The Benefits of Balance

From Information Projections to Variance Reduction

NeurIPS 2024

Team

Lang Liu
University of Washington

Ronak Mehta
University of Washington

Soumik Pal
University of Washington

Zaid Harchaoui
University of Washington

IFDS
Institute for Foundations of Data Science

NSF

IFML

UNIVERSITY of WASHINGTON

Learning Transferable Visual Models From Natural Language Supervision

Alec Radford [*1] Jong Wook Kim [*1] Chris Hallacy [1] Aditya Ramesh [1] Gabriel Goh [1] Sandhini Agarwal [1]
Girish Sastry [1] Amanda Askell [1] Pamela Mishkin [1] Jack Clark [1] Gretchen Krueger [1] Ilya Sutskever [1]

Unsupervised Learning of Visual Features by Contrasting Cluster Assignments

Mathilde Caron[1,2] Ishan Misra[2] Julien Mairal[1]
Priya Goyal[2] Piotr Bojanowski[2] Armand Joulin[2]
[1] Inria* [2] Facebook AI Research

SELF-LABELLING VIA SIMULTANEOUS CL
AND REPRESENTATION LEARNING

Yuki M. Asano Christian Rupprecht Andrea Vedaldi
Visual Geometry Group
ots.ox.ac.uk

DEMYSTIFYING CLIP DATA

Hu Xu[1] Saining Xie[2] Xiaoqing Ellen Tan[1] Po-Yao Huang[1] Russell Howes[1] Vasu Sharma[1]
Shang-Wen Li[1] Gargi Ghosh[1] Luke Zettlemoyer[1,3] Christoph Feichtenhofer[1]
[1]FAIR, Meta AI [2]New York University [3]University of Washington

DINOv2: Learning Robust Visual Features without Supervision

Maxime Oquab**, Timothée Darcet**, Théo Moutakanni**,
iec*, Vasil Khalidov*, Pierre Fernandez, Daniel Haziza,
Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba,
Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat,
l Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal[1],
tut*, Armand Joulin*, Piotr Bojanowski*
Meta AI Research [1]Inria
core team **equal contribution

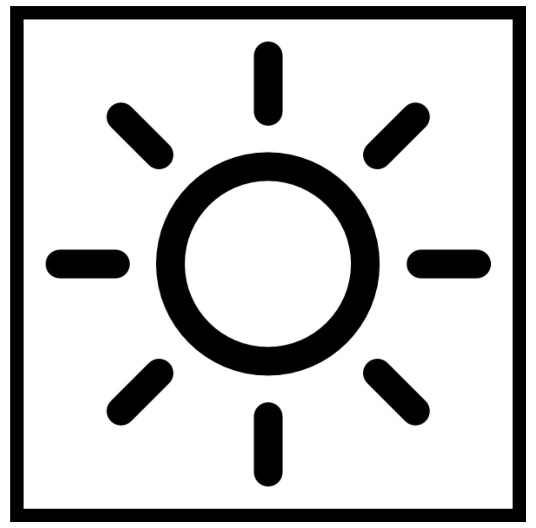
DATACOMP:
In search of the next generation of multimodal datasets

adre*[2], Gabriel Ilharco*[1], Alex Fang*[1], Jonathan Hayase[1],
nis[5], Thao Nguyen[1], Ryan Marten[7,9], Mitchell Wortsman[1],
eyu Zhang[1], Eyal Orgad[3], Rahim Entezari[10], Giannis Daras[5],
Vivek Ramanujan[1], Yonatan Bitton[11], Kalyani Marathe[1],
Stephen Mussmann[1], Richard Vencu[6], Mehdi Cherti[6,8], Ranjay Krishna[1],
Pang Wei Koh[1,12], Olga Saukh[10], Alexander Ratner[1,13], Shuran Song[2],
Hannaneh Hajishirzi[1,7], Ali Farhadi[1], Romain Beaumont[6],
Sewoong Oh[1], Alex Dimakis[5], Jenia Jitsev[6,8],
Yair Carmon[3], Vaishaal Shankar[4], Ludwig Schmidt[1,6,7]

