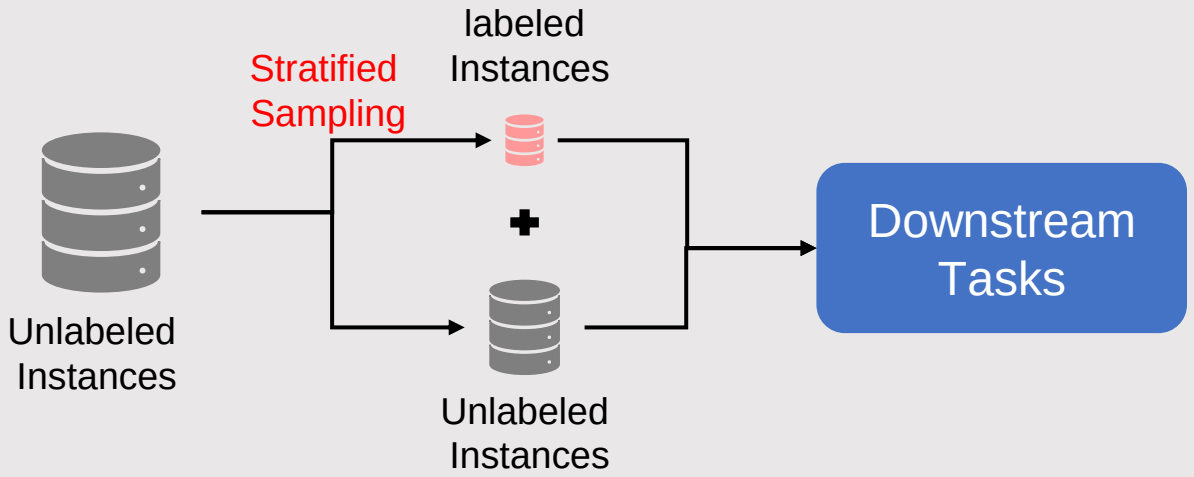
*UP-DP: **U**nsupervised **P**rompt Learning for **D**ata **P**re-Selection with Vision-Language Models*

Submitted to 37th Conference on Neural Information Processing Systems
Submission ID # 14432

Two Lines of Data Efficiency Learning
Semi-supervised learning (SSL)



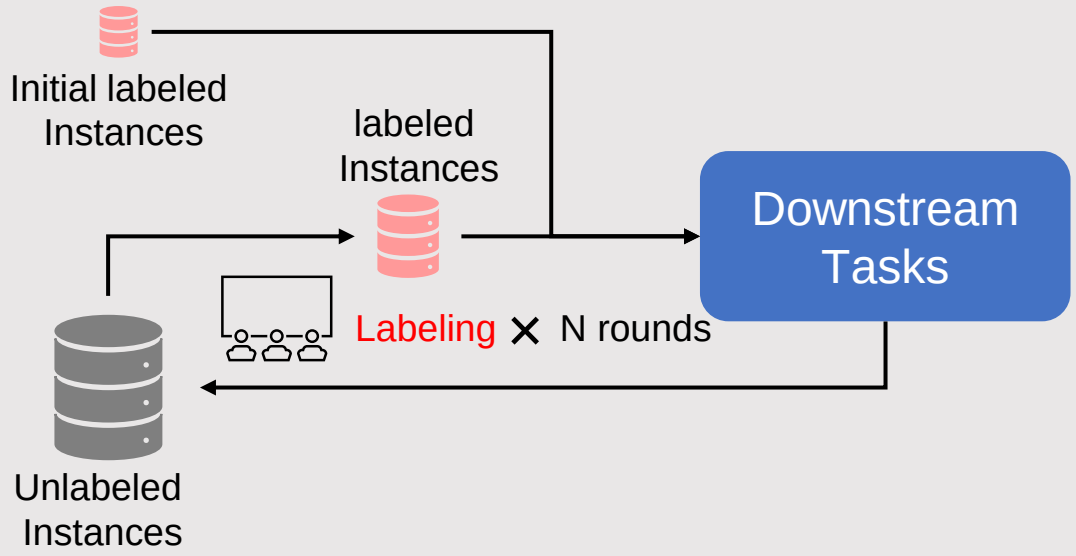
Semi-supervised learning (SSL) [1]:

- Relies on invalid assumption of **stratified sampling**
(Require equal sampling from each class)

[1] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. Advances in neural information processing systems, 33:596–608, 2020.432

