Additional Experiments for Rebuttal

	Method	EuroSAT	OxfordPets	DTD	Caltech101	FGVCAircraft	UCF101	Flowers102	Average
RN50	$f(\cdot)$	0.456 ± 0.040	**0.229 ± 0.002**	0.283 ± 0.011	0.324 ± 0.001	0.066 ± 0.001	0.184 ± 0.010	**0.137 ± 0.003**	0.240
	$g_I(\cdot)$	0.469 ± 0.038	0.214 ± 0.007	0.282 ± 0.007	0.324 ± 0.001	0.074 ± 0.012	0.179 ± 0.009	0.136 ± 0.003	0.240
	$g_C(\cdot)$	**0.487 ± 0.001**	0.204 ± 0.013	**0.283 ± 0.007**	**0.353 ± 0.002**	**0.076 ± 0.004**	**0.205 ± 0.006**	0.130 ± 0.001	**0.248**
RN101	$f(\cdot)$	0.350 ± 0.051	**0.211 ± 0.004**	0.250 ± 0.017	0.298 ± 0.001	0.056 ± 0.003	0.146 ± 0.013	0.133 ± 0.005	0.206
	$g_I(\cdot)$	0.408 ± 0.024	0.197 ± 0.011	0.245 ± 0.018	0.294 ± 0.006	0.070 ± 0.012	0.145 ± 0.010	**0.133 ± 0.002**	0.213
	$g_C(\cdot)$	**0.423 ± 0.007**	0.177 ± 0.012	**0.260 ± 0.010**	**0.312 ± 0.002**	**0.067 ± 0.006**	**0.190 ± 0.015**	0.120 ± 0.001	**0.221**
ViTB32	$f(\cdot)$	0.525 ± 0.010	**0.439 ± 0.011**	0.372 ± 0.002	0.352 ± 0.002	0.089 ± 0.001	0.214 ± 0.011	0.188 ± 0.008	0.311
	$g_I(\cdot)$	0.557 ± 0.006	0.429 ± 0.019	0.379 ± 0.002	0.355 ± 0.005	0.094 ± 0.013	0.207 ± 0.013	0.186 ± 0.006	0.315
	$g_C(\cdot)$	**0.584 ± 0.013**	0.380 ± 0.020	**0.385 ± 0.007**	**0.392 ± 0.002**	**0.098 ± 0.006**	**0.234 ± 0.010**	**0.217 ± 0.003**	**0.327**
ViTH14	$f(\cdot)$	0.577 ± 0.011	**0.567 ± 0.005**	0.392 ± 0.003	0.335 ± 0.001	0.116 ± 0.001	0.253 ± 0.017	0.206 ± 0.000	0.349
	$g_I(\cdot)$	0.596 ± 0.007	0.548 ± 0.025	0.394 ± 0.003	0.332 ± 0.003	0.131 ± 0.020	0.247 ± 0.005	0.203 ± 0.001	0.350
	$g_C(\cdot)$	**0.634 ± 0.014**	0.477 ± 0.021	**0.403 ± 0.009**	**0.371 ± 0.005**	**0.143 ± 0.003**	**0.287 ± 0.010**	**0.241 ± 0.001**	**0.365**
ViTG14	$f(\cdot)$	0.560 ± 0.024	**0.582 ± 0.008**	0.389 ± 0.004	0.339 ± 0.002	0.130 ± 0.005	0.254 ± 0.014	0.224 ± 0.002	0.354
	$g_I(\cdot)$	0.598 ± 0.004	0.566 ± 0.023	0.388 ± 0.006	0.336 ± 0.001	0.145 ± 0.021	0.252 ± 0.003	0.216 ± 0.004	0.357
	$g_C(\cdot)$	**0.609 ± 0.019**	0.503 ± 0.017	**0.402 ± 0.005**	**0.376 ± 0.004**	**0.161** ± 0.005	**0.289 ± 0.015**	**0.258 ± 0.002**	**0.371**

Table A: UP-DP Sampling Variants: (1) Medoid selection based on $f(\cdot)$ features, (2) Medoid selection based on $g_I(\cdot)$ features, (3) the highest probability instance from each cluster predicted by $g_C(\cdot)$.

Model	Random	USL-I	USL-M	Ours
Dinov2_ViTS	54.9	56.4	58.1	58.5
Dinov2_ViTB	55.9	57.3	58.3	58.8
Dinov2_ViTL	55.2	54.2	53.7	57.1
Dinov2_ViTG	47.9	51.9	51.4	53.6
Average	53.5	55.0	55.4	**57.0**

Table B: Experiment on Semantic Segmentation Task: The mIoU is used as the evaluation metric.

