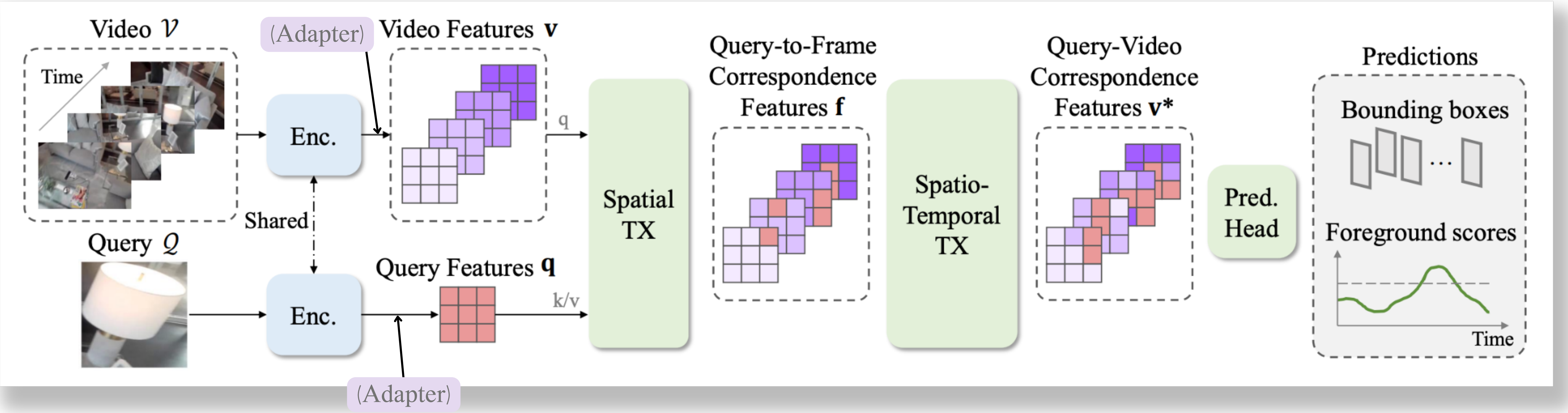
VISUAL QUERIES 2D LOCALIZATION

Team: DEEEEEEEEP
R12922169 廖哲賢
R12922164 陳祈安
R12942143 林翰莘
R12922172 郭旻展

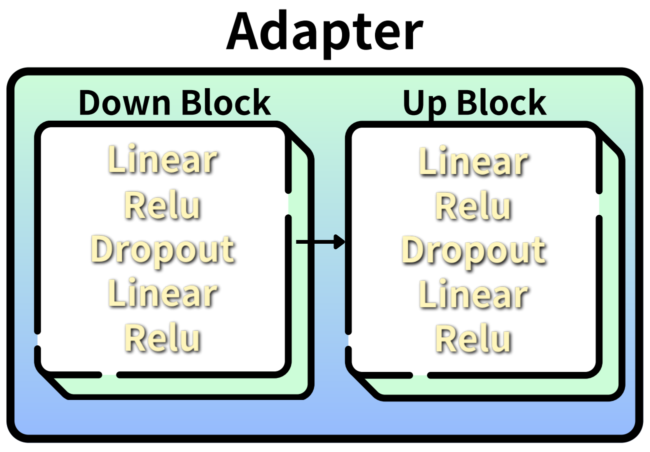
MODEL ARCHITECTURE

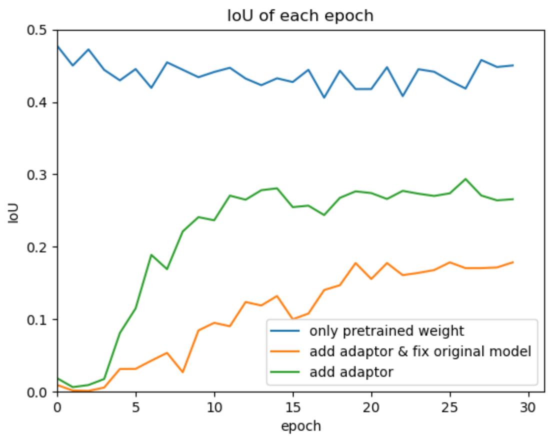


ABLATION STUDY

1.Parameter-Efficient Fine-tuning

	Training Time	*number of parameters*	*IoU at 30 epochs*
original model	*5hr 07m*	*42,678,844*	*0.45*
train adapter only	*5hr 30m*	*2,626,048*	*0.17*
train adapter & model	*5hr 40m*	*45,304,892*	*0.26*





2. Focal Loss vs. HNM

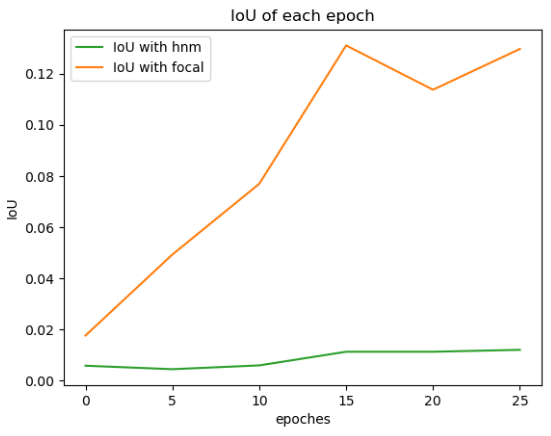
- Focal Loss

By increasing the loss weight for hard negative examples, the model is encouraged to focus more on challenging instances during training.

$$FL(p_t) = -\alpha(1 - p_t)^r \log(p_t)$$

- Hard Negative Mining

Hard negative mining involves collecting negative examples that are more challenging for the model to distinguish, and then using them to further train the model.

