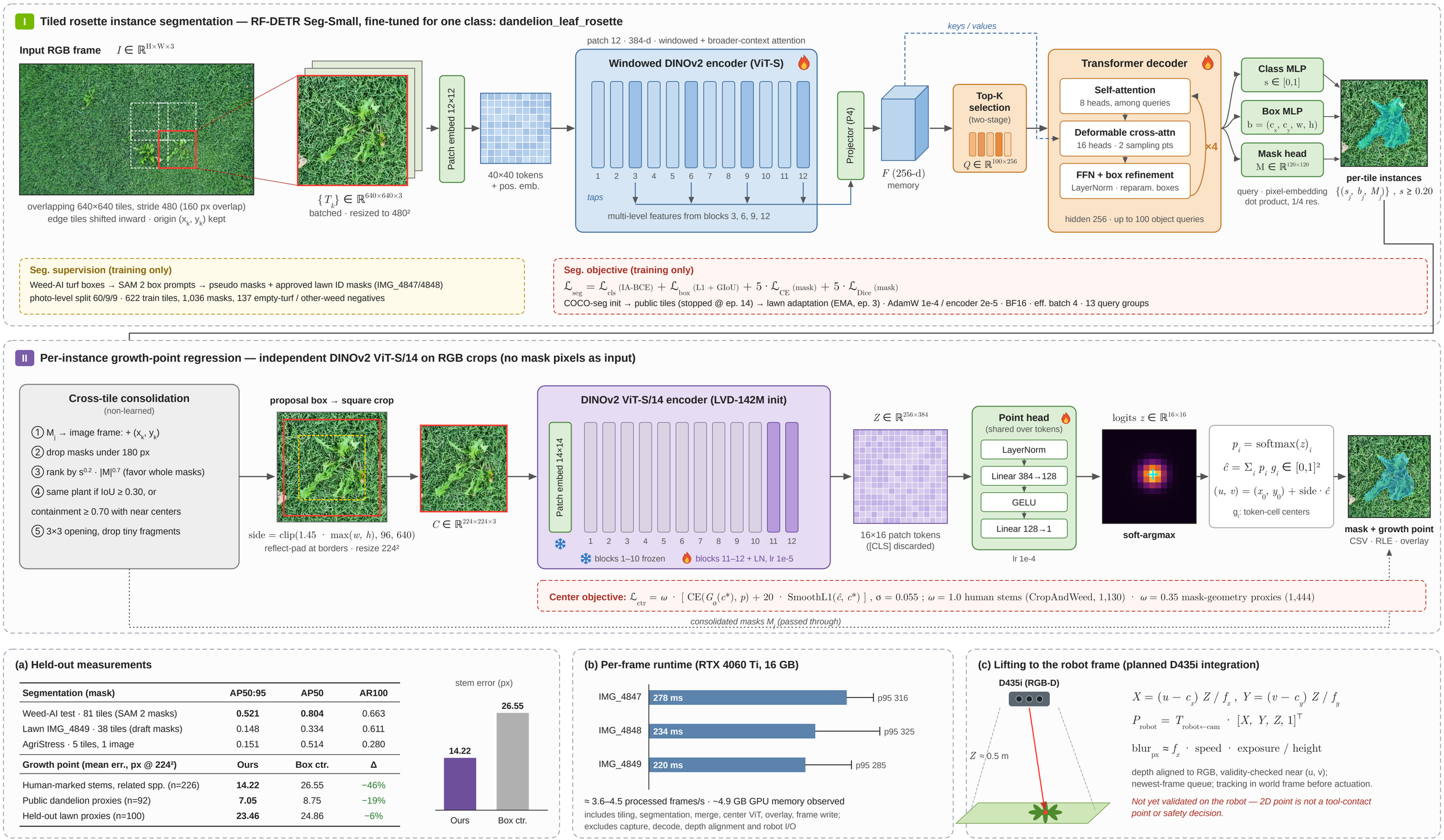


(a) Held-out measurements

Segmentation (mask)	AP50:95	AP50	AR100
Weed-AI test · 81 tiles (SAM 2 masks)	**0.521**	**0.804**	0.663
Lawn IMG_4849 · 38 tiles (draft masks)	0.148	0.334	0.611
AgriStress · 5 tiles, 1 image	0.151	0.514	0.280
Growth point (mean err., px @ 224²)	**Ours**	**Box ctr.**	**Δ**
Human-marked stems, related spp. (n=226)	**14.22**	26.55	−46%
Public dandelion proxies (n=92)	**7.05**	8.75	−19%
Held-out lawn proxies (n=100)	**23.46**	24.86	−6%

(b) Per-frame runtime (RTX 4060 Ti, 16 GB)

≈ 3.6–4.5 processed frames/s · ~4.9 GB GPU memory observed
includes tiling, segmentation, merge, center ViT, overlay, frame write; excludes capture, decode, depth alignment and robot I/O

(c) Lifting to the robot frame (planned D435i integration)

$X = (u - c_x)\, Z / f_x$, $Y = (v - c_y)\, Z / f_y$

$P_{robot} = T_{robot \leftarrow cam} \cdot [X, Y, Z, 1]^\top$

$\text{blur}_{px} \approx f_x \cdot \text{speed} \cdot \text{exposure} / \text{height}$

depth aligned to RGB, validity-checked near (u, v); newest-frame queue; tracking in world frame before actuation.

Not yet validated on the robot — 2D point is not a tool-contact point or safety decision.

segmentation ViT · DETR decoder / queries · conv / MLP heads · center ViT · non-learned op · fine-tuned · frozen · inference data flow · training only · encoder memory · tile / crop shown downstream

Figure 2. Overview of the two-network dandelion rosette perception pipeline. Stage I: an overhead RGB frame is cut into overlapping 640² tiles; a fine-tuned RF-DETR Seg-Small (windowed DINOv2-style ViT-S, patch 12, taps at blocks 3/6/9/12, single-scale P4 projector, two-stage top-K queries, 4-layer deformable decoder) predicts a score, box and quarter-resolution mask per query. Tile proposals are shifted to image coordinates, de-duplicated and cleaned. **Stage II**: each surviving proposal box defines a 1.45× square RGB crop that a separate DINOv2 ViT-S/14 (blocks 1–10 frozen) maps to a 16×16 token heatmap; a soft-argmax yields a continuous growth point (u, v), learned independently of the mask centroid. Example: held-out Weed-AI tile. The heatmap is a schematic of the soft-argmax target shape, not a recorded activation. Labels are pseudo / draft masks and mostly geometric center proxies, so the measurements in (a) reflect agreement with those references rather than verified botanical accuracy.