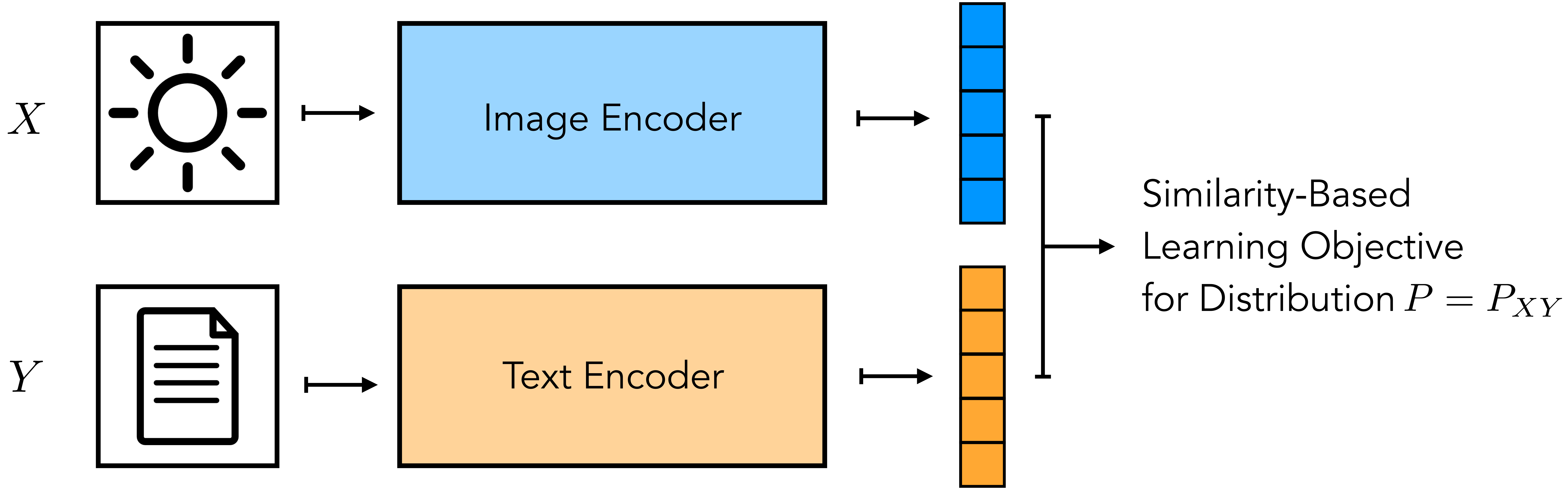
Discriminative clustering with representation learning with any ratio of labeled