Two Lines of Data Efficiency Learning

Active Learning (AL)

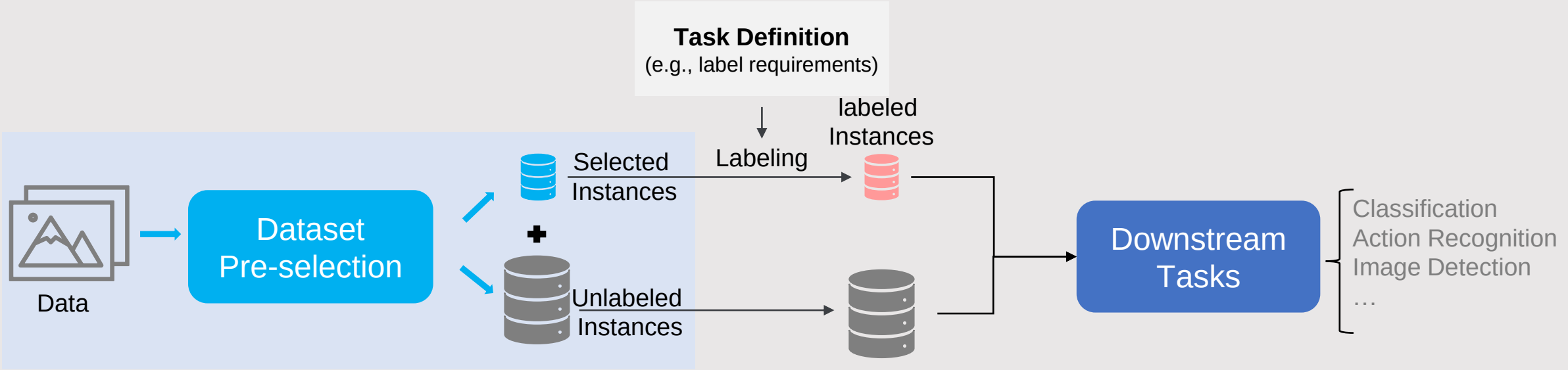


Active Learning (AL) [2]:

+ avoid of **stratified sampling**
- require an initial labeled set
- requires **multiple rounds** of training and labeling
- selected instances are task specific

[2] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. International Conference on Learning Representations 2018.

Our approach: Data Pre-Selection

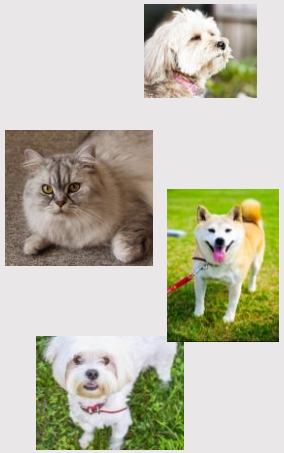


Data Pre-Selection (new task):

+ avoid of stratified sampling
+ no initial labeled set
+ a single pass
+ fit for various ***unknown*** tasks

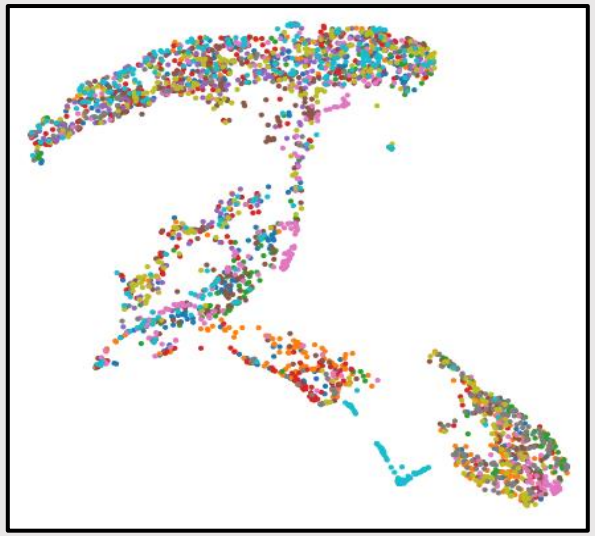
The Core Idea:

Select **diverse** and **representative** instances based on semantically meaningful multimodal feature from **adapted** BLIP-2 [3].



