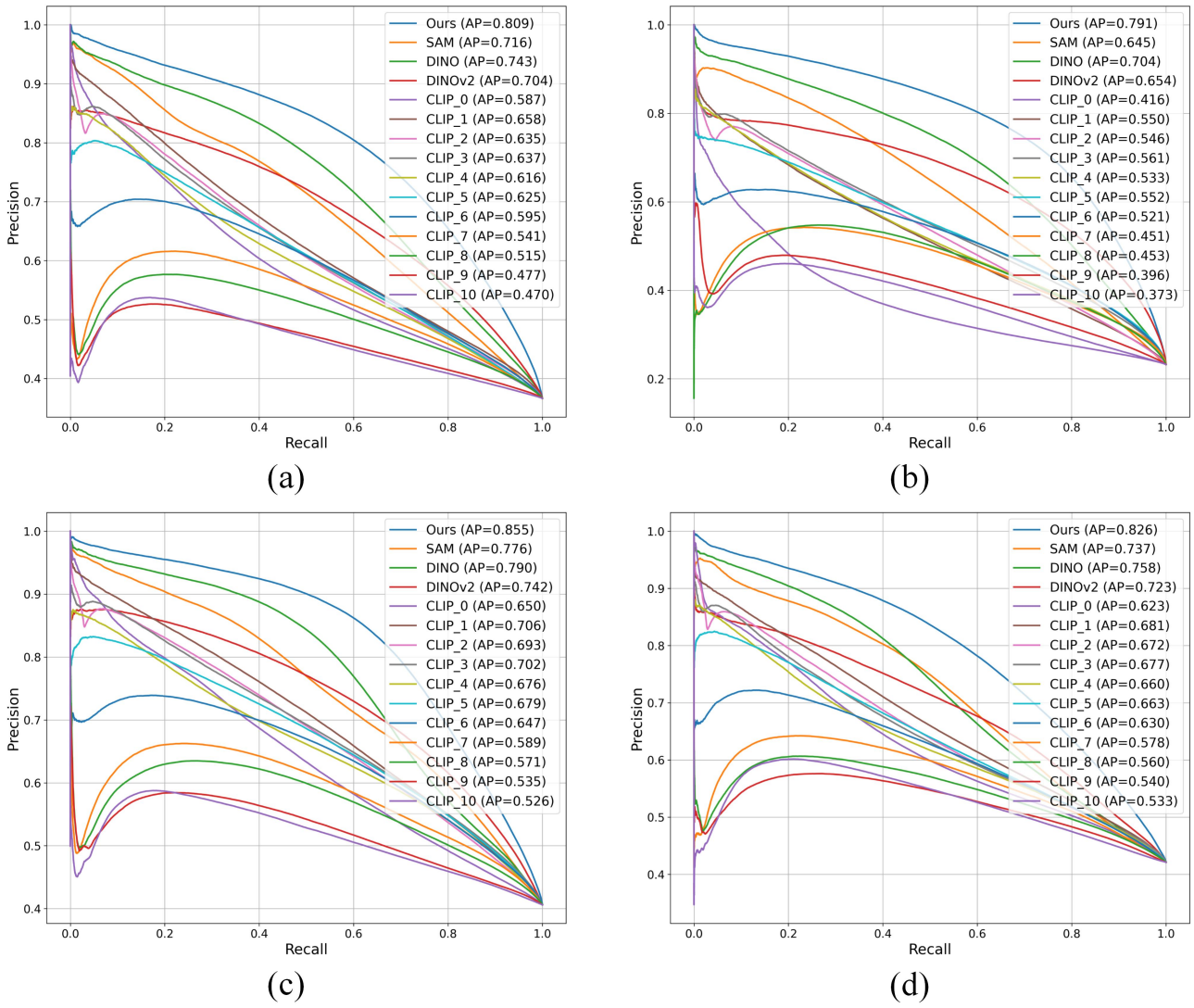

Figure 3. Precision-Recall Performance of Vision Foundation Models for Semantic Consistency Classification. Comparison of Average Precision (AP) between CLIP (layers 0-10) and other Vision Foundation Models (DINO, DINOV2, and SAM) across (a) ADE20K, (b) Cityscapes, (c) COCO-Stuff, and (d) Pascal Context datasets. Higher AP indicates better performance in distinguishing whether image patches belong to the same semantic category.