to unlabeled data

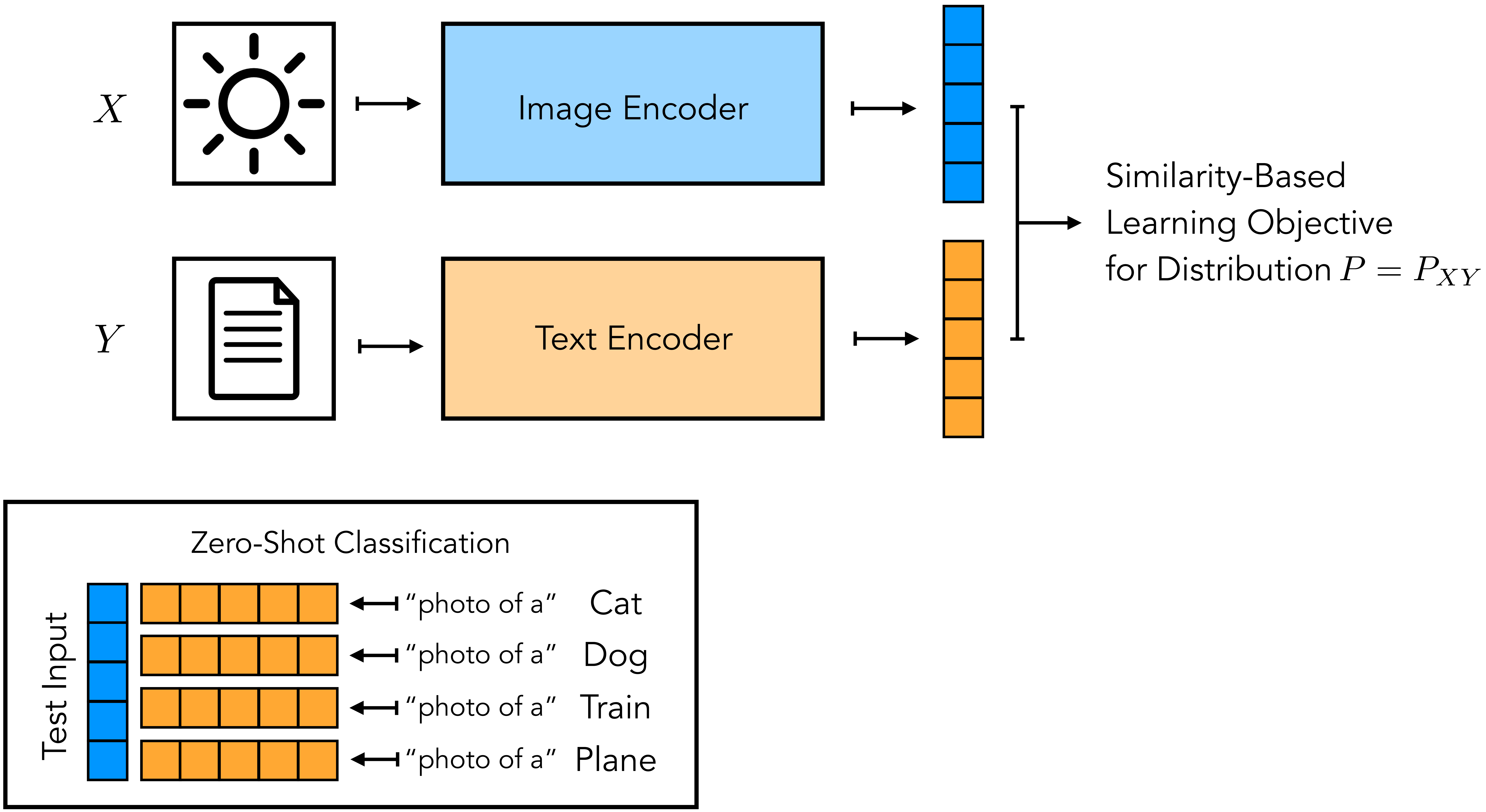
Corinne Jones[1] · Vincent Roulet[2] · Zaid Harchaoui[2]