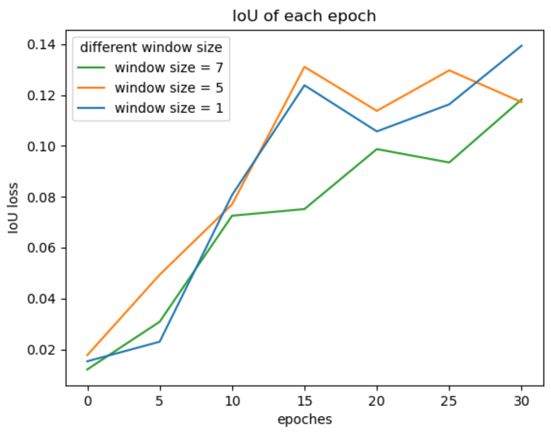


3. Windows size

- Focal Loss

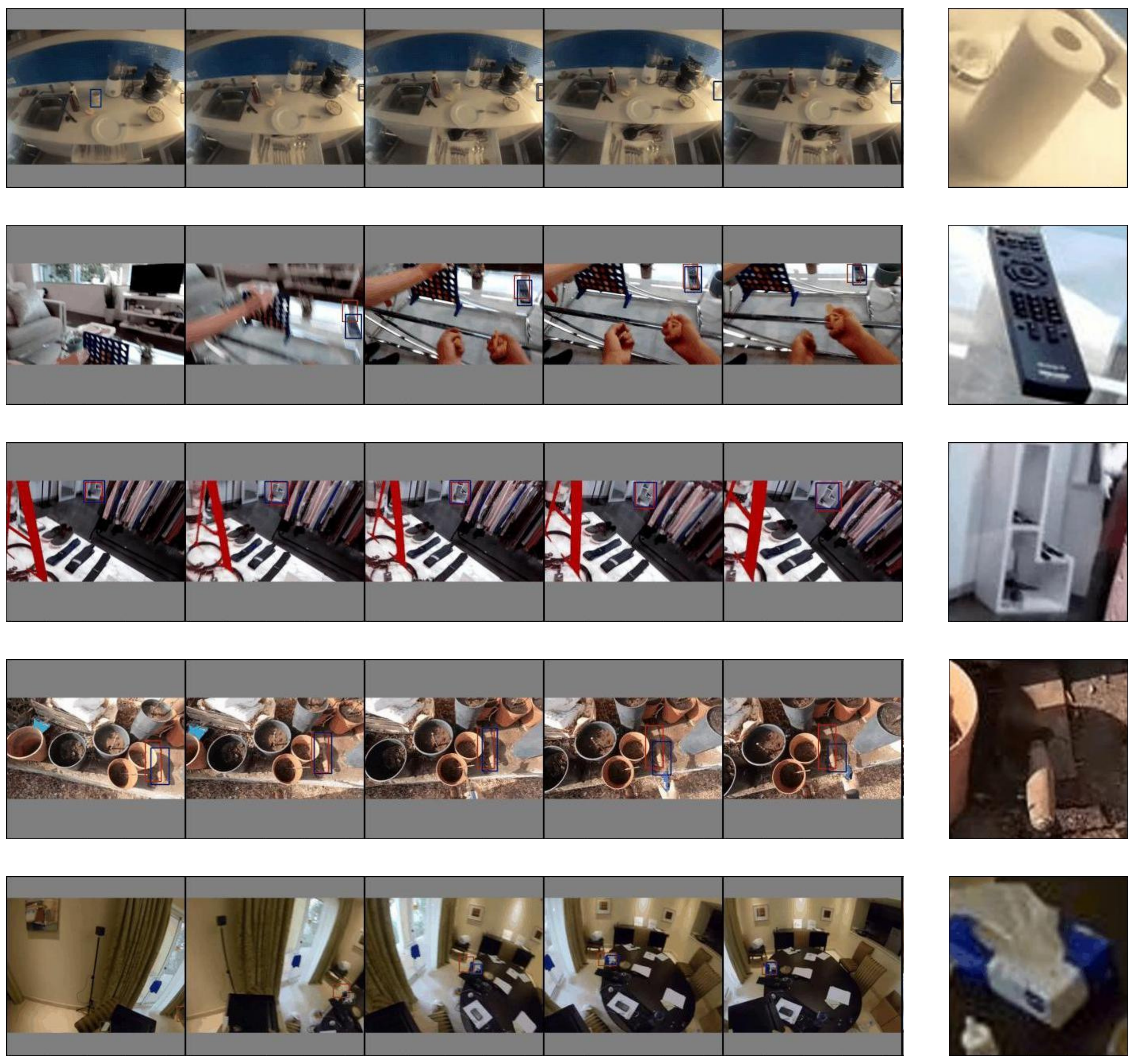
Learning rate: 0.0001, Schedular warmup iter: 1000, Total iteration: 60000, Batch size: 2, Data argument : query image random flip and random crop, Without pre-trained weight

	win1	*win5*	*win7*
iou	*0.139*	*0.131*	*0.118*
prob. acc	*0.507*	*0.440*	*0.480*



RESULT

stAP on Test set: 0.2897



Reference: https://github.com/hwjiang1510/VQLoC