

A Generalization Theory for Zero-Shot Prediction

International Conference on Machine Learning (ICML): Paper #4085
July 16, 2025



Ronak Mehta and Zaid Harchaoui

The Mystery of Foundation Models

Learning Transferable Visual Models From Natural Language Supervision

Alec Radford^{*1} Jong Wook Kim^{*1} Chris Hallacy¹ Aditya Ramesh¹ Gabriel Goh¹ Sandhini Agarwal¹
Girish Sastry¹ Amanda Askell¹ Pamela Mishkin¹ Jack Clark¹ Gretchen Krueger¹ Ilya Sutskever¹

GPT-4 Technical Report



DeepSeek-R1: Incentivizing Reasoning Capabilities Reinforcement Learning

DeepSeek-AI

research@deepseek.com

GPT-4.



The Llama 3 Herd of Models

Llama Team, AI @ Meta¹

¹A detailed contributor list can be found in the appendix of this paper.

Modern artificial intelligence (AI) systems are powered by foundation models. This paper presents a new set of foundation models, called Llama 3. It is a herd of language models that natively support multilinguality, coding, reasoning, and tool usage. Our largest model is a dense Transformer with 405B parameters and a context window of up to 128K tokens. This paper presents an extensive empirical evaluation of Llama 3. We find that Llama 3 delivers comparable quality to leading language models such as GPT-4 on a plethora of tasks. We publicly release Llama 3, including pre-trained and post-trained versions of the 405B parameter language model and our Llama Guard 3 model for input and output safety. The paper also presents the results of experiments in which we integrate image, video, and speech capabilities into Llama 3 via a compositional approach. We observe this approach performs competitively with the state-of-the-art on image, video, and speech recognition tasks. The resulting models are not yet being broadly released as they are still under development.

Date: July 23, 2024

Website: <https://llama.meta.com/>

Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture

Mahmoud Assran^{1,2,3*} Quentin Duval¹ Ishan Misra¹ Piotr Bojanowski¹
Pascal Vincent¹ Michael Rabbat^{1,3} Yann LeCun^{1,4} Nicolas Ballas¹

¹Meta AI (FAIR) ²McGill University ³ Mila, Quebec AI Institute ⁴New York University

DINOv2: Learning Robust Visual Features without Supervision

Iaxime Oquab**, Timothée Darcet**, Théo Moutakanni**,
Marc Szafraniec*, Vasil Khalidov*, Pierre Fernandez, Daniel Haziza,
Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba,
wes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat,
arma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal¹,
Patrick Labatut*, Armand Joulin*, Piotr Bojanowski*

Meta AI Research ¹Inria

*core team **equal contribution

The Mystery of Foundation Models

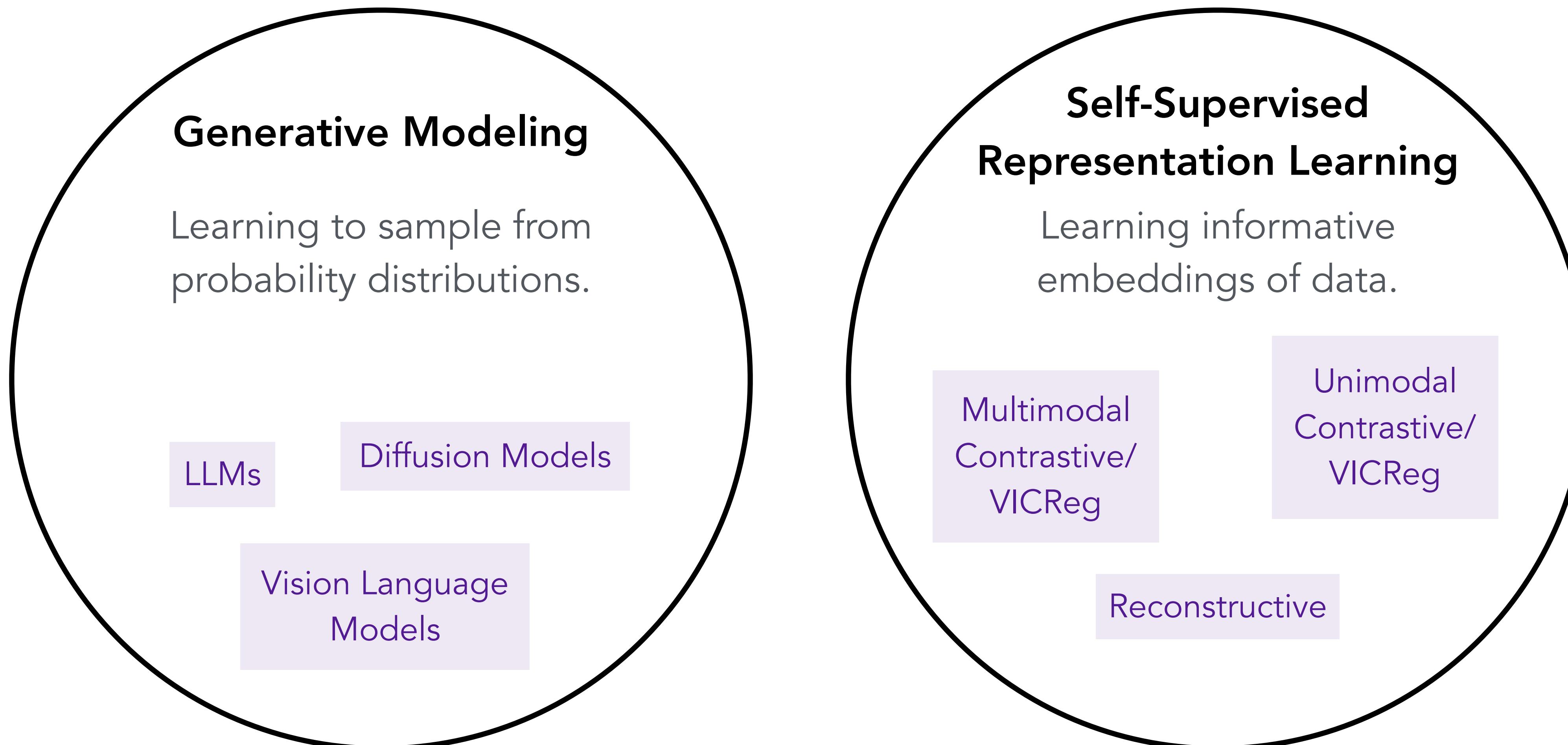
Generative Modeling

Learning to sample from probability distributions.