X

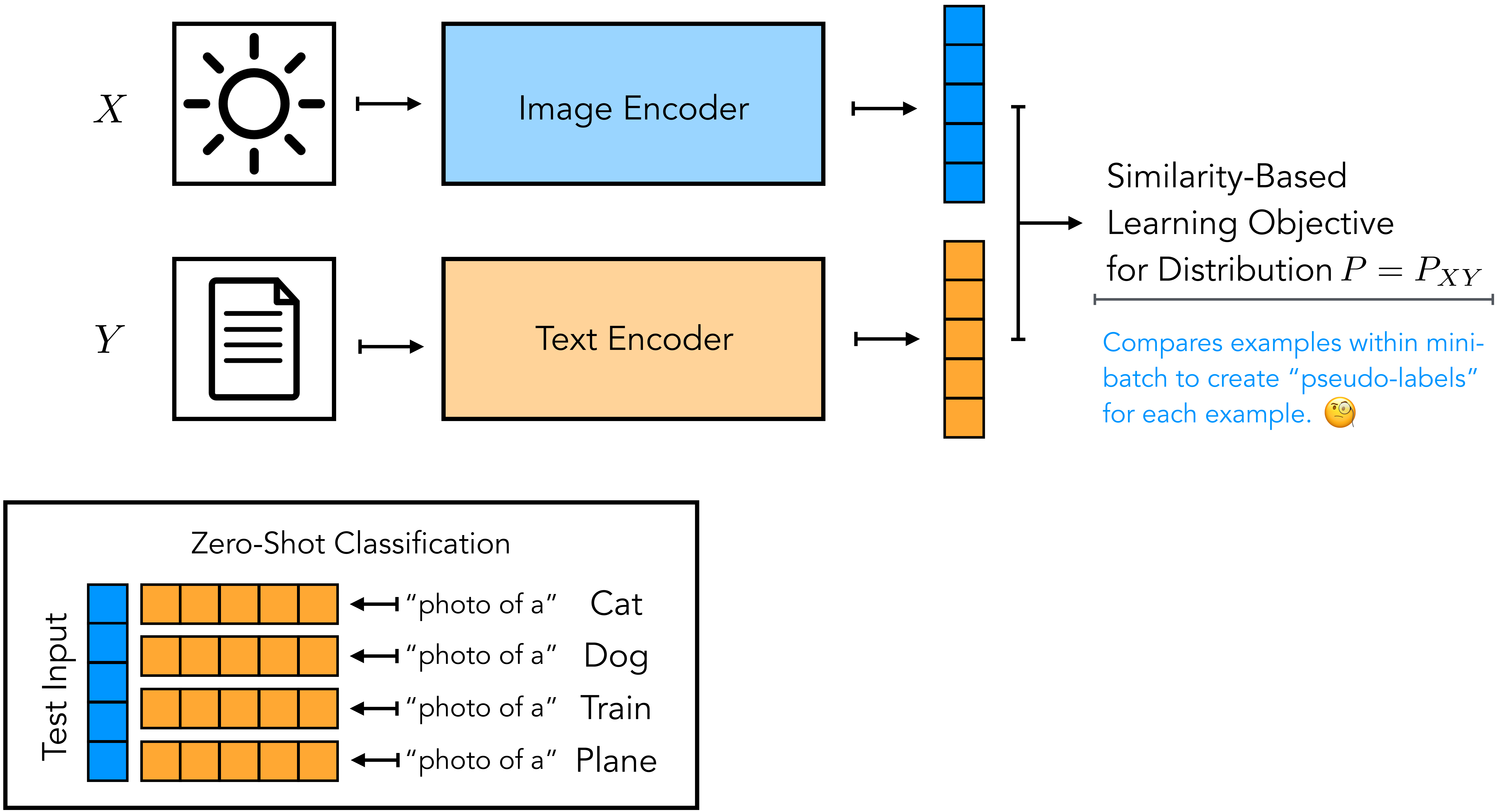


Y

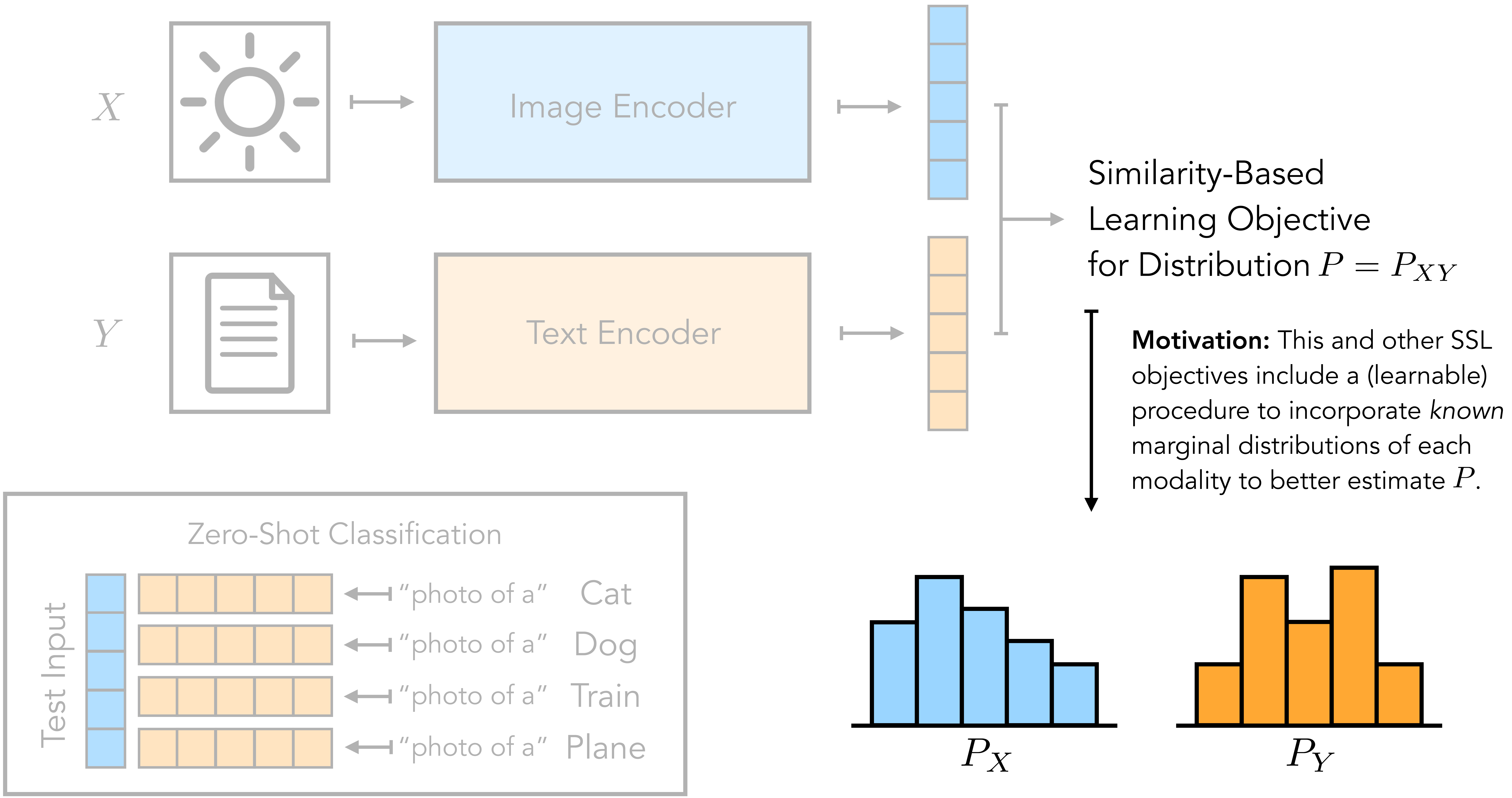

X
Image Encoder
Y
Text Encoder
Similarity-Based
Learning Objective
for Distribution $P = P_{XY}$



X
Image Encoder
Y
Text Encoder
Similarity-Based
Learning Objective
for Distribution $P = P_{XY}$
Zero-Shot Classification
Test Input
"photo of a" Cat
"photo of a" Dog
"photo of a" Train
"photo of a" Plane



X
Image Encoder
Y
Text Encoder
Similarity-Based
Learning Objective
for Distribution $P = P_{XY}$
Compares examples within mini-batch to create "pseudo-labels" for each example.
Zero-Shot Classification
Test Input
"photo of a" Cat
"photo of a" Dog
"photo of a" Train
"photo of a" Plane



X
Image Encoder
Y
Text Encoder
Similarity-Based Learning Objective for Distribution $P = P_{XY}$
Motivation: This and other SSL objectives include a (learnable) procedure to incorporate *known* marginal distributions of each modality to better estimate P.
Zero-Shot Classification
Test Input
"photo of a" Cat
"photo of a" Dog
"photo of a" Train
"photo of a" Plane
P_X
P_Y

Problem Setting

Data from **unknown** joint probability distribution.

$$(X_1, Y_1), \ldots, (X_n, Y_n) \sim P$$

Access to **known** marginal distributions.

$$(P_X, P_Y)$$

Goal: estimate the parameter:

$$P(h) = \mathbb{E}_{(X,Y)\sim P}\left[h(X, Y)\right]$$

and characterize how the marginals improve upon

$$P_n(h) = \frac{1}{n}\sum_{i=1}^{n} h(X_i, Y_i)$$

Problem Setting

Data from **unknown** joint probability distribution.

$$(X_1, Y_1), \ldots, (X_n, Y_n) \sim P$$

Access to **known** marginal distributions.

$$(P_X, P_Y)$$

Goal: estimate the parameter:

