A Generalised Validation Mechanism without Annotation

Initial Design Proposal

▼ Lack-of-Label Problem

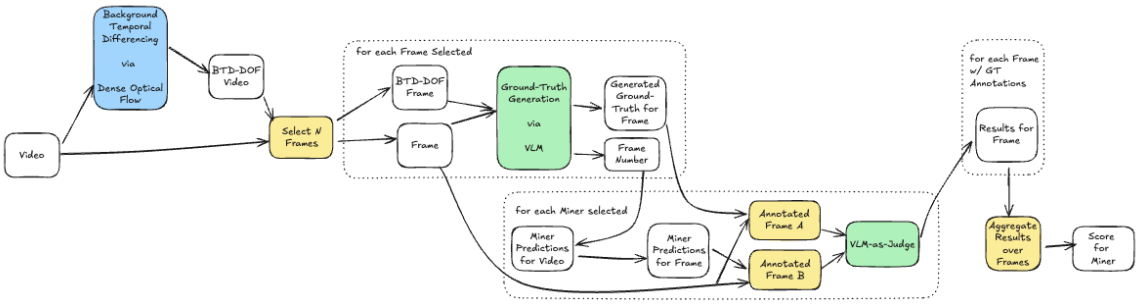
Most algorithms are typically trained in a fully or semi-supervised fashion, which relies on expensive training datasets of human-labelled images which are time-prohibitive to produce. Additional manual annotations are relied upon to evaluate these algorithms during training and after. Alternatively, a human-in-the-loop is employed to assess quality, but this is also slow and potentially unreliable as it depends on the visual inspection of a few selected samples instead of a systematic approach based on quantitative metrics.

▼ Proposed Solution

This document outlines a design proposal that objectively assess the quality of a miner's predictions (i.e. categorisation/classification, object detection and tracking methods) for videos across any domain in the absence of ground-truth annotations and human-in-the-loop strategies.

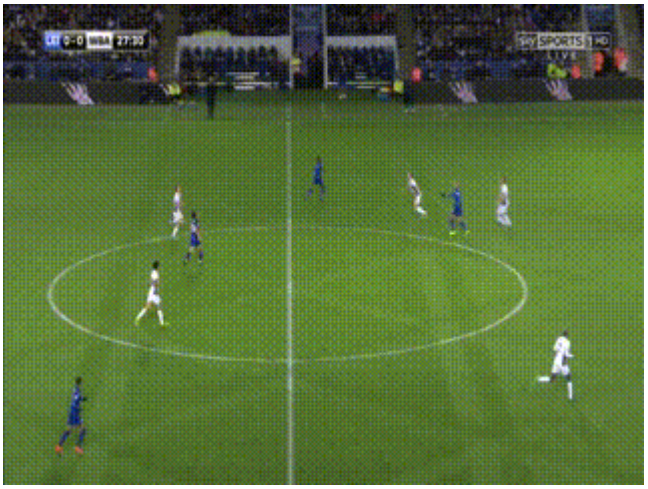
The design makes use of Visual (Large) Language Models (VLMs) to generate pseudo-annotations in an unsupervised manner and also to assess the quality of these annotations alongside a miner's annotations using the LLM-as-Judge paradigm.

The end-to-end design is shown below:



The main components at the heart of this pipeline are two VLMs (which can be the same model) which are doing quite different jobs. The first VLM is

generating pseudo ground-truth annotations while the second VLM is acting as a pairwise judge to compare these pseudo GT annotations against the predicted annotations by the miner and determine which are better. We illustrate the above flow using an example video clip of a football match:



