A Survey on Open-Vocabulary Detection

Jiewei Lyu, Sun Yat-sen University

WHY WE NEED OVD

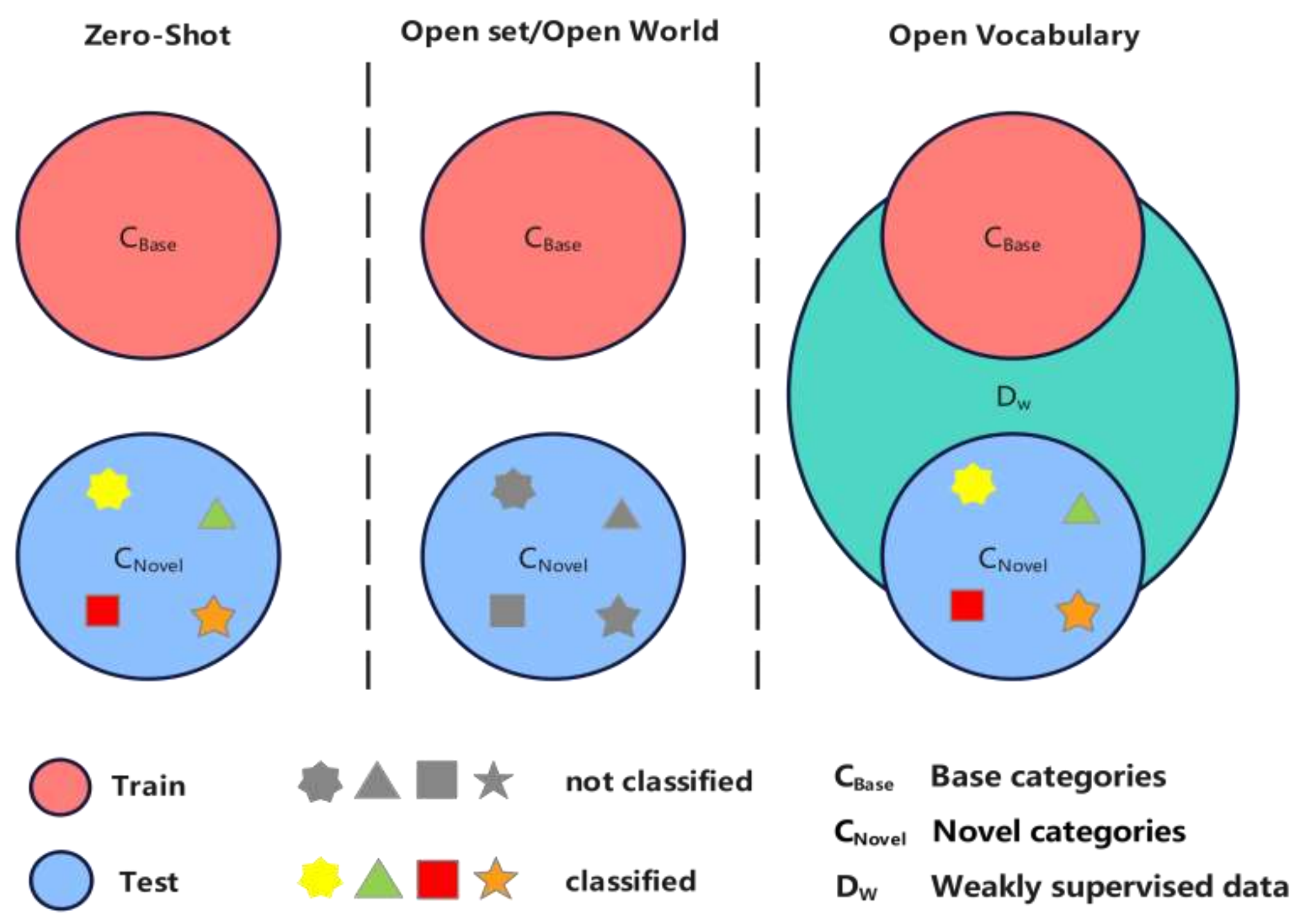
- **Traditional object detection models struggle to recognize unfamiliar categories in open-world scenarios, as they are limited by the predefined categories in closed-set training data.**