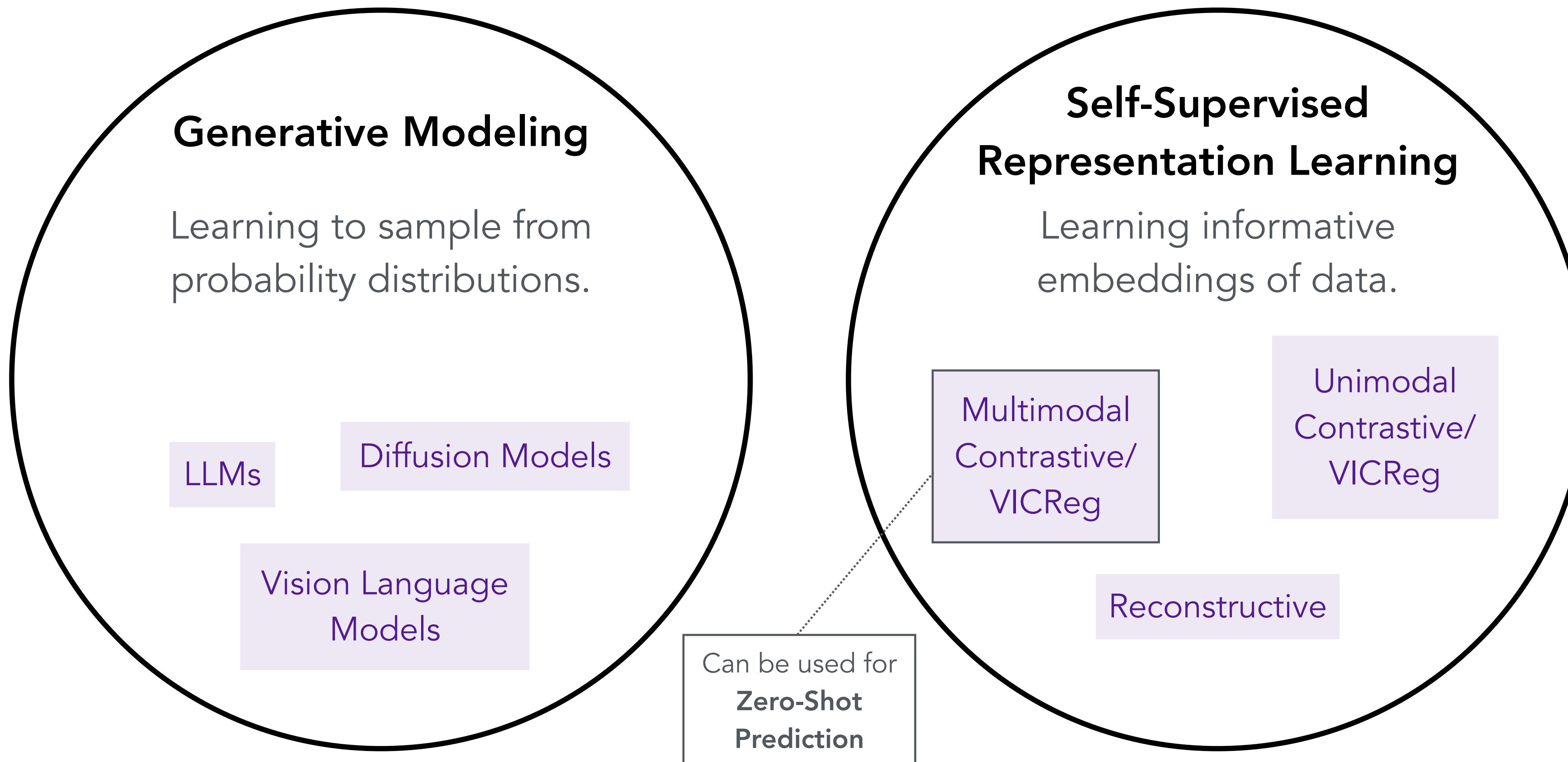
Self-Supervised Representation Learning

Learning informative embeddings of data.

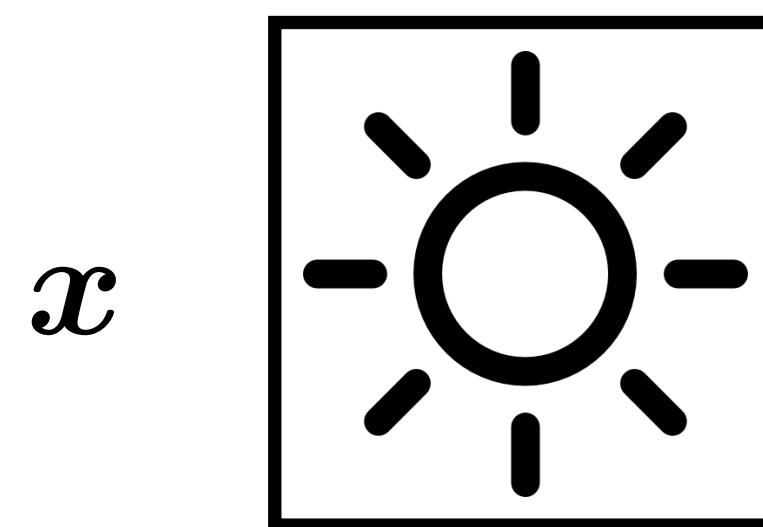
The Mystery of Foundation Models



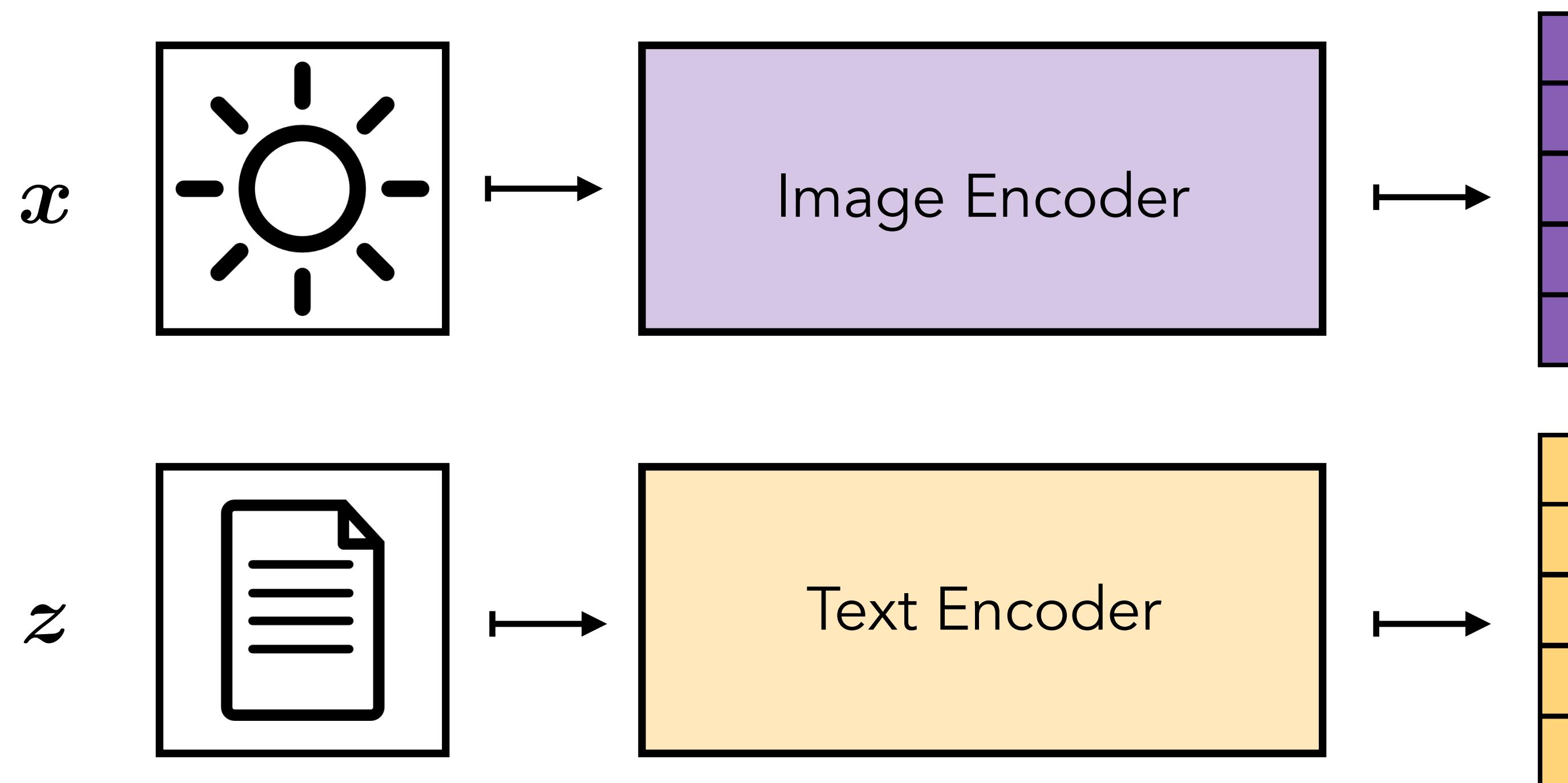
The Mystery of Foundation Models



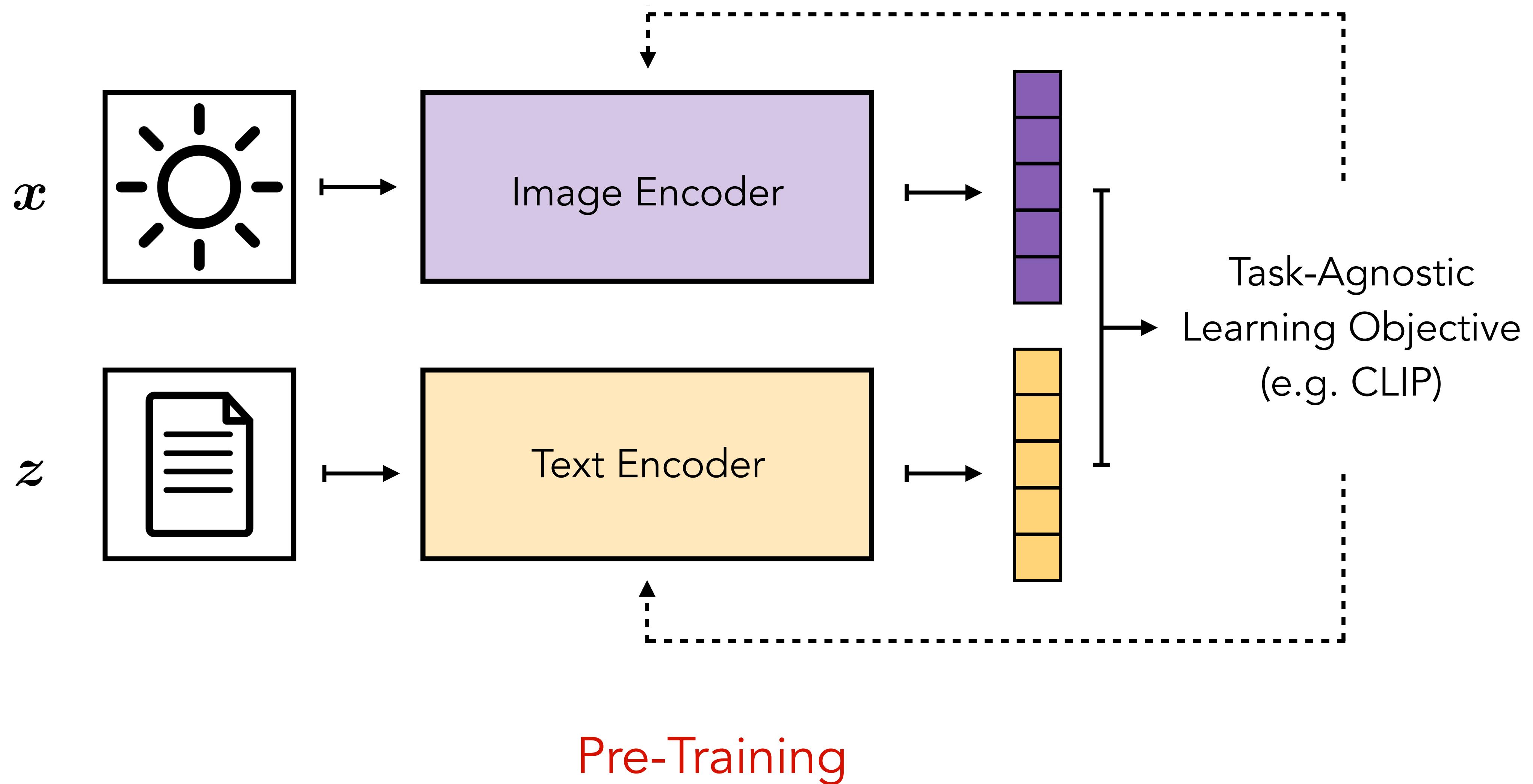
Foundation Models and Zero-Shot Prediction