Linear Probe Model		Random	USL-I	USL-M	Ours (BLIP-2 v1)	Ours (BLIP-2 v2)
CLIP	RN50	0.119 ± 0.021	0.131 ± 0.020	0.100 ± 0.037	0.169 ± 0.007	0.182 ± 0.010
	RN101	0.085 ± 0.018	0.103 ± 0.011	0.077 ± 0.040	0.152 ± 0.007	0.170 ± 0.008
	ViTB32	0.150 ± 0.035	0.187 ± 0.027	0.141 ± 0.044	0.221 ± 0.009	0.239 ± 0.006
	ViTH14	0.153 ± 0.027	0.196 ± 0.034	0.147 ± 0.052	0.231 ± 0.010	0.258 ± 0.004
	ViTG14	0.179 ± 0.028	0.200 ± 0.032	0.146 ± 0.050	0.233 ± 0.011	0.255 ± 0.005
BLIP-2	ViTL	0.416 ± 0.016	0.360 ± 0.010	0.503 ± 0.010	0.460 ± 0.005	0.477 ± 0.004
	ViTG	0.461 ± 0.009	0.387 ± 0.007	0.557 ± 0.016	0.504 ± 0.015	0.527 ± 0.002
Average		0.223	0.224	0.240	0.282	**0.302**

Table C: Impact of Different Versions of BLIP-2: BLIP-2 v1 utilizes ViTL, while BLIP-2 v2 utilizes ViTG as the image encoder. Note that the USL-M use the features learned by our method.

Linear Probe Model		BLIP-2 v1 1:3	BLIP-2 v1 1:1	BLIP-2 v1 3:1	BLIP-2 v2 1:3	BLIP-2 v2 1:1	BLIP-2 v2 3:1
CLIP	RN50	0.169 ± 0.007	0.151 ± 0.003	0.160 ± 0.006	0.182 ± 0.010	0.159 ± 0.025	0.174 ± 0.029
	RN101	0.152 ± 0.007	0.097 ± 0.002	0.130 ± 0.004	0.170 ± 0.008	0.111 ± 0.027	0.132 ± 0.027
	ViTB32	0.221 ± 0.009	0.209 ± 0.005	0.215 ± 0.011	0.239 ± 0.006	0.253 ± 0.028	0.267 ± 0.020
	ViTH14	0.231 ± 0.010	0.211 ± 0.005	0.226 ± 0.007	0.258 ± 0.004	0.266 ± 0.020	0.285 ± 0.012
	ViTG14	0.233 ± 0.011	0.213 ± 0.005	0.230 ± 0.008	0.255 ± 0.005	0.265 ± 0.024	0.282 ± 0.015
BLIP-2	ViTL	0.460 ± 0.005	0.507 ± 0.005	0.476 ± 0.005	0.477 ± 0.004	0.472 ± 0.008	0.490 ± 0.004
	ViTG	0.504 ± 0.015	0.530 ± 0.007	0.521 ± 0.004	0.527 ± 0.002	0.508 ± 0.007	0.533 ± 0.009
Average		0.281	0.274	0.280	0.301	0.290	**0.309**

Table D: Impact of Weights between Instance-level and Cluster-level Loss.

Linear Probe Model		50	100	150	200	250	300
CLIP	RN50	0.151 ± 0.064	0.182 ± 0.010	0.230 ± 0.004	0.249 ± 0.008	0.312 ± 0.001	0.316 ± 0.006
	RN101	0.124 ± 0.085	0.170 ± 0.008	0.196 ± 0.008	0.200 ± 0.014	0.263 ± 0.002	0.282 ± 0.011
	ViTB32	0.205 ± 0.042	0.239 ± 0.006	0.332 ± 0.011	0.332 ± 0.008	0.410 ± 0.001	0.437 ± 0.002
	ViTH14	0.208 ± 0.048	0.258 ± 0.004	0.338 ± 0.014	0.354 ± 0.008	0.434 ± 0.001	0.441 ± 0.005
	ViTG14	0.210 ± 0.044	0.255 ± 0.005	0.339 ± 0.013	0.360 ± 0.010	0.438 ± 0.004	0.442 ± 0.007
BLIP-2	ViTL	0.405 ± 0.003	0.477 ± 0.004	0.531 ± 0.003	0.572 ± 0.002	0.580 ± 0.003	0.604 ± 0.001
	ViTG	0.441 ± 0.003	0.527 ± 0.002	0.579 ± 0.002	0.609 ± 0.003	0.603 ± 0.001	0.643 ± 0.004
Ours Average		**0.249**	**0.301**	**0.364**	**0.382**	**0.434**	**0.452**
Random Average		0.139	0.223	0.287	0.338	0.385	0.425
USL-I Average		0.141	0.224	0.294	0.334	0.353	0.375
USL-M Average		0.196	0.240	0.313	0.351	0.416	0.444

Table E: Impact of Annotation Budget.

Linear Probe Model		Random	BLIP-2 w/o Prompt	BLIP-2
CLIP	RN50	0.119 ± 0.021	0.136 ± 0.005	0.183 ± 0.011
	RN101	0.086 ± 0.018	0.113 ± 0.010	0.171 ± 0.008
	ViTB32	0.150 ± 0.035	0.166 ± 0.008	0.239 ± 0.007
	ViTH14	0.153 ± 0.027	0.176 ± 0.003	0.259 ± 0.004
	ViTG14	0.178 ± 0.028	0.177 ± 0.002	0.256 ± 0.006
BLIP-2	ViTL	0.416 ± 0.016	0.314 ± 0.004	0.478 ± 0.004
	ViTG	0.461 ± 0.009	0.341 ± 0.008	0.527 ± 0.002
Average		0.223	0.203	**0.302**

Table F: Impact of Training without Prompt.