- **Existing datasets are typically small in scale, even the largest LVIS dataset annotates only 1,203 categories.**

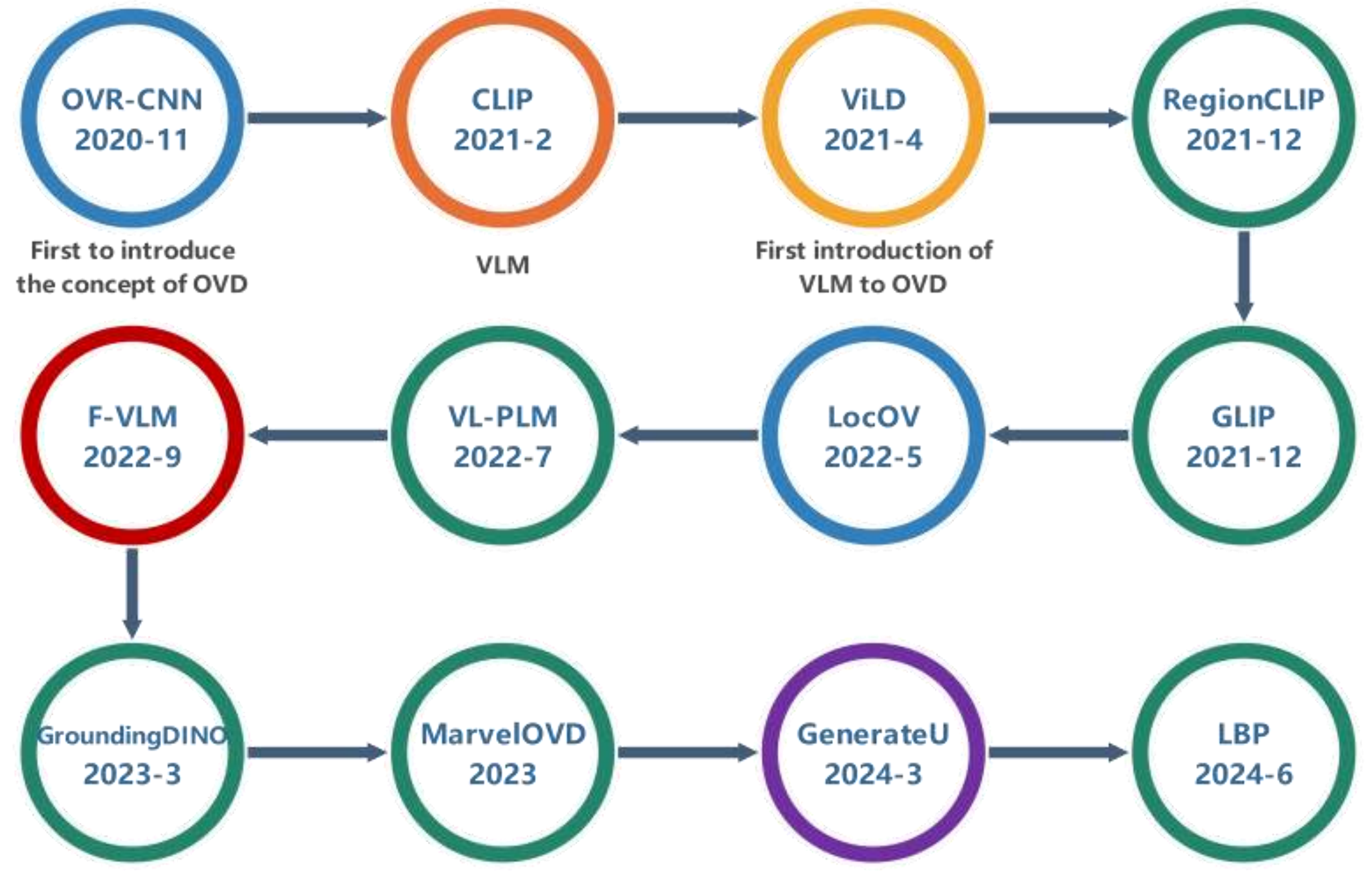
- **OVD addresses these limitations by combining images with natural language descriptions, allowing models to use a broader vocabulary during training. This allows the model to continuously update and recognize new objects and scenes, enhancing its ability to identify a broader range of unseen categories during inference.**

RELATED REASEARCH

- **Zero-Shot Detection**
- **Open Set Detection**
- **Open World Detection**



TIMELINE



VLMs

Inspired by advances in natural language processing, Vision-Language Models (VLMs) are pre-trained on large-scale image-text pairs available abundantly on the internet.

