子計畫三：3M Maker: 使用多模式、多任務、多事例學習開發可信賴多媒體處理技術

- The proposed Generalist YOLO achieves excellent results on multi-modal and multiple vision and language tasks, while maintaining higher accuracy than existing state-of-the-art generalist models with 135 times faster inference speed.

- We proposed Relaxed Optimizer and several efficient modules to design the Generalist YOLO.

Encoder	#param	FLOPs	AP^{box}	AP^{mask}	$mIoU^{sem}$	$mIoU^{stuff}$	PQ^{pano}	$B@4^{cap}$
GELAN	47.5M	~225.7G	50.3	41.8	48.7	56.8	40.4	36.8
+ Relaxed Optimizer	47.5M	~225.7G	51.9	42.6	50.0	56.5	41.7	38.8
+ PGI [35]	47.5M	~225.7G	52.1	42.6	50.1	56.4	42.1	39.1
+ Shallow Decoder	45.5M	~222.7G	52.2	42.9	49.4	56.8	42.2	**39.4**
+ Semantic Feature	37.3M	~195.2G	52.3	42.9	51.5	59.1	44.0	38.6
+ Distinctive Feature	37.3M	~195.2G	52.3	42.9	51.7	**59.4**	**44.3**	38.5
+ Light Head	**32.5M**	**~122.2G**	**52.4**	**43.0**	**52.0**	59.2	44.2	38.7

- The proposed Generalist YOLO do not need any pretrained model, and it gets prediction of all tasks in one-shot inference.

