapping the Swin-T Encoder with DINOv3

Contact Me:

Weiwei Liu

PhD Student, University of Trieste
weiwei.liu@phd.units.it



Contact card



LinkedIn



GitHub