$$P(h) = \mathbb{E}_{(X,Y)\sim P}\left[h(X,Y)\right]$$

and characterize how the marginals improve upon

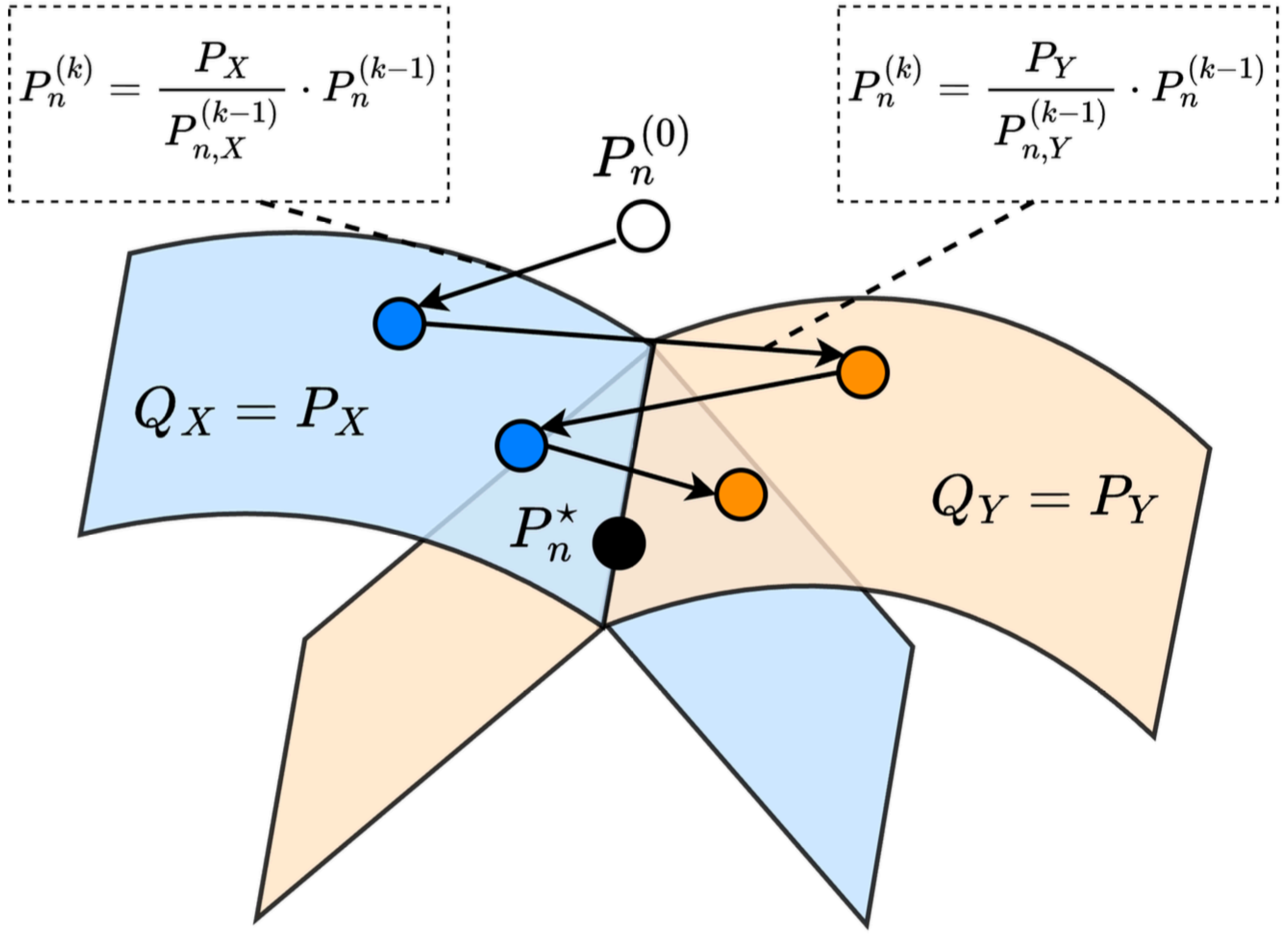
$$P_n(h) = \frac{1}{n}\sum_{i=1}^{n} h(X_i, Y_i)$$

Data Balancing

Rescale rows and columns by the desired marginals.

$$P_n^{(0)} = P_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{(X_i, Y_i)}$$

$$P_n^{(k)} = \begin{cases} \frac{P_X}{P_{n,X}^{(k-1)}} \cdot P_n^{(k-1)} & k \text{ odd} \\ \frac{P_Y}{P_{n,Y}^{(k-1)}} \cdot P_n^{(k-1)} & k \text{ even} \end{cases}$$



Problem Setting

Data from **unknown** joint probability distribution.

$$(X_1, Y_1), \ldots, (X_n, Y_n) \sim P$$

Access to **known** marginal distributions.

$$(P_X, P_Y)$$

Goal: estimate the parameter:

$$P(h) = \mathbb{E}_{(X,Y)\sim P}\left[h(X, Y)\right]$$

and characterize how the marginals improve upon

$$P_n(h) = \frac{1}{n}\sum_{i=1}^{n} h(X_i, Y_i)$$

Data Balancing

Rescale rows and columns by the desired marginals.

$$P_n^{(0)} = P_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{(X_i, Y_i)}$$

$$P_n^{(k)} = \begin{cases} \frac{P_X}{P_{n,X}^{(k-1)}} \cdot P_n^{(k-1)} & k \text{ odd} \\ \frac{P_Y}{P_{n,Y}^{(k-1)}} \cdot P_n^{(k-1)} & k \text{ even} \end{cases}$$

1. How does the balanced distribution improve upon the empirical measure **theoretically**?
2. What are the **practical** implications for SSL objectives such as CLIP?

Problem Setting

Data from **unknown** joint probability distribution.

$$(X_1, Y_1), \ldots, (X_n, Y_n) \sim P$$

Access to **known** marginal distributions.

$$(P_X, P_Y)$$

Goal: estimate the parameter:

$$P(h) = \mathbb{E}_{(X,Y)\sim P}\left[h(X,Y)\right]$$

and characterize how the marginals improve upon

$$P_n(h) = \frac{1}{n}\sum_{i=1}^{n} h(X_i, Y_i)$$

Data Balancing

Rescale rows and columns by the desired marginals.