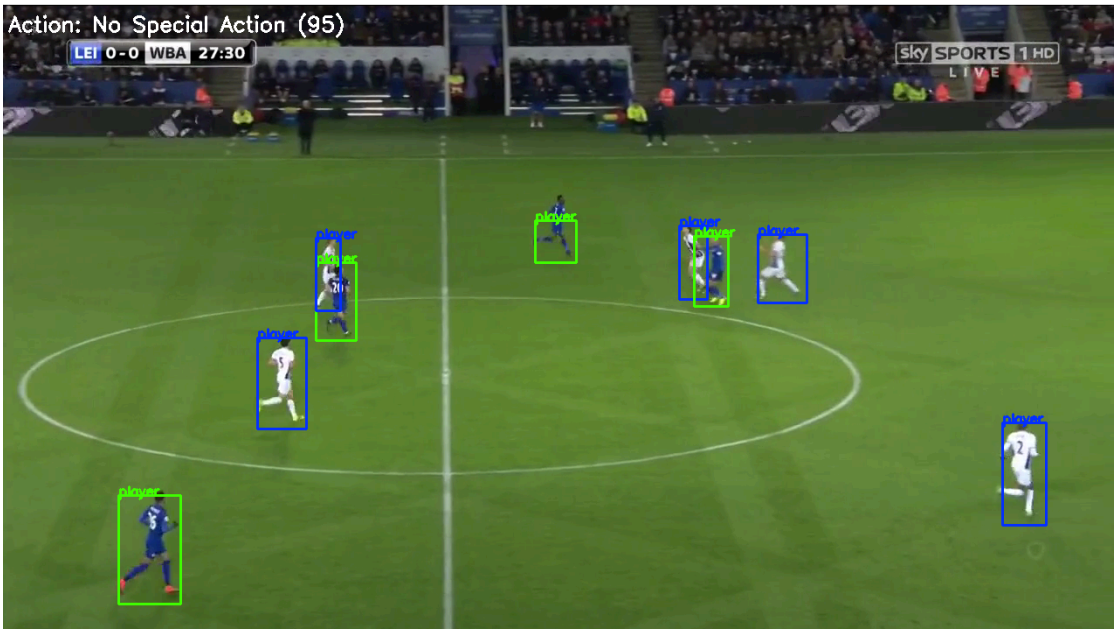
▼ Background Temporal Differencing: Optical Flow

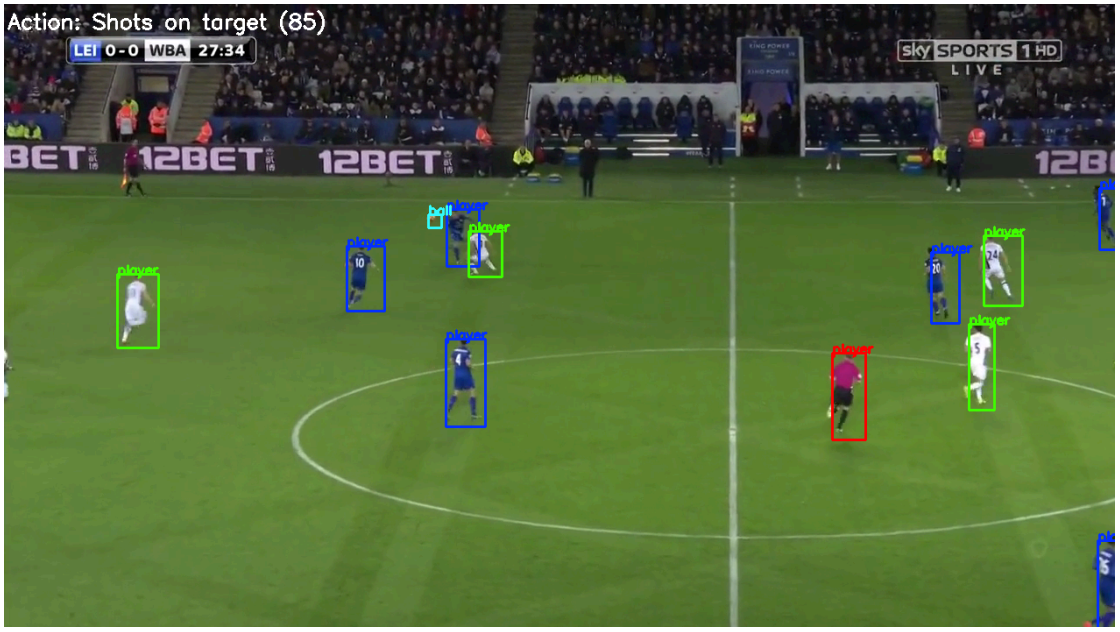
Since we are passing a single image at a time to the VLM, we lose the temporal information (ie the motion of the objects) that comes with the video. While the image encodes the spatial information at that moment in time, we encode the temporal context using background subtracting temporal differencing based on dense optical flow.