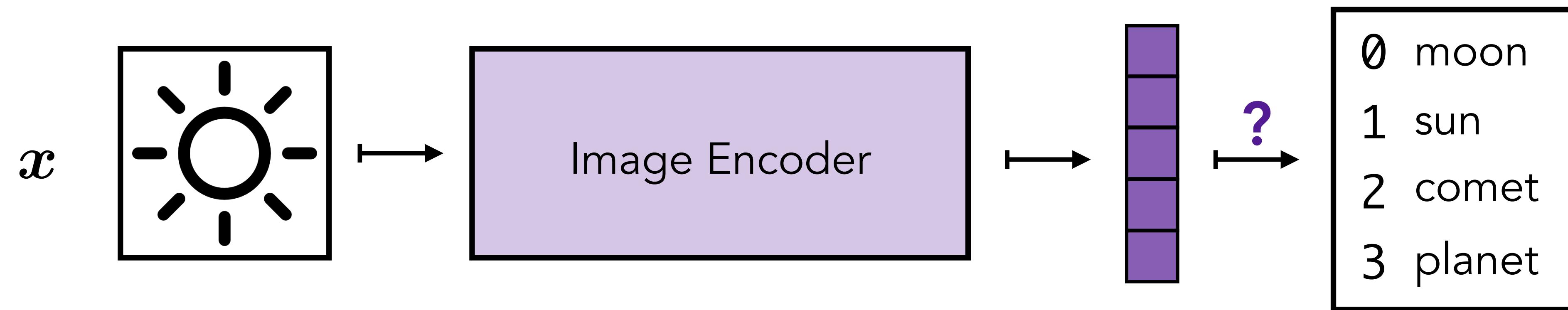
Foundation Models and Zero-Shot Prediction