$$P_n^{(0)} = P_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{(X_i, Y_i)}$$

$$P_n^{(k)} = \begin{cases} \frac{P_X}{P_{n,X}^{(k-1)}} \cdot P_n^{(k-1)} & k \text{ odd} \\ \frac{P_Y}{P_{n,Y}^{(k-1)}} \cdot P_n^{(k-1)} & k \text{ even} \end{cases}$$

Theorem. The iterates of balancing satisfy

$$\mathbb{E}_P \left| P_n^{(k)}(h) - P(h) \right|^2 = \frac{\sigma_k^2}{n} + \tilde{O}\left(\frac{k^6}{n^{3/2}}\right) \to \frac{\sigma_0^2 - \sigma_{\text{gap}}^2}{n}$$

Novel recursion formula for estimation error that is of independent interest (OT, data-centric ML, etc.)

Problem Setting

Data from **unknown** joint probability distribution.

$$(X_1, Y_1), \ldots, (X_n, Y_n) \sim P$$

Access to **known** marginal distributions.

$$(P_X, P_Y)$$

Goal: estimate the parameter:

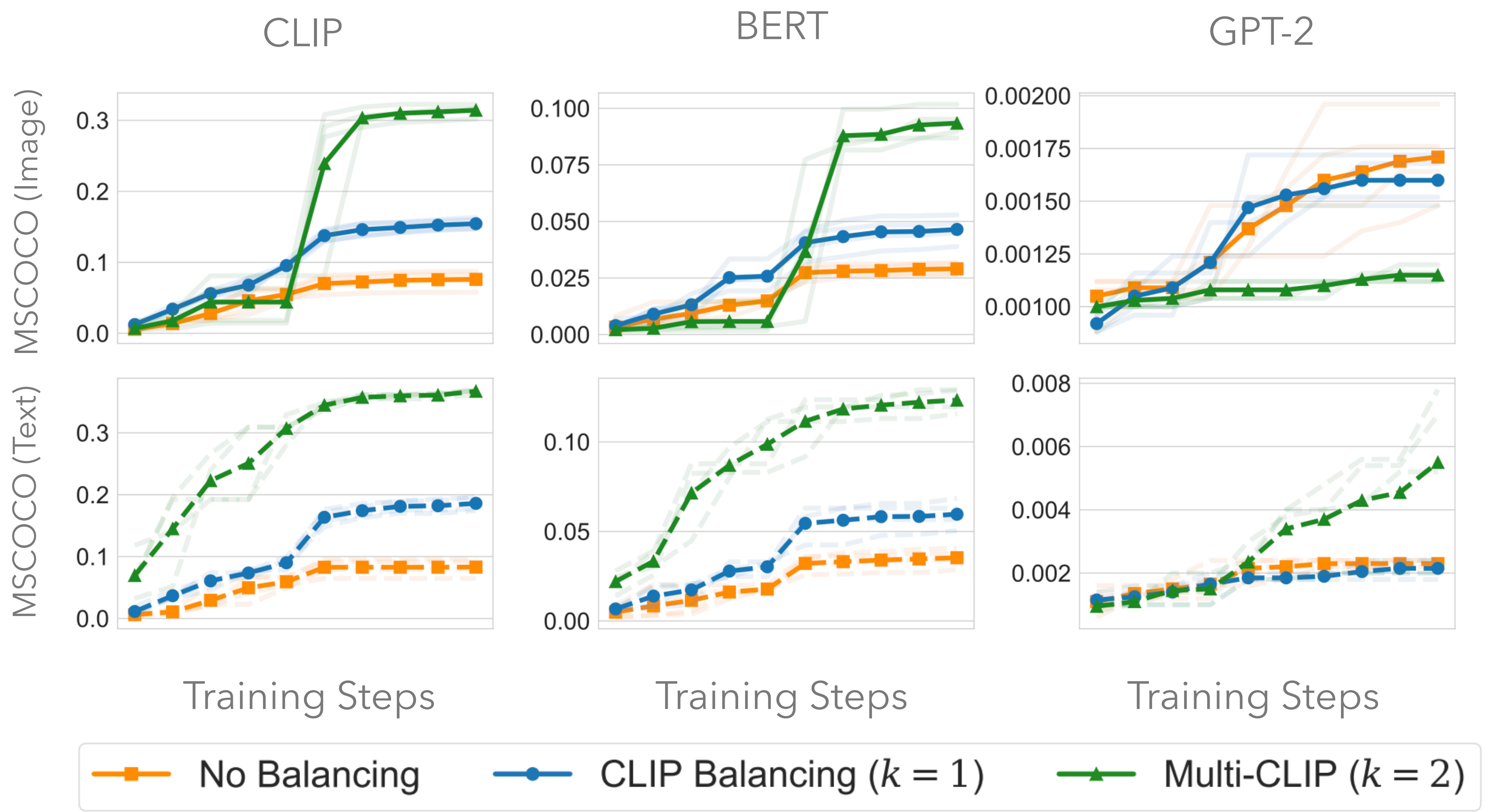
$$P(h) = \mathbb{E}_{(X,Y)\sim P}\left[h(X,Y)\right]$$

