Performance Comparison Across Methods and Datasets
(Top-1 Accuracy %)

Method	ImageNet	ChestX-ray	EuroSAT	Oxford Pets	Food-101	Average
CLIP Baseline	68.3	42.1	51.2	83.5	79.8	65.0
CLIP Ensemble (80 prompts)	69.1	43.8	53.6	85.2	81.3	66.6
DCLIP (LLM prompts)	69.8	45.2	54.1	86.1	82.1	67.5
WaffleCLIP	70.2	44.9	54.8	86.8	82.7	67.9
AutoCLIP	70.9	46.3	56.2	87.9	83.4	68.9
TPT (Test-time tuning)	71.4	47.1	57.0	88.5	84.2	69.6
Ours (w/o adaptation)	**71.8**	**48.7**	**58.3**	**88.9**	**84.8**	**70.5 (+5.5)**
Ours (full)	**72.5**	**48.9**	**59.5**	**89.8**	**85.6**	**71.3 (+6.3)**
Ours (+ 500 samples)	**74.1**	**52.7**	**63.4**	**91.8**	**87.3**	**73.9 (+8.9)**