- **Goal:** Learn image-text correlations.

- **Method**: First use text encoders and image encoders to extract features, and then learn visual-language correlations according to specific pre-training objectives.

C
L
I
P

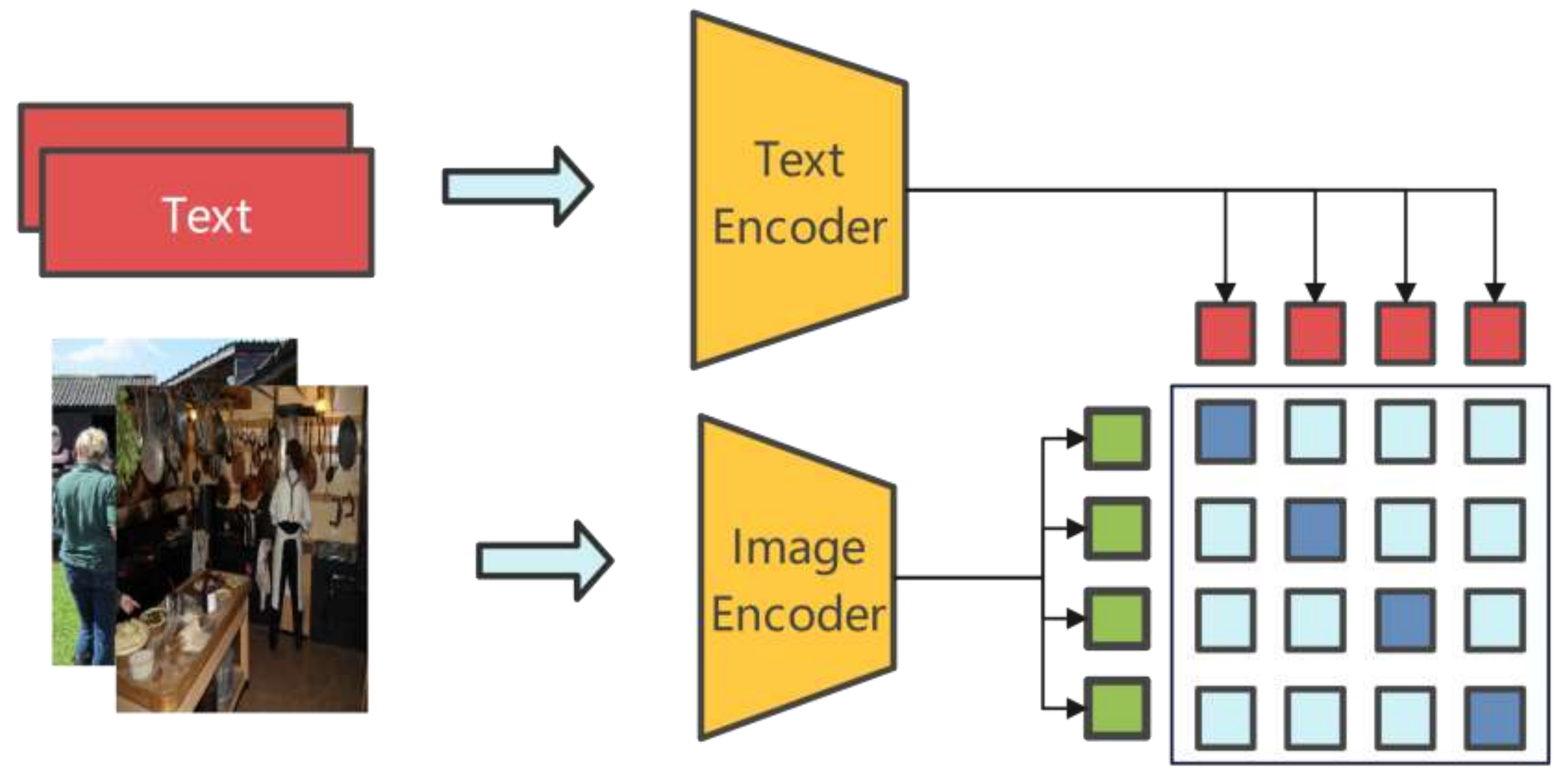
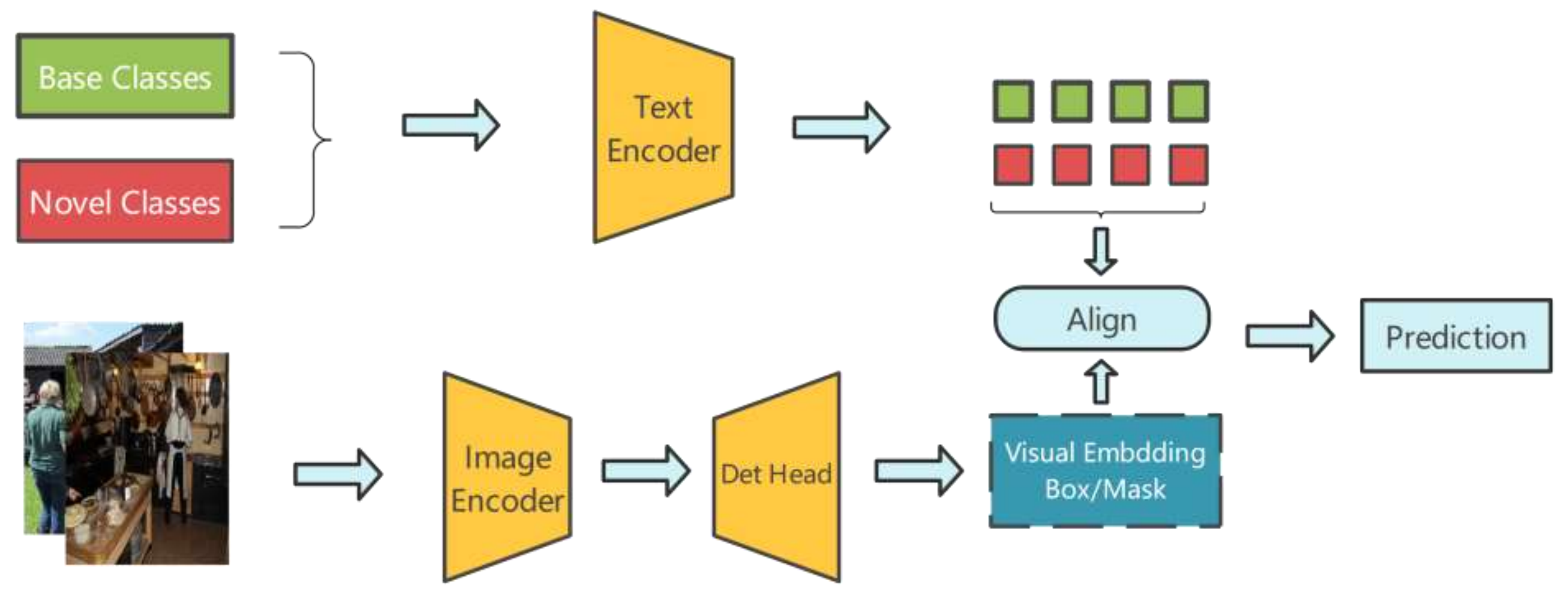