OxfordPets Dataset

Multimodal Feature From BLIP-2 with **Learned** Prompt

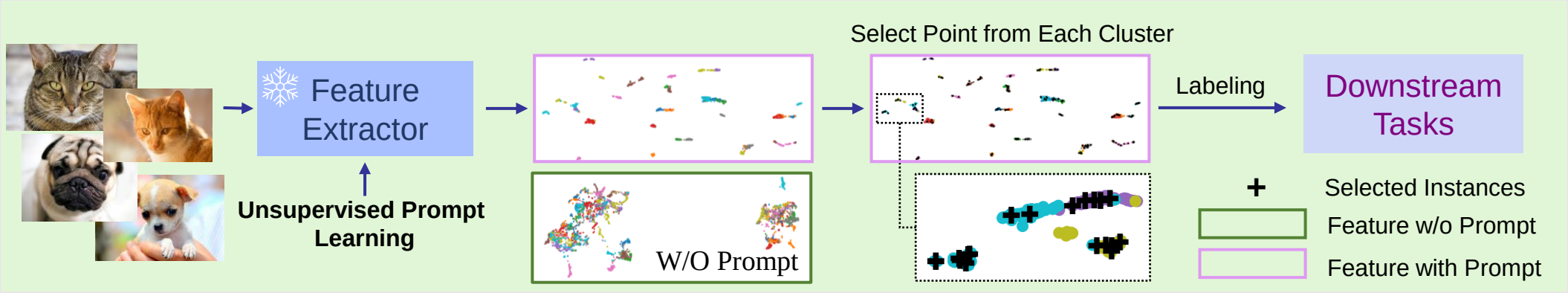


Multimodal Feature with Random Prompt

Image Feature Only

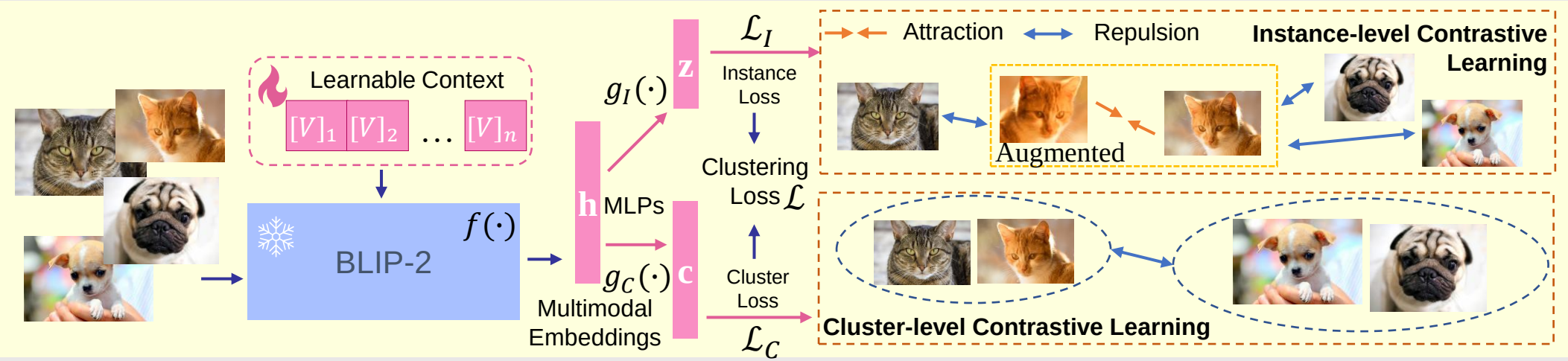
[3] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597, 2023.

Data Pre-selection Workflow



- Extract multimodal features for clustering
- Select representative and diverse instances for labeling
  - Medoid in each cluster
  - Instance with highest probability in each cluster

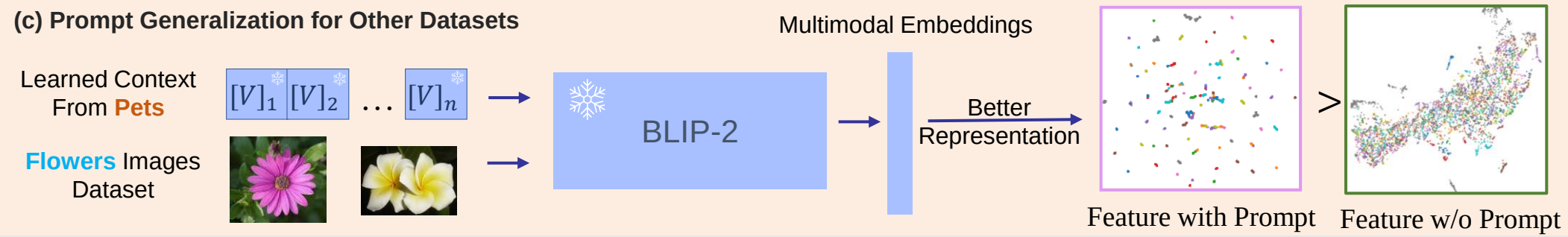
Unsupervised Prompt Learning
Adapt BLIP-2 for Data Pre-selection