Foundation Models and Zero-Shot Prediction

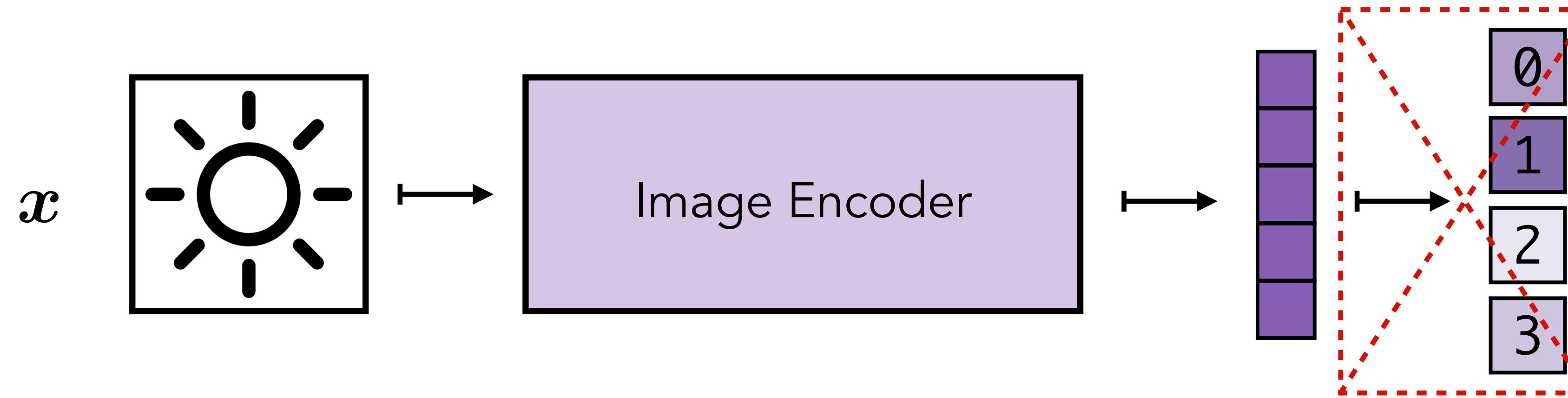


Foundation Models and Zero-Shot Prediction



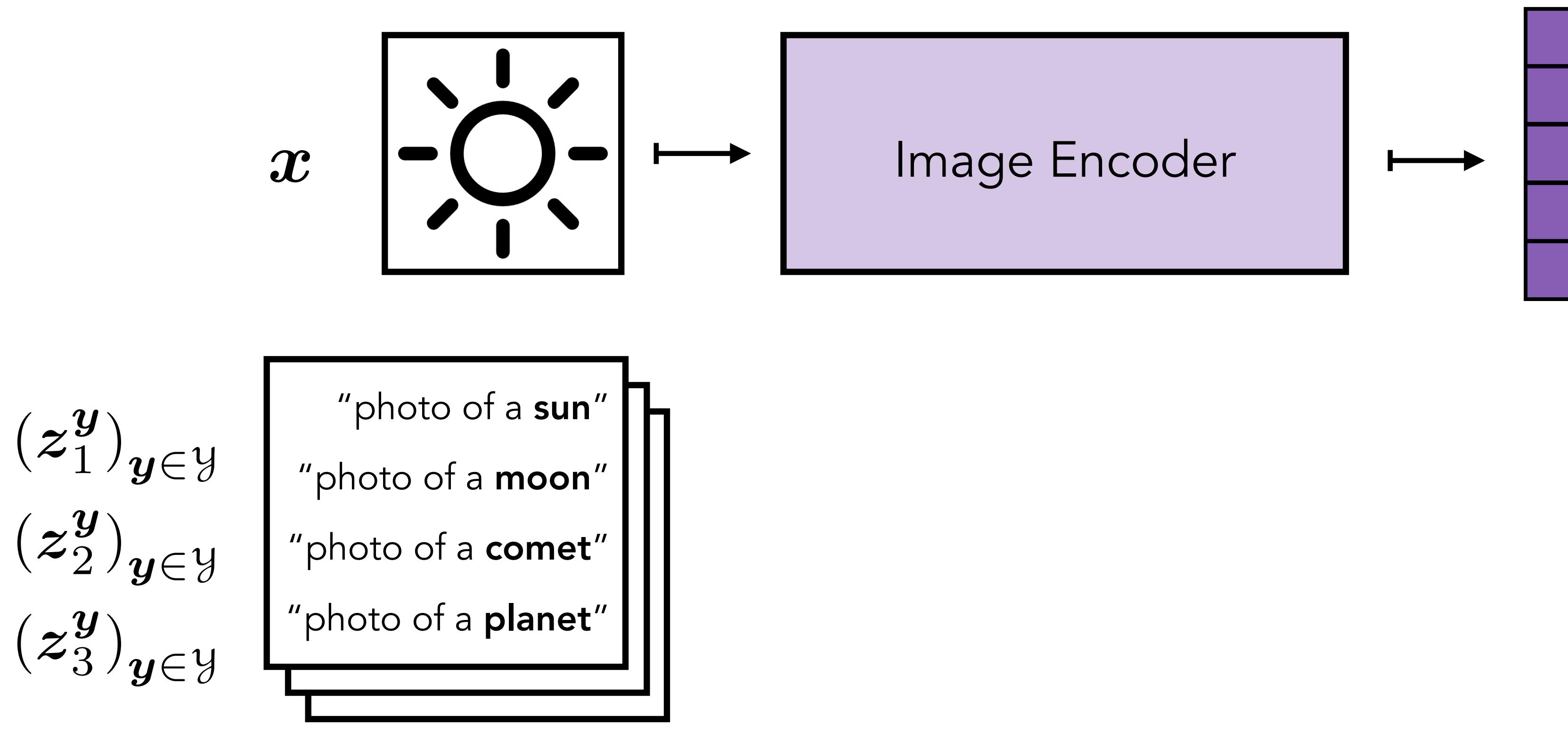
Evaluation

Foundation Models and Zero-Shot Prediction



Evaluation

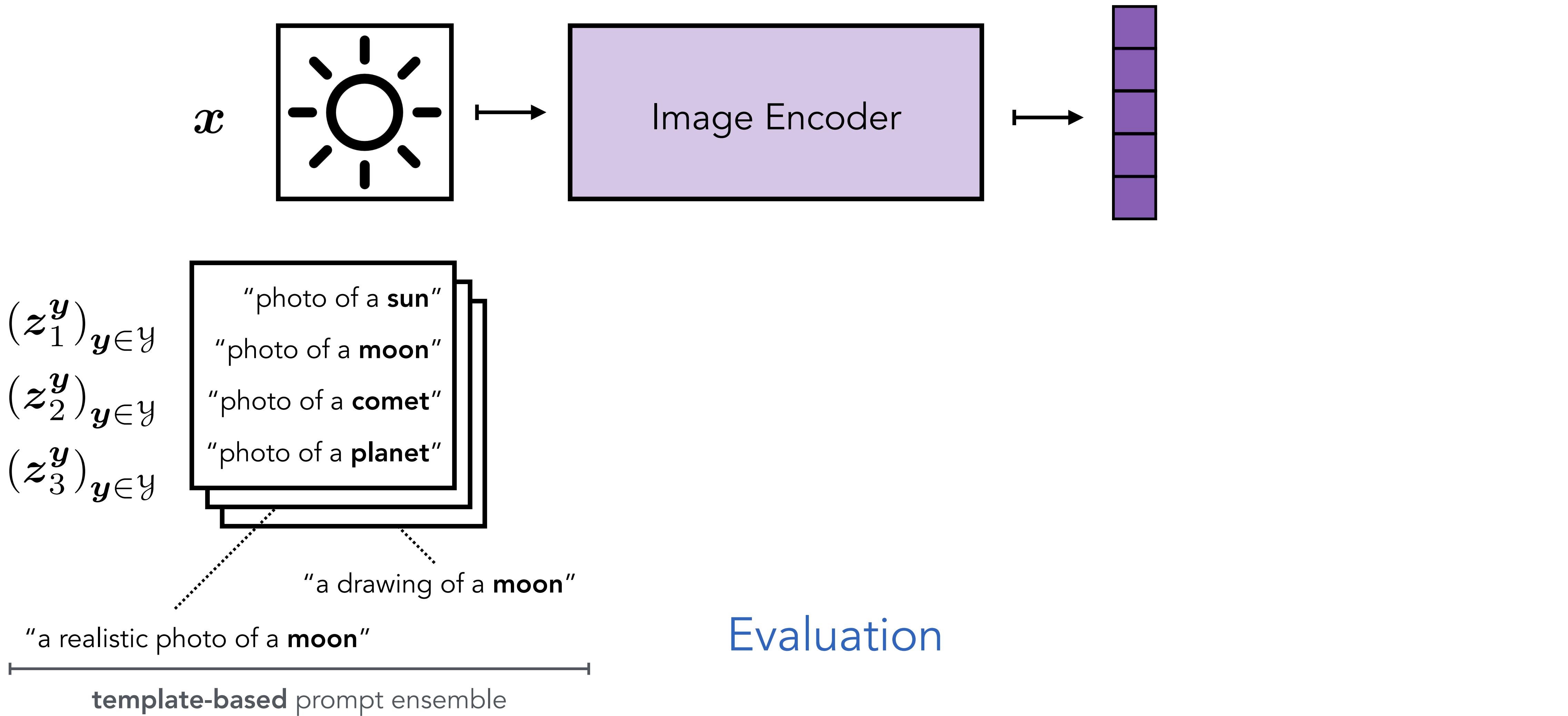
Foundation Models and Zero-Shot Prediction



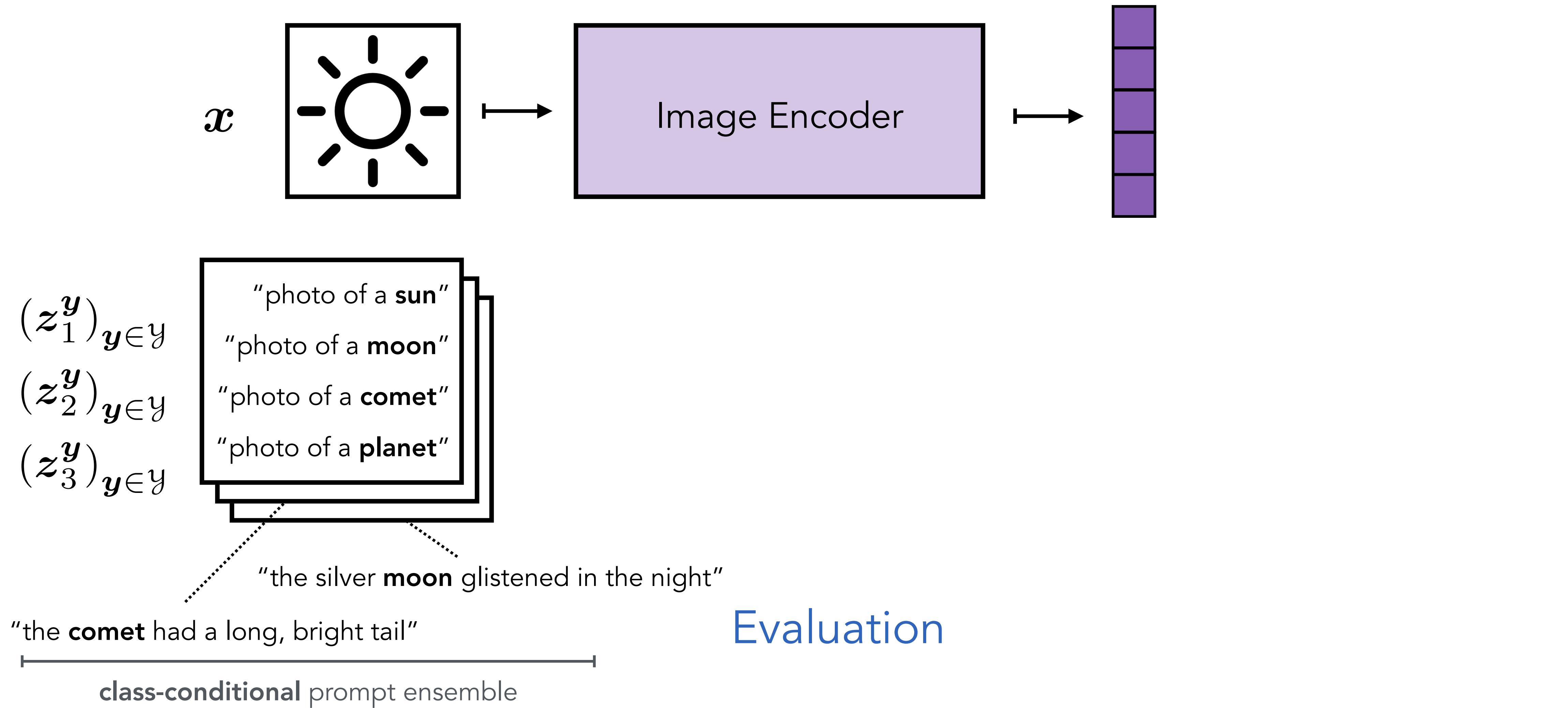
Idea: Convert labels into
prompts (pseudo-captions)