IMAGE-TEXT PAIR

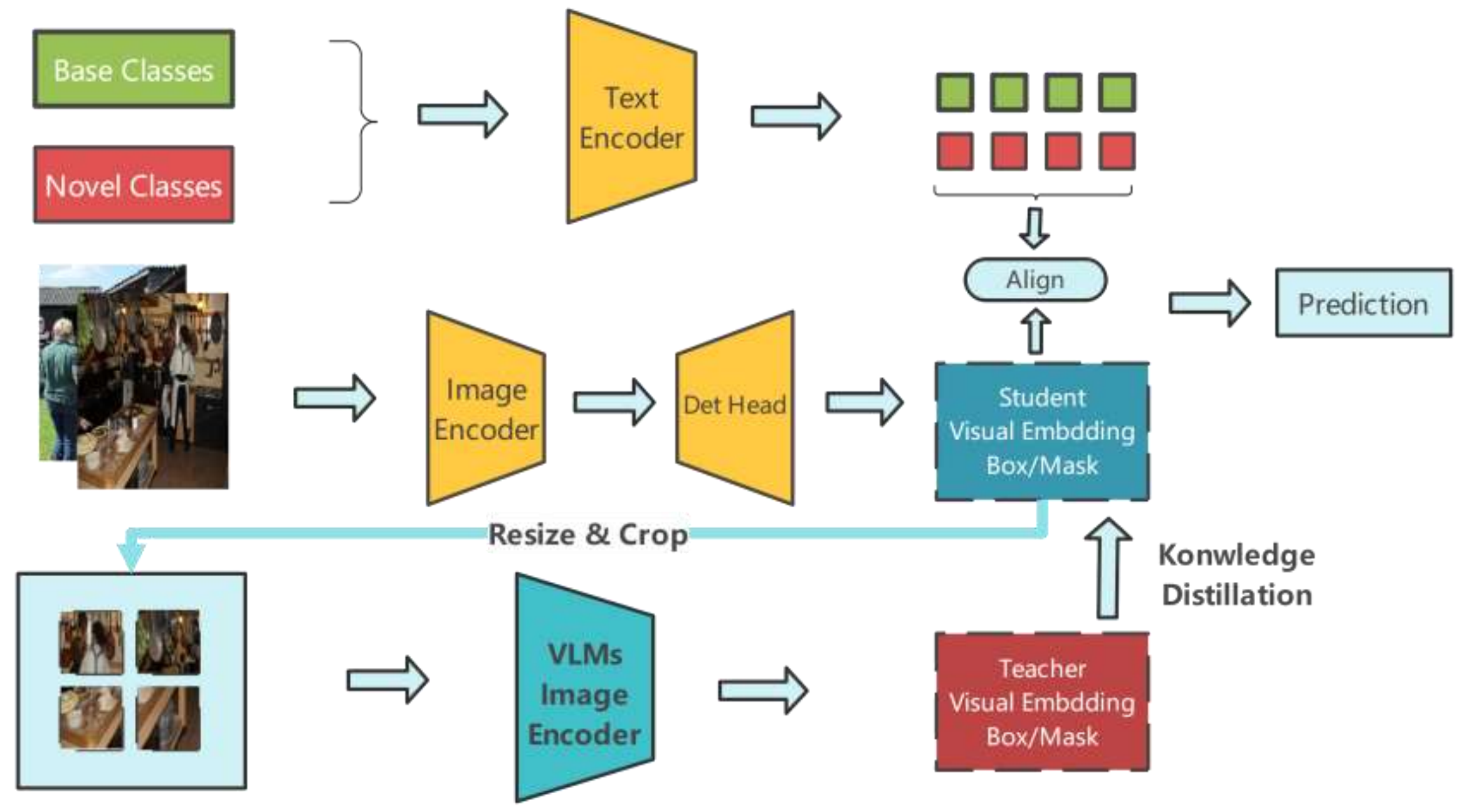
- **OVR-CNN:** First to introduce the concept of OVD