- Three cooperatively learned elements
  - Learnable prompts
  - Instance-level MLP head $g_I(\cdot)$
  - Cluster-level MLP head $g_C(\cdot)$
- Unsupervised clustering objectives [4]
  - Instance-level contrastive loss
  - Cluster-level contrastive loss

[4] Yunfan Li, Peng Hu, Zitao Liu, Dezhong Peng, Joey Tianyi Zhou, and Xi Peng. Contrastive clustering. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, pages 8547–8555, 2021

Prompt Generalization across Other Datasets



- Learned prompt from a single dataset can be **directly** applied to enhance the feature quality of other datasets

Experiments

- Downstream task
  - Image classification - linear probe on CLIP image encoders
- Datasets
  - EuroSAT
  - OxfordPets
  - DTD
  - Caltech101
  - FGVAircraft
  - UCF101
  - Flowers102
- Domain generalization task
  - K-Nearest Neighbors classifier on extracted feature

Linear Probe Results of CLIP

	Method	EuroSAT	OxfordPets	DTD	Caltech101	FGVCAircraft	UCF101	Flowers102	Average
RN50	Random	0.221 ± 0.055	0.193 ± 0.032	0.202 ± 0.016	0.336 ± 0.021	0.057 ± 0.009	0.113 ± 0.012	0.121 ± 0.015	0.178
	USL-I	0.309 ± 0.052	0.161 ± 0.020	0.188 ± 0.028	0.297 ± 0.008	0.048 ± 0.008	0.133 ± 0.013	0.104 ± 0.043	0.177
	USL-M	0.289 ± 0.044	0.191 ± 0.038	0.185 ± 0.030	0.286 ± 0.008	0.064 ± 0.010	0.108 ± 0.027	0.130 ± 0.026	0.179
	Ours $f(\cdot)$	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
	Ours $g_I(\cdot)$	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
	Ours $g_C(\cdot)$	**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	Random	0.217 ± 0.062	0.193 ± 0.021	0.139 ± 0.026	0.274 ± 0.023	0.042 ± 0.013	0.093 ± 0.014	0.116 ± 0.040	0.153
	USL-I	0.257 ± 0.038	0.149 ± 0.019	0.145 ± 0.049	0.263 ± 0.012	0.042 ± 0.006	0.101 ± 0.017	0.090 ± 0.042	0.150
	USL-M	0.209 ± 0.030	0.190 ± 0.052	0.144 ± 0.035	0.257 ± 0.010	0.058 ± 0.013	0.072 ± 0.037	0.111 ± 0.026	0.149
	Ours $f(\cdot)$	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
	Ours $g_I(\cdot)$	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
	Ours $g_C(\cdot)$	**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	Random	0.332 ± 0.044	0.340 ± 0.024	0.286 ± 0.049	0.371 ± 0.014	0.071 ± 0.015	0.150 ± 0.013	0.175 ± 0.024	0.246
	USL-I	0.422 ± 0.096	0.299 ± 0.038	0.275 ± 0.028	0.318 ± 0.006	0.074 ± 0.006	0.164 ± 0.020	0.179 ± 0.036	0.247
	USL-M	0.412 ± 0.077	0.365 ± 0.033	0.290 ± 0.034	0.310 ± 0.010	0.091 ± 0.008	0.132 ± 0.035	0.188 ± 0.025	0.255
	Ours $f(\cdot)$	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
	Ours $g_I(\cdot)$	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
	Ours $g_C(\cdot)$	**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	Random	0.482 ± 0.099	0.404 ± 0.042	0.278 ± 0.034	0.330 ± 0.016	0.118 ± 0.025	0.174 ± 0.024	0.229 ± 0.031	0.288
	USL-I	0.504 ± 0.103	0.359 ± 0.035	0.284 ± 0.035	0.298 ± 0.016	0.108 ± 0.010	0.206 ± 0.022	0.227 ± 0.024	0.284
	USL-M	0.505 ± 0.070	0.434 ± 0.029	0.304 ± 0.037	0.301 ± 0.015	0.128 ± 0.013	0.180 ± 0.036	0.221 ± 0.018	0.296
	Ours $f(\cdot)$	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
	Ours $g_I(\cdot)$	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
	Ours $g_C(\cdot)$	**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	Random	0.402 ± 0.045	0.421 ± 0.027	0.305 ± 0.033	0.334 ± 0.022	0.109 ± 0.014	0.198 ± 0.035	0.216 ± 0.032	0.284
	USL-I	0.499 ± 0.106	0.383 ± 0.029	0.285 ± 0.033	0.302 ± 0.013	0.113 ± 0.014	0.208 ± 0.019	0.243 ± 0.026	0.290
	USL-M	0.481 ± 0.087	0.461 ± 0.026	0.305 ± 0.038	0.298 ± 0.017	0.133 ± 0.013	0.184 ± 0.040	0.238 ± 0.024	0.300
	Ours $f(\cdot)$	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
	Ours $g_I(\cdot)$	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
	Ours $g_C(\cdot)$	**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**
Zero Shot BLIP-2		0.111 (0.100)	0.081 (0.027)	0.123 (0.021)	0.379 (0.010)	0.113 (0.010)	0.070 (0.010)	0.114 (0.010)	0.141