Evaluation

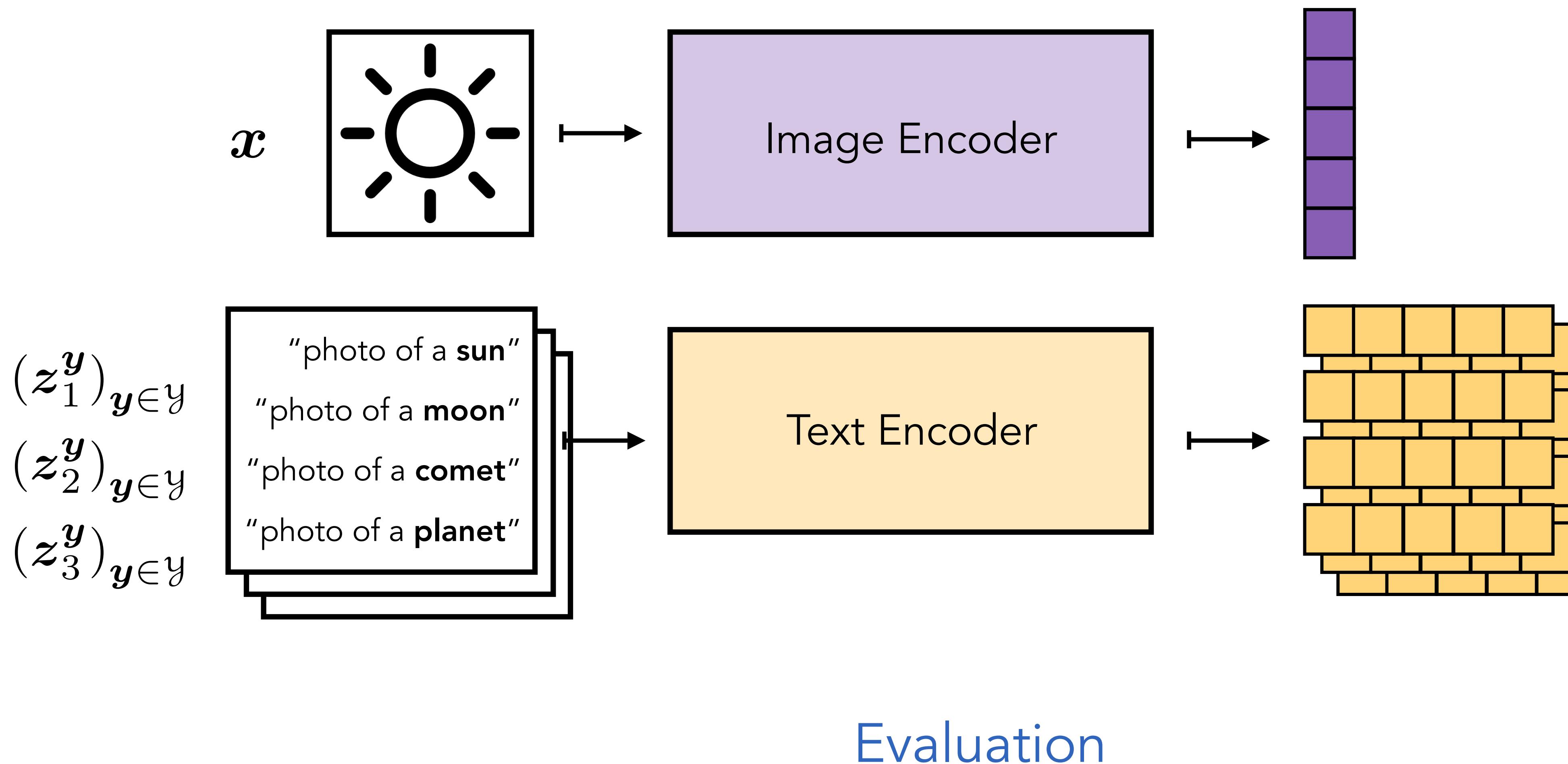
Foundation Models and Zero-Shot Prediction



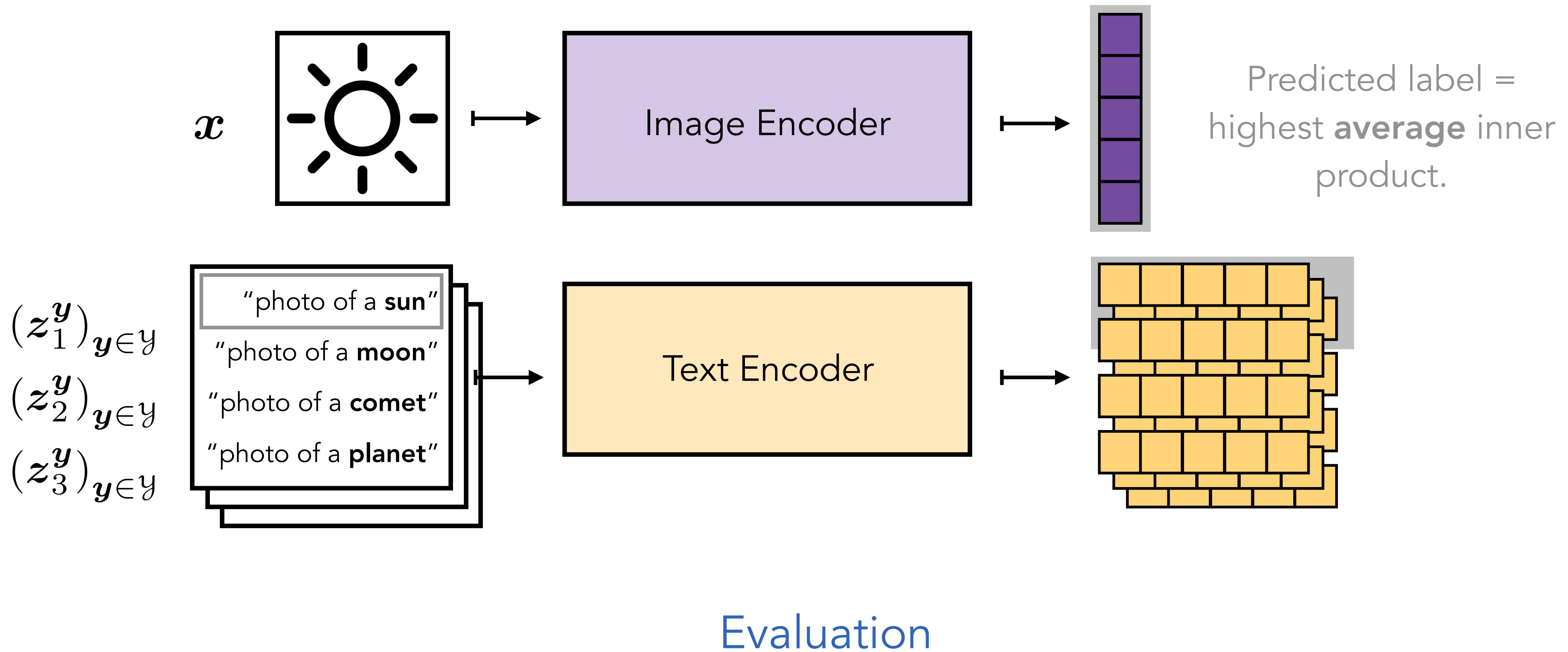
Foundation Models and Zero-Shot Prediction



Foundation Models and Zero-Shot Prediction



Foundation Models and Zero-Shot Prediction



Context

- **Pre-training/prompting techniques** advanced significantly in applications.
- How do we understand/perform theoretical **generalization analysis of ZSP?**

**UNDERSTANDING TRANSFERABLE REPRESENTATION
LEARNING AND ZERO-SHOT TRANSFER IN CLIP**

Zixiang Chen^{†*}, Yihe Deng^{‡*}, Yuanzhi Li[◊], Quanquan Gu[‡]
[‡]Department of Computer Science, University of California, Los Angeles

**Language in a Bottle: Language Model Guided Concept Bottlenecks
for Interpretable Image Classification**

Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin,

Enhancing CLIP with GPT-4: Harnessing Visual Descriptions as Prompts

Mayug Maniparambil, Chris Vorster, Derek Molloy, Noel Murphy,
 Kevin McGuinness, Noel E. O'Connor
 ML Labs, Dublin City University,
 Dublin, Ireland

Generating customized prompts for zero-shot image classification

Sarah Pratt^{1*} Ian Covert¹ Rosanne Liu^{2,3} Ali Farhadi¹

¹University of Washington ²Google DeepMind ³ML Collective

Context

- Pre-training/prompting techniques advanced significantly in applications.
- How do we understand/perform theoretical generalization analysis of ZSP?