- **LocVO:** Refine the utilization of image-caption data



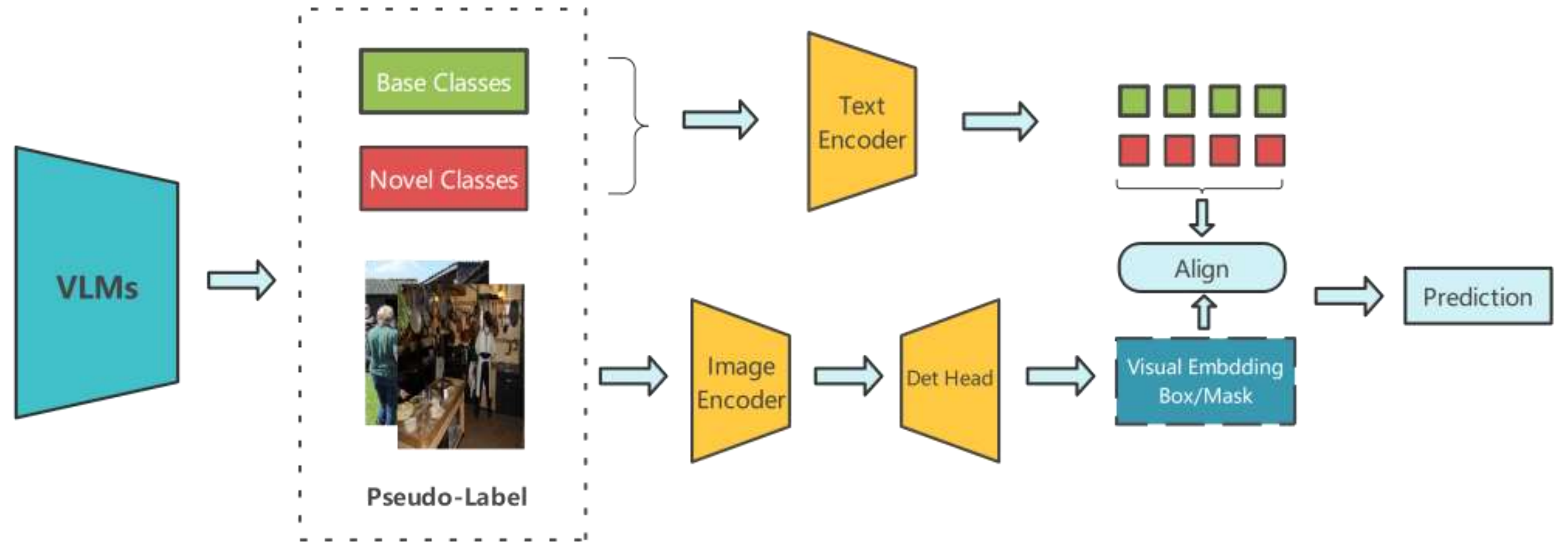
KONWLEDGE DISTILLATION & TRANSFER LEARNING

- **ViLD:** First introduced the pretrained multimodal model CLIP to enhance OVD performance**.**



- **F-VLM:**Trains the detector head on a frozen VLM backbone.

PSEUDO LABEL



- **RegionCLIP:**Capture fine-grained alignment.

- **GLIP:** Unify phrase grounding and object detection tasks

- **GroundingDINO:**Upgrade detector to a Transformer-based

- **LBP:**Better differentiate background and novel.

Improving Pseudo-Label Quality

- **VL-PLM:**Combine RPN scores

- **MarvelOVD:**Adaptive weighting mechanism and hierarchical label assignment

GENERATIVE

- **GenerateU:Transform OVD into a generative problem.**

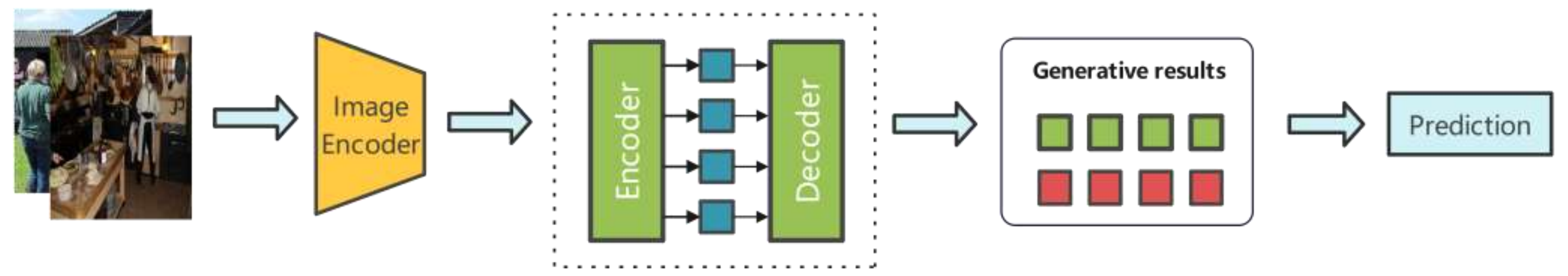


Image Encoder
Encoder
Decoder
Generative results
Prediction