Supplementary Table S1. Per-class precision, recall, and F1 with 95% credible intervals (CrI) for the selected models on the held-out test sets (Task1, RT-DETR-L; Task2, RT-DETR-L; Task3, YOLO11-L-seg). Counts and point estimates correspond to a fixed confidence of 0.25 and an IoU matching threshold of 0.5 (Tables 7, 9, and 11). For Task3, box and mask outputs yielded identical counts and therefore identical values.

Task	Class	TP / FP / FN	Precision (95% CrI)	Recall (95% CrI)	F1 (95% CrI)
T1	Holding pipette tip	263 / 0 / 0	1.000 (0.990–1.000)	1.000 (0.991–1.000)	1.000 (0.993–1.000)
	Needle tip	257 / 6 / 6	0.977 (0.954–0.990)	0.977 (0.953–0.990)	0.977 (0.960–0.987)
	Oocyte 40×	263 / 0 / 3	1.000 (0.990–1.000)	0.989 (0.970–0.997)	0.994 (0.983–0.998)
	Oocyte 4×	76 / 0 / 0	1.000 (0.968–1.000)	1.000 (0.968–1.000)	1.000 (0.976–1.000)
	Macro-average	—	**0.994 (0.983–0.996)**	**0.991 (0.980–0.994)**	**0.993 (0.984–0.994)**
T2	Meniscus	62 / 1 / 0	0.984 (0.928–0.998)	1.000 (0.961–1.000)	0.992 (0.956–0.998)
	Needle tip	374 / 8 / 10	0.979 (0.961–0.990)	0.974 (0.954–0.987)	0.977 (0.963–0.985)
	Sperm in needle	268 / 68 / 23	0.798 (0.752–0.838)	0.921 (0.886–0.948)	0.855 (0.823–0.882)
	Macro-average	—	**0.920 (0.895–0.935)**	**0.965 (0.946–0.974)**	**0.941 (0.924–0.949)**
T3	Head	424 / 2 / 7	0.995 (0.985–0.999)	0.984 (0.968–0.993)	0.990 (0.980–0.994)
	Tail	424 / 1 / 8	0.998 (0.989–1.000)	0.981 (0.965–0.991)	0.990 (0.980–0.994)
	Macro-average	—	**0.996 (0.990–0.999)**	**0.983 (0.972–0.989)**	**0.990 (0.983–0.993)**

To estimate uncertainty, 95% Bayesian credible intervals for each metric were derived via Monte Carlo sampling (200,000 draws) from the posterior Dirichlet distribution, parameterized by a symmetric Jeffreys prior ($\alpha = 0.5$) over the observed true-positive (TP), false-positive (FP), and false-negative (FN) counts. For precision and recall, this formulation is analytically equivalent to the equal-tailed Jeffreys prior interval for a binomial proportion. Intervals for macro-averages were computed jointly from the empirical distribution of the same draws. The simulation was executed natively within Microsoft Excel (Office 365, Version 2610). Importantly, these intervals treat instances as identically and independently distributed; because consecutive frames and multi-frame sequences from the same patient exhibit inherent clinical correlation, the assumption of independence likely understates the true uncertainty.