Human Interaction with Common Object Detection Project (HICO-DET)

Conduct an analysis to find optimal solution:

* **find pros and cons of different models**
* **research on models that are supposed to be used for the project**
* **Qwen multi modal language model (proposed model)**

|  |  |
| --- | --- |
| **Model (year)** | **HICO-DET and V-COCO mAP** |
| DiffHOI (2023) |  |
| HOI Transformer (CVPR’21) |  |
| QPIC (CVPR’21) |  |

**DiffHOI:** It is one of the models with the highest reported mAP scores and provided pre-trained weights to the public for further training to achieve a better model performance. However, it is extremely resource intensive (i.e. large GPU memory and computing processes) and complex since it has a very heavy pipeline. Hence, it has the best performance out of the models.

<https://github.com/IDEA-Research/DiffHOI>

**HOI Transformer:** The model leverages on End-to-End pipeline where identifying objects and inferring relationships are done in a single stage. Hence, the model has a better contextual reasoning when identifying relationships in a picture/video thus, a competitive mAP score as well. Since it processes the picture as a whole for contextual reasoning, it has a slower inference and this leads to a longer wait time because of delayed performance. Additionally, users of this model have to install separate dependencies as the source code is not integrated into easy-access hubs like Hugging Face Model Hub, TorchVision, or MMDetection.

<https://github.com/bbepoch/HoiTransformer>

**QPIC (Query-based Pairwise Interaction Classification):** Similar to HOI Transformer, it is also a single-stage detection model which directly predicts human-object interaction triplets using learnable queries. It has an edge over two-stages pipeline models like DiffHOI by generating fairly accurate results with a shorter inference time. However, both DiffHOI and QPIC relies on the same backbone visual encoder Swin-Large which requires extremely high memory and computational requirements. This is a model which works best with high batch sizes and high counts of frequent (non-rare) interactions because it has poor performance with rare interactions due to small representation.

<https://github.com/hitachi-rd-cv/qpic>