Just One Glance Detect Kinetic Object – VQ2D Task

Team : JOKOandherfriends
Ling Chun Chen R12942078 Yu Hong Zhou R12921053 Wen Tzu Chang R12921A11 Shun Gui Wang R11921099

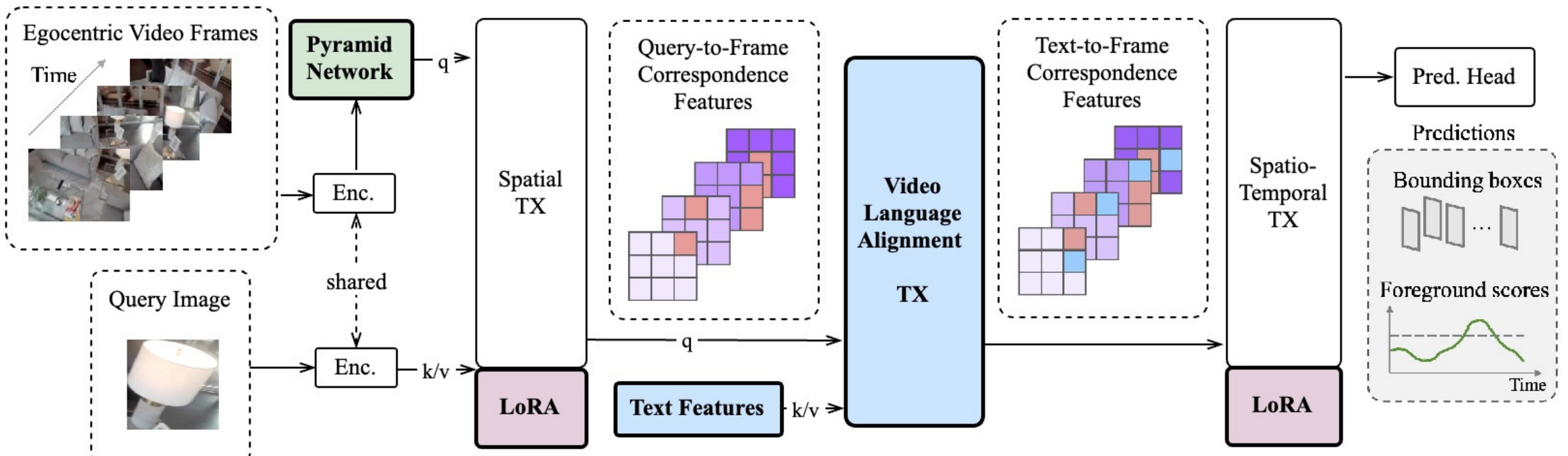


Fig1: Just One Glance Detect Kinetic Object (JOKO) Model.
An end-to-end framework that performs spatio-temporal searches[1] with text integration and image scale improvement.