UNDERSTANDING TRANSFERABLE REPRESENTATION LEARNING AND ZERO-SHOT TRANSFER IN CLIP

Zixiang Chen^{†*}, Yihe Deng^{‡*}, Yuanzhi Li[△], Quanquan Gu[‡]

[†]Department of Computer Science, University of California, Los Angeles

Language in a Bottle: Language Model Guided Concept Bottlenecks for Interpretable Image Classification

Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin,

Enhancing CLIP with GPT-4: Harnessing Visual Descriptions as Prompts

Mayug Maniparambil, Chris Vorster, Derek Molloy, Noel Murphy,

Kevin McGuinness, Noel E. O'Connor

ML Labs, Dublin City University,
Dublin, Ireland

Generating customized prompts for zero-shot image classification

Sarah Pratt^{1*}

Ian Covert¹

Rosanne Liu^{2,3}

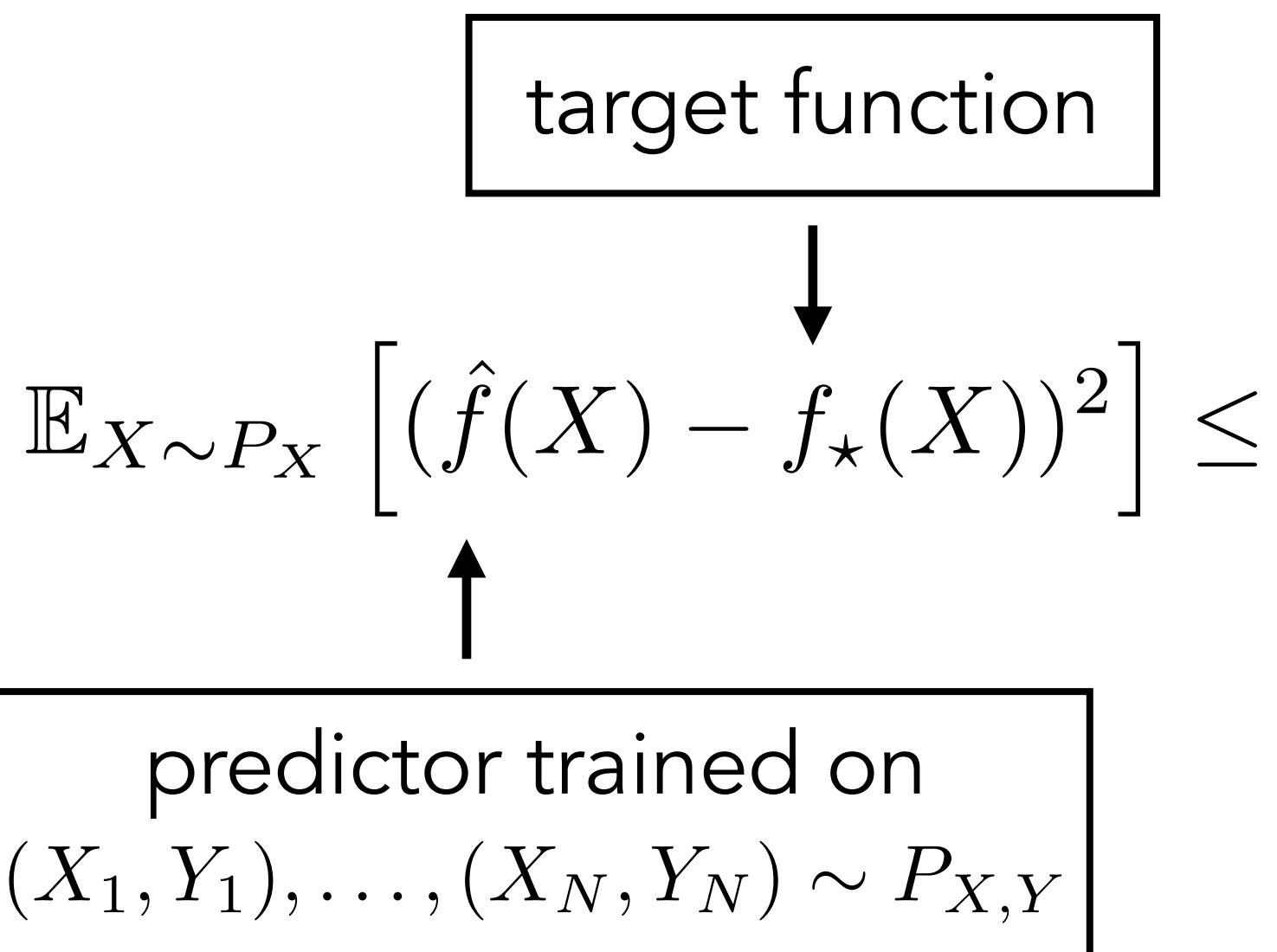
Ali Farhadi¹

¹University of Washington

²Google DeepMind

³ML Collective

Generalization Analysis (Supervised Learning)



Context

- Pre-training/prompting techniques advanced significantly in applications.
- How do we understand/perform theoretical generalization analysis of ZSP?

UNDERSTANDING TRANSFERABLE REPRESENTATION LEARNING AND ZERO-SHOT TRANSFER IN CLIP

Zixiang Chen^{†*}, Yihe Deng^{†*}, Yuanzhi Li[◊], Quanquan Gu[‡]

[†]Department of Computer Science, University of California, Los Angeles

Language in a Bottle: Language Model Guided Concept Bottlenecks for Interpretable Image Classification

Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin,

Enhancing CLIP with GPT-4: Harnessing Visual Descriptions as Prompts

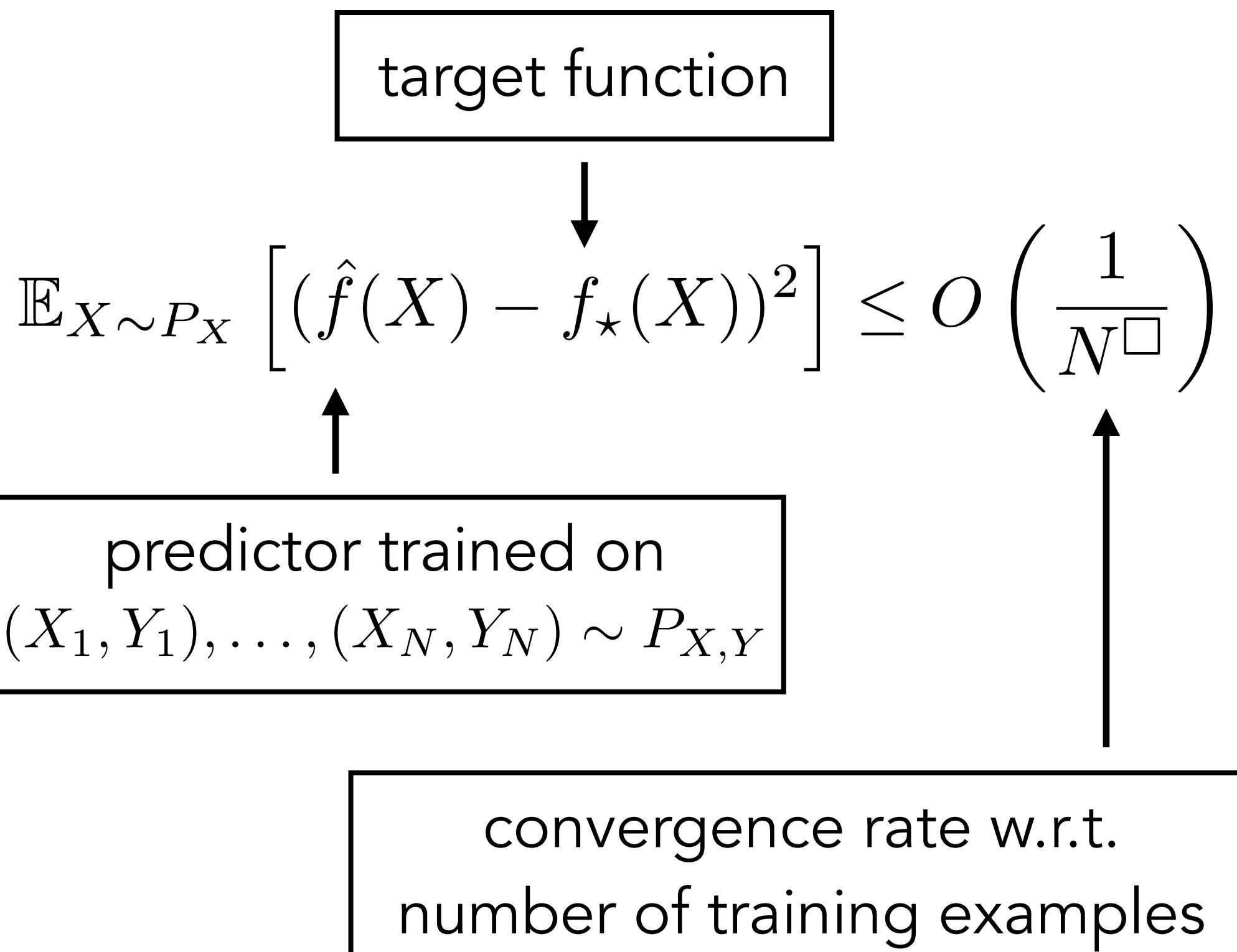
Mayug Maniparambil, Chris Vorster, Derek Molloy, Noel Murphy,
Kevin McGuinness, Noel E. O'Connor
ML Labs, Dublin City University,
Dublin, Ireland

Generating customized prompts for zero-shot image classification

Sarah Pratt^{1*} Ian Covert¹ Rosanne Liu^{2,3} Ali Farhadi¹

¹University of Washington ²Google DeepMind ³ML Collective

Generalization Analysis (Supervised Learning)



Context

- Pre-training/prompting techniques advanced significantly in applications.
- How do we understand/perform theoretical generalization analysis of ZSP?