and characterize how the marginals improve upon

$$P_n(h) = \frac{1}{n}\sum_{i=1}^{n} h(X_i, Y_i)$$

Balancing mini-batches to improve the stability of the CLIP training objective.

Using a balanced objective increases zero-shot retrieval (recall) across datasets and embedding architectures.



Problem Setting

Data from **unknown** joint probability distribution.

$$(X_1, Y_1), \ldots, (X_n, Y_n) \sim P$$

Access to **known** marginal distributions.

$$(P_X, P_Y)$$

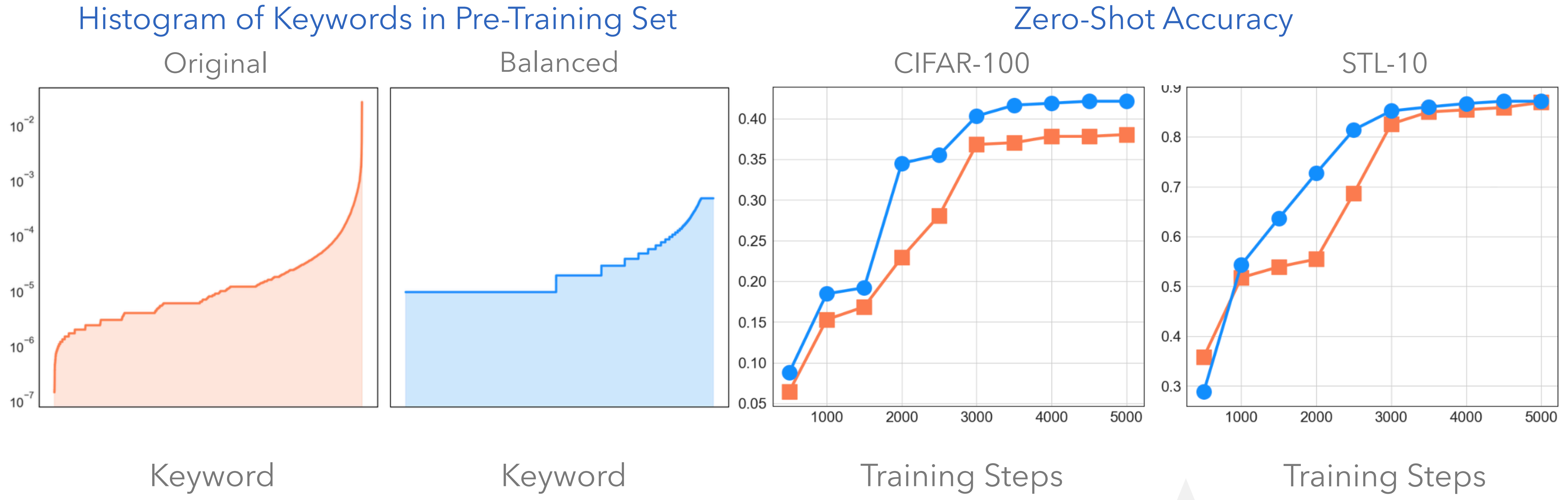
Goal: estimate the parameter:

$$P(h) = \mathbb{E}_{(X,Y)\sim P}\left[h(X,Y)\right]$$

and characterize how the marginals improve upon

$$P_n(h) = \frac{1}{n}\sum_{i=1}^{n} h(X_i, Y_i)$$

CLIP models trained on the balanced pre-training set improve over those trained on the original.



Balancing at scale improves performance on zero-shot classification.

Problem Setting

Data from **unknown** joint probability distribution.

$$(X_1, Y_1), \ldots, (X_n, Y_n) \sim P$$

Access to **known** marginal distributions.

$$(P_X, P_Y)$$

Goal: estimate the parameter:

$$P(h) = \mathbb{E}_{(X,Y)\sim P}\left[h(X, Y)\right]$$

and characterize how the marginals improve upon

$$P_n(h) = \frac{1}{n}\sum_{i=1}^{n} h(X_i, Y_i)$$

Thank you!