UP-DP outperforms the SOTA Methods

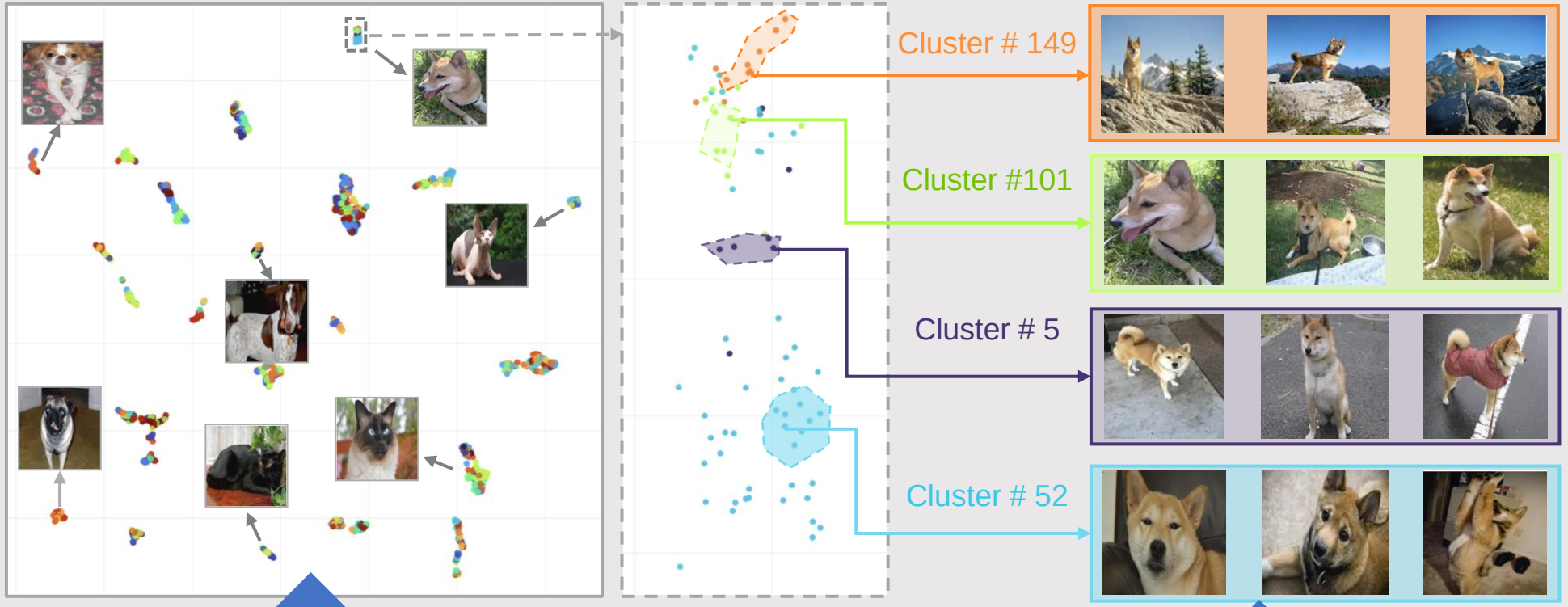
- $f(\cdot), g_I(\cdot), g_C(\cdot)$ represents different sampling strategies