UNDERSTANDING TRANSFERABLE REPRESENTATION LEARNING AND ZERO-SHOT TRANSFER IN CLIP

Zixiang Chen^{†*}, Yihe Deng^{‡*}, Yuanzhi Li[△], Quanquan Gu[‡]

[†]Department of Computer Science, University of California, Los Angeles

Language in a Bottle: Language Model Guided Concept Bottlenecks for Interpretable Image Classification

Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin,

Enhancing CLIP with GPT-4: Harnessing Visual Descriptions as Prompts

Mayug Maniparambil, Chris Vorster, Derek Molloy, Noel Murphy,

Kevin McGuinness, Noel E. O'Connor

ML Labs, Dublin City University,
Dublin, Ireland

Generating customized prompts for zero-shot image classification

Sarah Pratt^{1*}

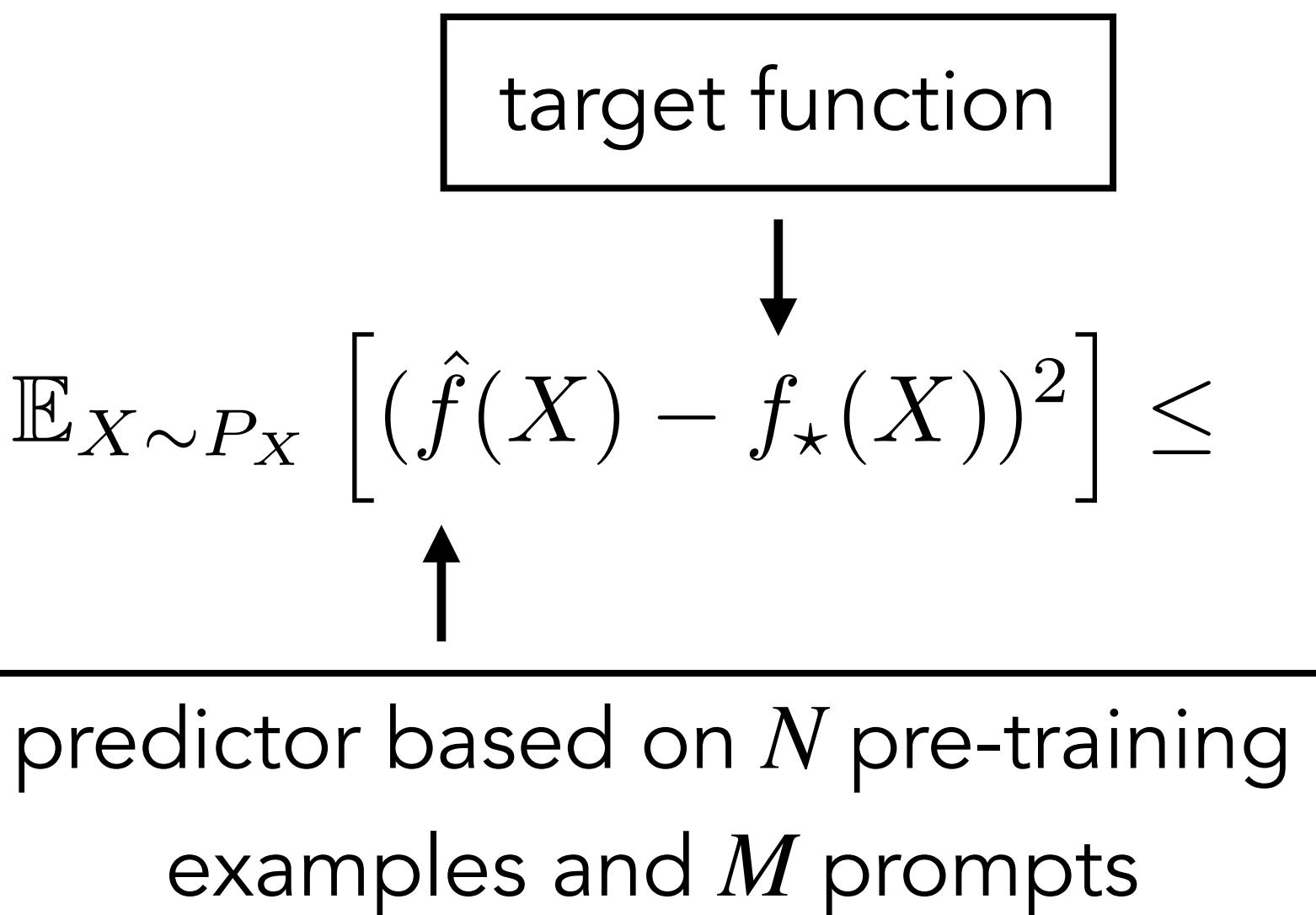
Ian Covert¹

Rosanne Liu^{2,3}

Ali Farhadi¹

¹University of Washington ²Google DeepMind ³ML Collective

Generalization Analysis (Zero-Shot Prediction)



Context

- Pre-training/prompting techniques advanced significantly in applications.
- How do we understand/perform theoretical generalization analysis of ZSP?

UNDERSTANDING TRANSFERABLE REPRESENTATION LEARNING AND ZERO-SHOT TRANSFER IN CLIP

Zixiang Chen^{†*}, Yihe Deng^{†*}, Yuanzhi Li[◊], Quanquan Gu[‡]

[†]Department of Computer Science, University of California, Los Angeles

Language in a Bottle: Language Model Guided Concept Bottlenecks for Interpretable Image Classification

Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin,

Enhancing CLIP with GPT-4: Harnessing Visual Descriptions as Prompts

Mayug Maniparambil, Chris Vorster, Derek Molloy, Noel Murphy,
Kevin McGuinness, Noel E. O'Connor
ML Labs, Dublin City University,
Dublin, Ireland

Generating customized prompts for zero-shot image classification

Sarah Pratt^{1*} Ian Covert¹ Rosanne Liu^{2,3} Ali Farhadi¹

¹University of Washington ²Google DeepMind ³ML Collective

Generalization Analysis (Zero-Shot Prediction)

target function

$$\mathbb{E}_{X \sim P_X} \left[(\hat{f}(X) - f_\star(X))^2 \right] \lesssim \frac{1}{N^{\alpha}} + \frac{1}{M^{\alpha}} + ?$$

predictor based on N pre-training examples and M prompts

convergence rate w.r.t. N, M , and fundamental limits of ZSP

Context

- **Pre-training/prompting techniques** advanced significantly in applications.
- How do we understand/perform theoretical **generalization analysis of ZSP?**

UNDERSTANDING TRANSFERABLE REPRESENTATION LEARNING AND ZERO-SHOT TRANSFER IN CLIP

Zixiang Chen^{†*}, Yihe Deng^{‡*}, Yuanzhi Li[△], Quanquan Gu[‡]

[†]Department of Computer Science, University of California, Los Angeles

Language in a Bottle: Language Model Guided Concept Bottlenecks for Interpretable Image Classification

Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin,

Enhancing CLIP with GPT-4: Harnessing Visual Descriptions as Prompts

Mayug Maniparambil, Chris Vorster, Derek Molloy, Noel Murphy,
Kevin McGuinness, Noel E. O'Connor
ML Labs, Dublin City University,
Dublin, Ireland

