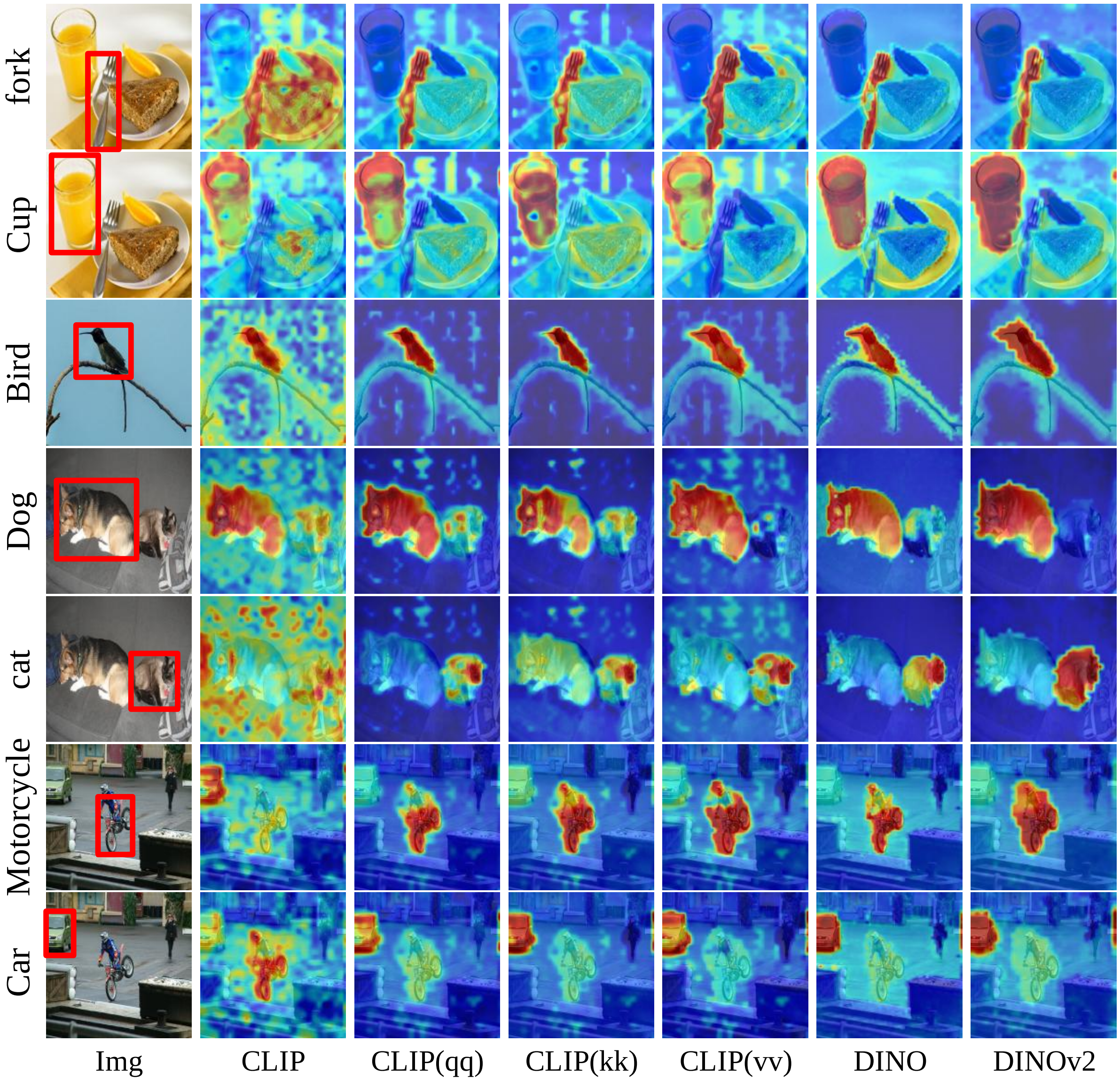

*Figure 6.* Feature Visualization Comparison of Vision Encoder Architectures. Qualitative comparison of features extracted by CLIP, DINO, and DINOV2 visual encoders. CLIP(qq), CLIP(kk), and CLIP(vv) represent different attention weight recombination strategies in CLIP's final layer. Standard CLIP outputs (column 2) show limited localization capability, while self-self attention modifications (columns 3-5) improve localization despite noise artifacts. DINO and DINOV2 (columns 6-7) demonstrate superior object-level feature localization.