Domain Generalization Results of Learned Prompts

Features	Caltech101	DTD	FGVCAircraft	Flowers102	OxfordPets	UCF101	EuroSAT	Average
Caltech101_Prompt	0.975	0.768	0.485	0.987	0.922	**0.885**	0.957	0.834
DTD_Prompt	**0.979**	**0.809**	0.482	0.990	0.889	0.878	0.964	0.836
FGVCAircraft_Prompt	0.974	0.768	0.518	0.986	0.852	0.862	0.963	0.829
Flowers102_Prompt	0.978	0.806	**0.583**	**0.995**	**0.946**	0.876	0.952	**0.864**
OxfordPets_Prompt	0.976	0.746	0.520	0.989	0.945	0.871	0.960	0.840
UCF101_Prompt	0.970	0.766	0.423	0.978	0.786	0.863	0.962	0.795
EuroSAT_Prompt	0.955	0.794	0.415	0.975	0.501	0.865	**0.969**	0.746
Empty_Prompt	0.756	0.501	0.322	0.719	0.492	0.744	0.933	0.606
Initi_Prompt	0.771	0.534	0.270	0.757	0.421	0.779	0.919	0.592
Image	0.974	0.763	0.349	0.984	0.681	0.875	0.961	0.760

Learned Prompt can be generalized across different datasets

Visualization of Features and Cluster Assignments



High level feature can distinguish between various types of pets

Cluster levels feature can capture more detailed information i.e., background and postures

Interpreting the Learned Prompts

EuroSAT	OxfordPets	DTD	Caltech101	FGVCAircraft	UCF101	Flowers102
makes	shark	several	learn	diamond	healthy	learn
wood	brigham	learn	butterfly	offers	learn	**butterfly**
takes	saint	putting	saint	learn	attends	angel
healthy	laying	add	add	**plane**	performs	saint
ku	elegant	six	three	del	speaks	personality
single	chicken	three	2020	river	newly	picking
have	posing	q	2019	**100**	physical	adding
november	kate	vector	2000	**40**	provides	gold
holds	bee	che	speaks	have	choose	**spider**
holding	attends	some	with	get	serves	giant

A few **words** is related to their respective task

Some shared words may lead to generalization

Majority of the words lack coherent meaning: Prompt encode meanings beyond the scope of the existing vocabulary

Ablation Study
Context Length

	Context Length	EuroSAT	OxfordPets	DTD	Caltech101	FGVCAircraft	UCF101	Flowers102	Average
RN50	16	0.359 ± 0.037	0.207 ± 0.006	0.246 ± 0.022	0.312 ± 0.002	0.058 ± 0.009	0.181 ± 0.014	0.104 ± 0.015	0.210
	4	**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**
RN101	16	0.335 ± 0.035	0.188 ± 0.017	0.226 ± 0.025	0.286 ± 0.003	0.049 ± 0.015	0.132 ± 0.012	0.087 ± 0.015	0.186
	4	**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**
ViTB32	16	0.462 ± 0.006	0.400 ± 0.023	0.320 ± 0.014	0.336 ± 0.007	0.077 ± 0.010	0.212 ± 0.019	0.185 ± 0.013	0.285
	4	**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**
ViTH14	16	0.494 ± 0.007	0.478 ± 0.018	0.348 ± 0.011	0.318 ± 0.004	0.101 ± 0.010	0.249 ± 0.018	0.200 ± 0.011	0.313
	4	**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**
ViTG14	16	0.501 ± 0.006	0.507 ± 0.015	0.352 ± 0.009	0.324 ± 0.007	0.108 ± 0.009	0.251 ± 0.019	0.210 ± 0.010	0.322
	4	**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**

