Method	Pros	Cons
Pretrained CLIP (on b-boxes from object detection)	-No data required -No training required	-Most design algorithm for ranking based on b-boxes -Probably less accurate than trained models
Self-Supervised (e.g., SimCLR, EsViT)	-No labeled data required -Probably more accurate than untrained models	-Training required -Might capture features that are not related to clothes, because of the lack of label (can be used on b-boxes to fix the issue)
Classification Network		-Training required -Labeled data required
Adding Extra Layer to Pretrained CLIP	-Probably great performance on the features that are included in captions	-Caption required for images (can be extracted from object detection features) -Only considers features that are included in captions for measuring similarity