SMALL OBJECT DETECTION

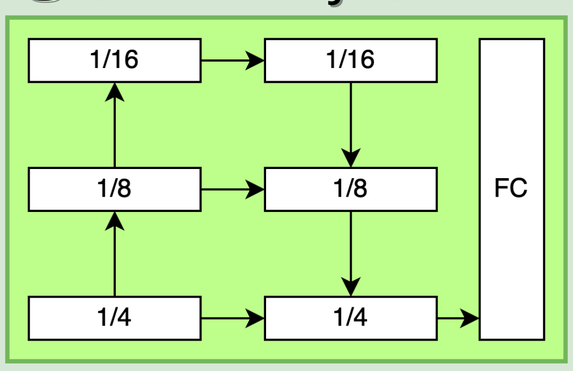


Fig 2 : FPN after ResNet50 pretrained on Ego4D full dataset

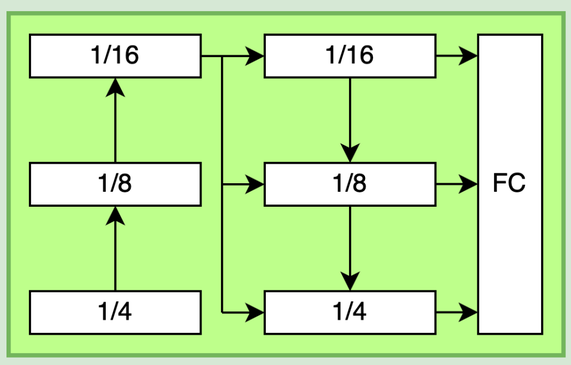
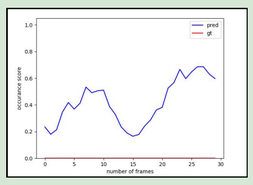
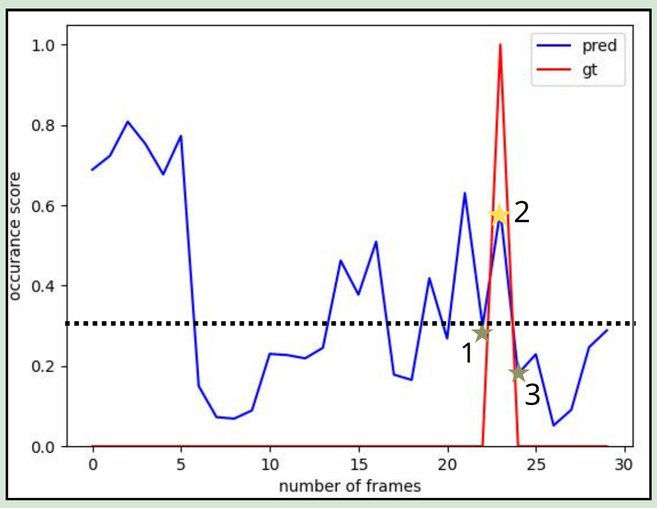


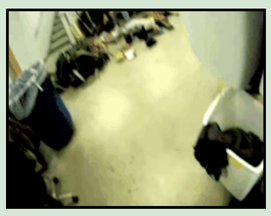
Fig 3 : Simple Feature Pyramid after ViT

Query Image



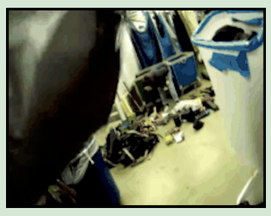
VQLoc can't catch the object.