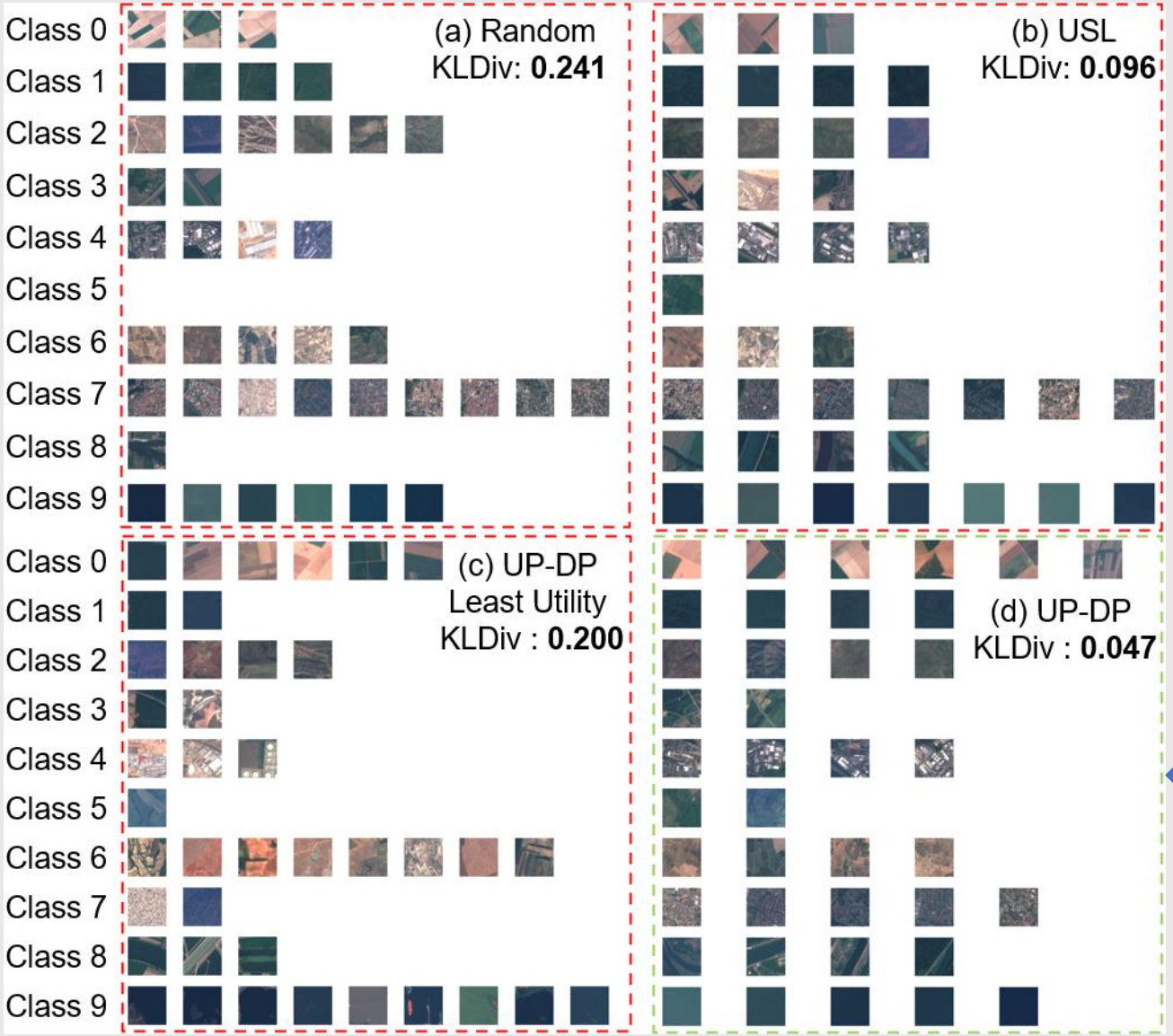
Short context length is better
Long context length may cause overfitting