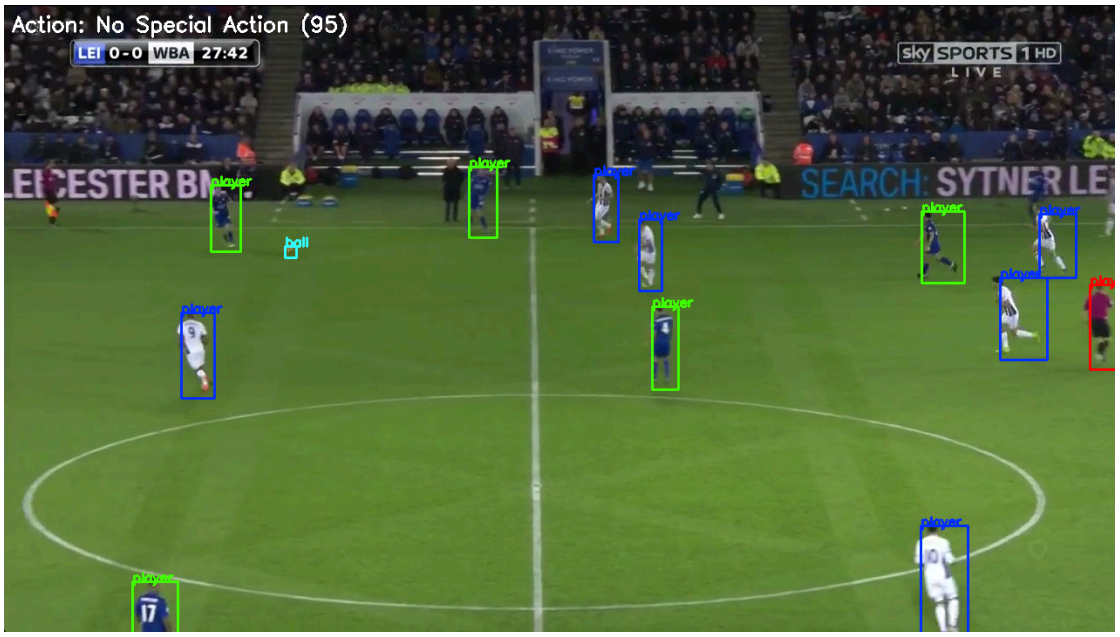
▼ Pseudo-Ground-Truth: VLM

Whenever a frame is passed to the VLM for analysis, it is accompanied by the temporal context extracted via optical flow. E.g.

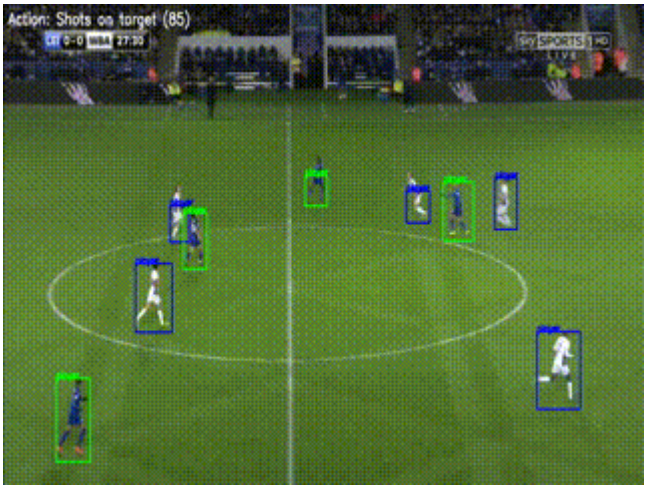
We prompt the VLM to generate annotations for the objects and action labels, etc. e.g.