1

GT: 0.0
Pred: 0.299

2

GT: 1.0
Pred: 0.587

3

GT: 0.0
Pred: 0.180

TEXT INTEGRATION - CLIP

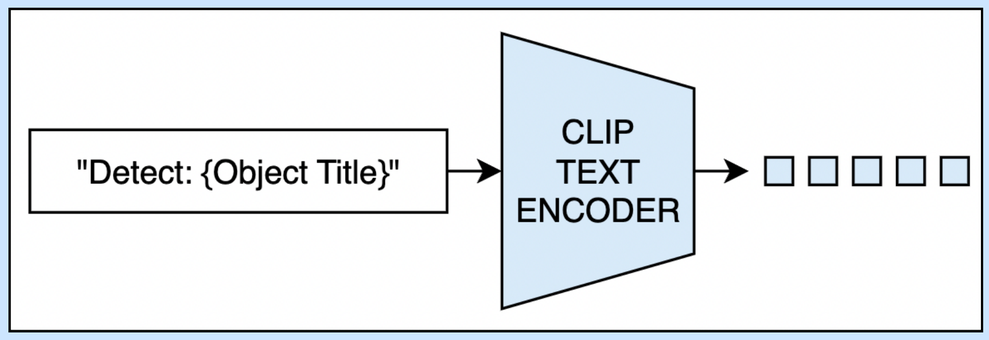


Fig 4 : Clip Implementation

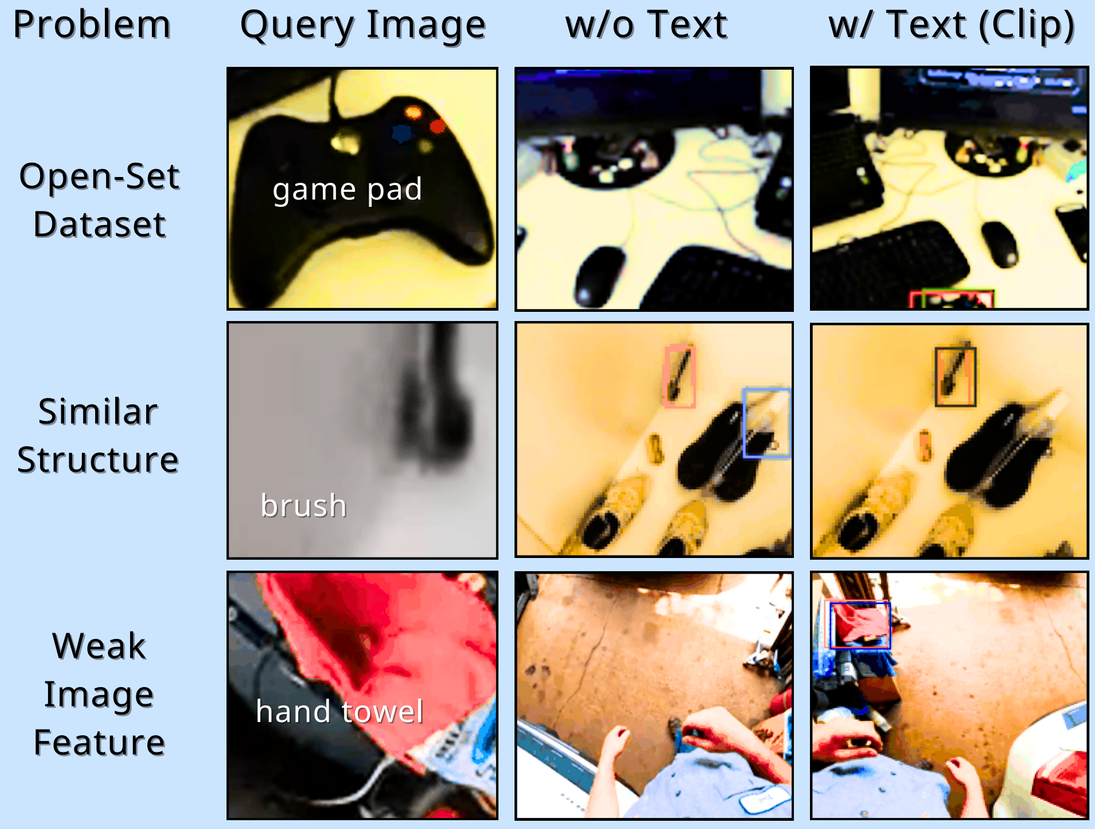


Fig 5 : Text Integration After and Before Comparison

SPEED ACCELERATION

Setting	w/o Lora	w/ Lora	Text	Dino w/SPF	Resnet50 w/FPN	WIOU
Training time (s)	29200	13491	23661	23870	9483	20079

LOSS FUNCTION - WIOU

$$\mathcal{L}_{WIoUv1} = \mathcal{R}_{WIoU}\mathcal{L}_{IoU}$$
$$\mathcal{R}_{WIoU} = \exp\left(\frac{(x-x_{gt})^2 + (y-y_{gt})^2}{(W_g^2 + H_g^2)^*}\right)$$

Equ 1: WIOU Our Loss Function

$$L_{img} = L_{bbox} + \lambda_p \cdot L_{prob} \quad L_{bbox}(\hat{\mathcal{B}}, \mathbf{b}) = \Sigma_{\hat{b}\in\hat{\mathcal{B}}^p} L_{reg}(\hat{\mathbf{b}}, \mathbf{b})$$
$$L_{reg}(\hat{\mathbf{b}}, \mathbf{b}) = ||\hat{\mathbf{b}}_c - \mathbf{b}_c|| + ||\hat{\mathbf{b}}_h - \mathbf{b}_h|| + ||\hat{\mathbf{b}}_w - \mathbf{b}_w|| + \lambda_{giou} \cdot L_{giou}(\hat{\mathbf{b}}, \mathbf{b})$$

Equ 2: VQLoC Loss Function

EXPERIMENT

Setting	stPA25
w/o pretrain	9.7941E-06
w/ Lora	0.3589464205
Text	0.3117680327
FPN	8.4839E-05
WIOU	3.7427E-06
PT on Complete Ego4D dataset	0.3923362006

Reference: [1] Jiang, Hanwen, Santhosh Kumar Ramakrishnan, and Kristen Grauman. "Single-Stage Visual Query Localization in Egocentric Videos." arXiv preprint arXiv:2306.09324 (2023). [2] Lin, Tsung-Yi, et al. "Feature pyramid networks for object detection." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017. [3] Li, Yanghao, et al. "Exploring plain vision transformer backbones for object detection." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022. [4] Tong, Zanjia, et al. "Wise-IoU: Bounding Box Regression Loss with Dynamic Focusing Mechanism." arXiv preprint arXiv:2301.10051 (2023). [5] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PMLR, 2021.