Statistical Significance Analysis
(Paired t-test vs. Our Full Method, ImageNet)

Method	Accuracy	p-value	Stars	Significance
CLIP Baseline	68.3 ± 0.5	<0.001	***	Highly Sig.
CLIP Ensemble	69.1 ± 0.4	<0.001	***	Highly Sig.
DCLIP	69.8 ± 0.5	<0.001	***	Highly Sig.
WaffleCLIP	70.2 ± 0.4	<0.001	***	Highly Sig.
AutoCLIP	70.9 ± 0.5	<0.001	***	Highly Sig.
TPT	71.4 ± 0.6	<0.001	***	Highly Sig.
Ours (w/o adapt)	71.8 ± 0.4	<0.01	**	Significant
Ours (full)	**74.1 ± 0.3**	**—**	**—**	**Reference**

*** $p < 0.001$ (highly significant) | ** $p < 0.01$ (significant) | * $p < 0.05$ (marginally significant)
All tests performed with n=3 runs, 1000 samples per run