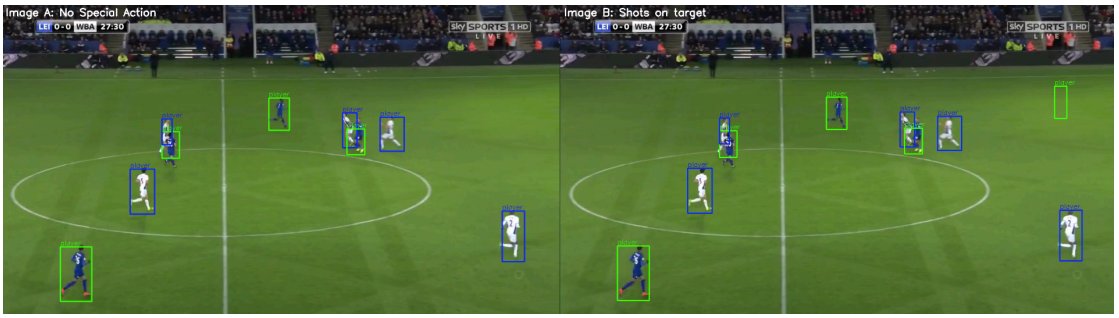
VLMs are generalisable, but they can be slow and expensive, so only a few frames will be selected from a video. However, generating the annotations for an entire video clip is possible. E.g. The video below has every 10th frame annotated by the VLM. (The lag is due to only every 10th frame being annotated. The colour changes on bboxes can be ignored for now as this is only because object tracking is not yet implemented and the generated cluster ids do not have a consistent reference point across frames)



The results are fairly reliable and consistent but there are small errors at times. This is why we have the VLM-as-Judge component to simultaneously verify the generated annotations along side the annotations predicted by a miner

▼ VLM-as-Judge

An example frame is passed to the VLM-as-Judge. The first image (image A) is annotated with the generated pseudo-GT predictions from the VLM in the previous step and the second image (image B) is annotated with the miner's predictions.



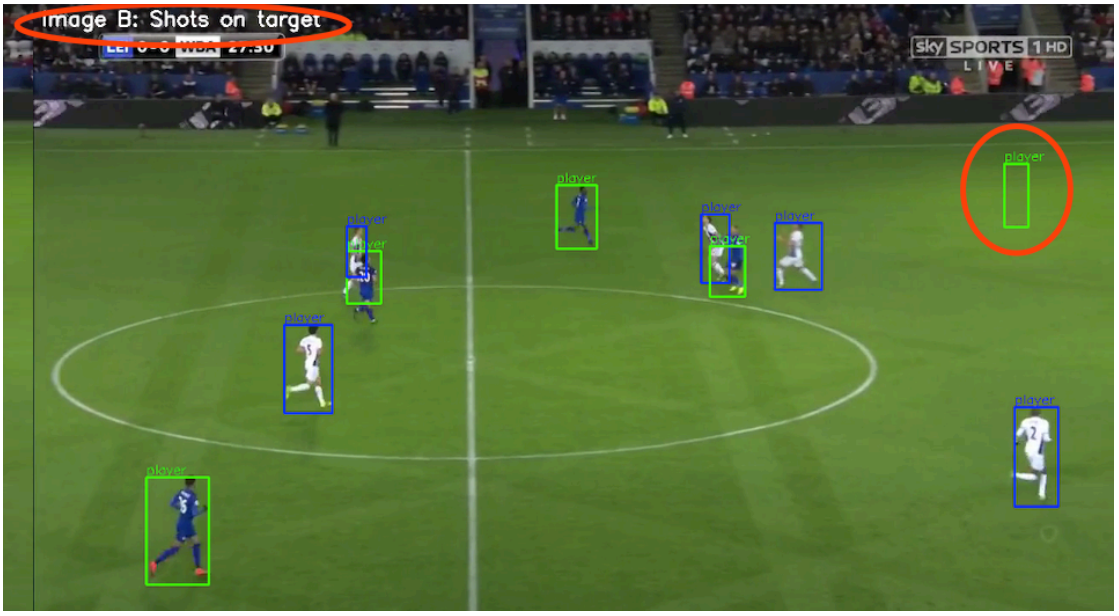
The VLM-as-Judge returns the following scores:


```
{
  "detections": {
    "thought": "Both images have similar bounding boxes around the p
layers, indicating that the detection of players is consistent between
the two. However, Image B has an additional bounding box on the rig
ht side that does not seem to enclose a player, suggesting an error in
detection.",
```

```
    "justification": "Image A correctly identifies all players with boundi
ng boxes, while Image B includes an extra, unnecessary bounding b
ox. This indicates that Image A has more accurate detection.",
    "winner": "Image A"
  },
  "bboxes": {
    "thought": "The bounding boxes in Image A are tightly fitted aroun
d the players, accurately representing their positions. In contrast, Im
age B has a bounding box that is not aligned with any player, which i
s incorrect.",
    "justification": "The bounding boxes in Image A are precise and co
rrectly placed, whereas Image B has an inaccurate bounding box, ma
king Image A superior in terms of bounding box accuracy.",
    "winner": "Image A"
  },
  "labels": {
    "thought": "Both images label the players correctly. There are no d
iscrepancies in the labels between the two images.",
    "justification": "Since both images label the players accurately wit
hout any errors, there is no difference in labeling quality.",
    "winner": "Tie"
  },
  "category": {
    "thought": "Image A is labeled as 'No Special Action,' which seem
s appropriate for the scene depicted. Image B is labeled 'Shots on tar
get,' which does not match the scene as there are no visible shots or
targets.",
    "justification": "The category label in Image A accurately describe
s the scene, while the label in Image B is misleading and does not co
rrespond to the image content.",
    "winner": "Image A"
  }
}
```