Ablation Study
Annotation Budget

	Method	EuroSAT	OxfordPets	DTD	Caltech101	FGVCAircraft	UCF101	Flowers102	Average
RN50	Random	0.323 ± 0.069	0.292 ± 0.023	0.270 ± 0.008	0.362 ± 0.010	0.062 ± 0.005	0.175 ± 0.041	0.151 ± 0.011	0.234
	USL-I	0.389 ± 0.042	0.234 ± 0.041	0.257 ± 0.017	0.349 ± 0.013	0.070 ± 0.007	0.188 ± 0.014	0.153 ± 0.026	0.234
	USL-M	0.386 ± 0.047	0.298 ± 0.032	0.269 ± 0.043	0.340 ± 0.021	0.075 ± 0.012	0.188 ± 0.009	0.133 ± 0.031	0.241
	Ours	**0.398 ± 0.010**	**0.307 ± 0.007**	**0.342 ± 0.010**	**0.399 ± 0.005**	**0.084 ± 0.001**	**0.245 ± 0.006**	**0.187 ± 0.007**	**0.280**
RN101	Random	0.291 ± 0.049	0.251 ± 0.028	0.216 ± 0.031	0.319 ± 0.016	0.058 ± 0.012	0.142 ± 0.028	0.136 ± 0.018	0.202
	USL-I	0.325 ± 0.047	0.225 ± 0.047	0.205 ± 0.027	0.291 ± 0.012	0.066 ± 0.011	0.144 ± 0.014	0.124 ± 0.028	0.197
	USL-M	0.319 ± 0.062	**0.298 ± 0.047**	0.213 ± 0.031	0.294 ± 0.018	0.064 ± 0.018	0.149 ± 0.017	0.109 ± 0.026	0.207
	Ours	**0.325 ± 0.000**	0.294 ± 0.005	**0.305 ± 0.010**	**0.345 ± 0.004**	**0.077 ± 0.000**	**0.224 ± 0.005**	**0.159 ± 0.005**	**0.247**
ViTB32	Random	0.438 ± 0.066	0.488 ± 0.070	0.400 ± 0.024	0.427 ± 0.019	0.085 ± 0.007	0.201 ± 0.011	0.270 ± 0.040	0.330
	USL-I	0.496 ± 0.039	0.438 ± 0.030	0.377 ± 0.017	0.393 ± 0.017	0.097 ± 0.010	0.223 ± 0.023	0.263 ± 0.028	0.327
	USL-M	**0.552 ± 0.054**	**0.524 ± 0.034**	0.409 ± 0.016	0.390 ± 0.021	0.102 ± 0.020	0.223 ± 0.011	0.237 ± 0.035	0.348
	Ours	0.541 ± 0.006	0.488 ± 0.019	**0.427 ± 0.001**	**0.448 ± 0.004**	**0.104 ± 0.004**	**0.274 ± 0.003**	**0.309 ± 0.001**	**0.370**
ViTH14	Random	0.505 ± 0.110	0.560 ± 0.029	0.421 ± 0.024	0.425 ± 0.009	0.152 ± 0.012	0.282 ± 0.027	0.290 ± 0.038	0.376
	USL-I	0.585 ± 0.052	0.518 ± 0.036	0.408 ± 0.017	0.369 ± 0.015	0.147 ± 0.011	0.301 ± 0.026	0.313 ± 0.040	0.377
	USL-M	**0.647 ± 0.057**	**0.629 ± 0.032**	0.435 ± 0.020	0.372 ± 0.025	0.160 ± 0.027	0.292 ± 0.008	0.294 ± 0.039	0.404
	Ours	0.615 ± 0.028	0.600 ± 0.018	**0.446 ± 0.005**	**0.435 ± 0.006**	**0.148 ± 0.009**	**0.323 ± 0.008**	**0.327 ± 0.001**	**0.413**
ViTG14	Random	0.567 ± 0.062	0.592 ± 0.039	0.440 ± 0.026	0.413 ± 0.008	0.165 ± 0.026	0.279 ± 0.023	0.341 ± 0.023	0.400
	USL-I	0.578 ± 0.044	0.546 ± 0.032	0.411 ± 0.017	0.372 ± 0.019	0.157 ± 0.010	0.312 ± 0.025	0.336 ± 0.037	0.387
	USL-M	**0.627 ± 0.052**	**0.658 ± 0.031**	0.436 ± 0.017	0.371 ± 0.023	**0.171 ± 0.023**	0.303 ± 0.004	0.316 ± 0.034	0.412
	Ours	0.614 ± 0.007	0.636 ± 0.014	**0.445 ± 0.007**	**0.434 ± 0.005**	0.165 ± 0.011	**0.326 ± 0.009**	**0.342 ± 0.002**	**0.423**

Our method still outperform others under larger annotation budget setting

Visualization on Selected Samples



More balanced Selection

Thank you for reviewing!

Back-up: Datasets Statistics

Dataset	Classes	Train	Val	Test	Hand-crafted prompt
EuroSAT	10	13,500	5,400	8,100	"a centered satellite photo of [CLASS]."
OxfordPets	37	2,944	736	3,669	"a photo of a [CLASS], a type of pet."
DTD	47	2,820	1,128	1,692	"[CLASS] texture."
Caltech101	100	4,128	1,649	2,465	"a photo of a [CLASS]."
FGVCAircraft	100	3,334	3,333	3,333	"a photo of a [CLASS], a type of aircraft."
UCF101	101	7,639	1,898	3,783	"a photo of a person doing [CLASS]."
Flowers102	102	4,093	1,633	2,463	"a photo of a [CLASS], a type of flower."