Notice that although the two annotations are very similar, the miner’s predictions (image B) has an additional bbox incorrectly predicted in the top right and the action classification is “Shots on target” (as opposed to

“No Special Action” as predicted by the pseudo-GT annotations in image A).



The VLM-as-Judge correctly picks up on both of these mistakes and ultimately considers them when deciding which annotation is better

> Image B includes an extra, unnecessary bounding box. This indicates that Image A has more accurate detection.

> Image B is labeled 'Shots on target,' which does not match the scene as there are no visible shots or targets.

▼ Trade Offs

Hybrid Design vs VLM to generate GT labels Only

Is there really is a need to use the VLM twice for two different jobs? Isn’t having a VLM generate pseudo GT annotations enough to compare and evaluate a miner’s predictions? While this is possible to use the generated GT annotations in this way, as we would with human annotations, however, this would require us to trust these generated annotations fully - which is risky. Having a second VLM-as-Judge does not only serve to assess the quality of the annotations predicted by a miner, it simultaneously verifies the pseudo GT annotations generated by the first VLM too!

Hybrid Design vs VLM-as-Judge Only

If we cannot rely on the generated pseudo GT annotations, why not just skip using the first VLM altogether and jump to using the VLM-as-judge, simplifying the end-to-end flow considerably. Unfortunately, this would mean the complexity of the judging task would become increasingly difficult as the VLM judge would need to assign a score almost arbitrarily to a miner's predictions without referring to any ground-truth at all. This would make the accuracy and consistency of the judgement very unreliable. This is overcome by providing a second set of annotations (i.e. from the pseudo GT annotations), and the beautiful part is that these reference annotations need not be perfect because the judge is only determining which annotation is better! With reference annotations, the judge's job simplifies to: is this annotation better than our reference annotation? And a simpler judgement task means the VLM will generate more accurate and consistent answers.

▼ Future Directions

- Tracking: adding object ids to track each object across frames
- Multiple VLM models: we currently use a single VLM model to predict the Ground Truth labels but we could use multiple VLM models and infer the GT from them all for potentially more reliable GT predictions
- Multiple VLM-as-Judge: we currently use a single VLM-as Judge but we can use multiple VLM judges to debate one another (as suggested in the paper)
- Multiple Miners: we currently evaluate a single miner at a time but there is potential to compare a group of miners at once
- Intelligently selecting frames in a video: currently we select N frames randomly from the video but it may be better to select a window of frames in close time proximity or some other selection method