Generating customized prompts for zero-shot image classification

Sarah Pratt^{1*} Ian Covert¹ Rosanne Liu^{2,3} Ali Farhadi¹

¹University of Washington ²Google DeepMind ³ML Collective

Contributions

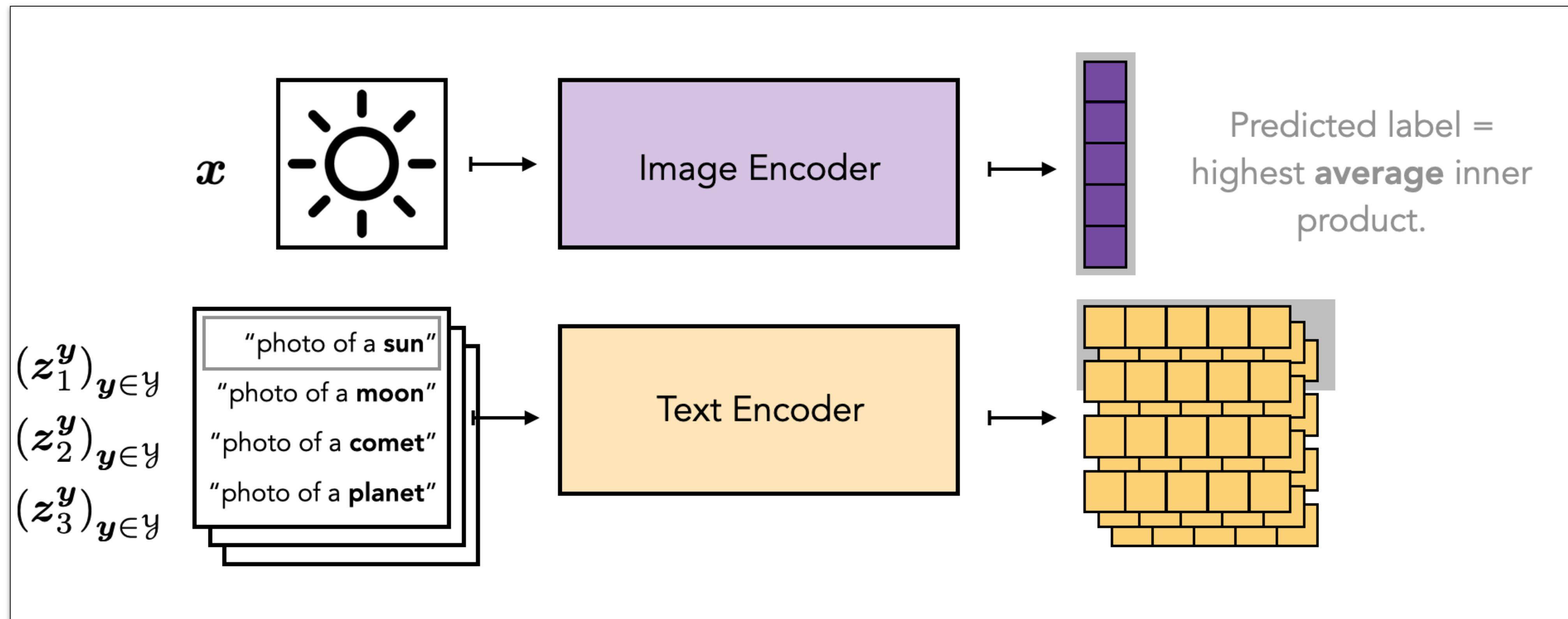
1. Theoretical framework to formalize zero-shot prediction (ZSP) and obtain its generalization analysis.
2. Two proof strategies which apply to different classes of methods.
3. Key quantities for success of ZSP: **residual dependence**, **prompt bias**, **sample complexity**, and **prompt complexity**.

$$\mathbb{E}_{X \sim P_X} \left[(f_\star(X) - \hat{f}(X))^2 \right] \leq$$

direct predictor ZSP procedure

$$\mathbb{E}_{X \sim P_X} \left[(f_\star(X) - \hat{f}(X))^2 \right] \leftarrow$$

direct predictor ZSP procedure



$$\mathbb{E}_{X \sim P_X} [(f_\star(X) - \hat{f}(X))^2] \leq 2\mathbb{E}_{X \sim P_X} [(f_\star(X) - \bar{f}(X))^2] + 2\mathbb{E}_{X \sim P_X} [(\bar{f}(X) - \hat{f}(X))^2]$$

direct predictor

ZSP procedure

population version of ZSP
(based on distributions instead of samples)

$$\mathbb{E}_{X \sim P_X} [(f_\star(X) - \hat{f}(X))^2] \leq 2\mathbb{E}_{X \sim P_X} [(f_\star(X) - \bar{f}(X))^2] + 2\mathbb{E}_{X \sim P_X} [(\bar{f}(X) - \hat{f}(X))^2]$$

information-theoretic error

learning error

direct predictor ZSP procedure

population version of ZSP
(based on distributions instead of samples)

$$\mathbb{E}_{X \sim P_X} [(f_\star(X) - \hat{f}(X))^2] \leq 2\mathbb{E}_{X \sim P_X} [(f_\star(X) - \bar{f}(X))^2] + 2\mathbb{E}_{X \sim P_X} [(\bar{f}(X) - \hat{f}(X))^2]$$

The diagram illustrates the decomposition of the total error into two components. A horizontal yellow bar spans the width of the equation. Two vertical dotted lines drop from the terms $f_\star(X) - \hat{f}(X)$ and $\bar{f}(X) - \hat{f}(X)$ down to the yellow bar. A pink horizontal bracket above the yellow bar is labeled "information-theoretic error". A green horizontal bracket to the right of the yellow bar is labeled "learning error". Below the yellow bar, the text "population version of ZSP (based on distributions instead of samples)" is written.

information-theoretic error

learning error

population version of ZSP
(based on distributions instead of samples)

direct predictor ZSP procedure

Roadmap of Theoretical Analysis

1. Define \bar{f} in terms of pre-training, evaluation, and prompting distribution.
2. Upper bound **information-theoretic error** using dependence relationships between images, captions, and labels.
3. Define class of estimators \hat{f} , and bound **learning error** using tools from statistical learning theory.

$$\mathbb{E}_{X \sim P_X} [(f_\star(X) - \hat{f}(X))^2] \leq 2\mathbb{E}_{X \sim P_X} [(f_\star(X) - \bar{f}(X))^2] + 2\mathbb{E}_{X \sim P_X} [(\bar{f}(X) - \hat{f}(X))^2]$$

The diagram illustrates the decomposition of the total error into two components. A horizontal yellow bar spans the width of the equation. Two vertical dotted lines drop from the terms $f_\star(X)$ and $\hat{f}(X)$ to the labels "direct predictor" and "ZSP procedure" respectively. A pink horizontal bracket above the yellow bar spans from the "direct predictor" label to the term $(f_\star(X) - \bar{f}(X))^2$, labeled "information-theoretic error". A green horizontal bracket above the yellow bar spans from the term $(\bar{f}(X) - \hat{f}(X))^2$ to the "learning error" label. Below the yellow bar, the text "population version of ZSP (based on distributions instead of samples)" is centered.

information-theoretic error

learning error

population version of ZSP
(based on distributions instead of samples)

Roadmap of Theoretical Analysis

1. Define \bar{f} in terms of pre-training, evaluation, and prompting distribution.
2. Upper bound **information-theoretic error** using dependence relationships between images, captions, and labels.
3. Define class of estimators \hat{f} , and bound **learning error** using tools from statistical learning theory.

X = image

Y = label

Z = caption

$$\mathbb{E}_{X \sim P_X} [(f_\star(X) - \hat{f}(X))^2] \leq 2\mathbb{E}_{X \sim P_X} [(f_\star(X) - \bar{f}(X))^2] + 2\mathbb{E}_{X \sim P_X} [(\bar{f}(X) - \hat{f}(X))^2]$$

The diagram illustrates the decomposition of the total error into two components. A horizontal dotted line is divided into two segments: a pink segment labeled "information-theoretic error" and a green segment labeled "learning error". Dotted lines from the endpoints of these segments point to the terms in the inequality above. Below the pink segment, there are two vertical dotted lines with labels: "direct predictor" on the left and "ZSP procedure" on the right. A bracket below the green segment spans the entire length of the green line and is labeled "population version of ZSP (based on distributions instead of samples)".

Roadmap of Theoretical Analysis

1. Define \bar{f} in terms of pre-training, evaluation, and prompting distribution.
2. Upper bound **information-theoretic error** using dependence relationships between images, captions, and labels.
3. Define class of estimators \hat{f} , and bound **learning error** using tools from statistical learning theory.

X = image

Y = label

Z = caption

$$\mathbb{E}_{X \sim P_X} [(f_\star(X) - \hat{f}(X))^2] \leq 2\mathbb{E}_{X \sim P_X} [(f_\star(X) - \bar{f}(X))^2] + 2\mathbb{E}_{X \sim P_X} [(\bar{f}(X) - \hat{f}(X))^2]$$

information-theoretic error

learning error

direct predictor ZSP procedure

population version of ZSP
(based on distributions instead of samples)

Roadmap of Theoretical Analysis

1. Define \bar{f} in terms of pre-training, evaluation, and prompting distribution.
2. Upper bound **information-theoretic error** using dependence relationships between images, captions, and labels.
3. Define class of estimators \hat{f} , and bound **learning error** using tools from statistical learning theory.

X = image

Y = label

Z = caption

$P_{X,Y}$
Evaluation

$$\mathbb{E}_{X \sim P_X} [(f_\star(X) - \hat{f}(X))^2] \leq 2\mathbb{E}_{X \sim P_X} [(f_\star(X) - \bar{f}(X))^2] + 2\mathbb{E}_{X \sim P_X} [(\bar{f}(X) - \hat{f}(X))^2]$$

direct predictor
ZSP procedure

information-theoretic error
learning error

population version of ZSP
(based on distributions instead of samples)

Roadmap of Theoretical Analysis

1. Define \bar{f} in terms of pre-training, evaluation, and prompting distribution.
2. Upper bound **information-theoretic error** using dependence relationships between images, captions, and labels.
3. Define class of estimators \hat{f} , and bound **learning error** using tools from statistical learning theory.

X = image

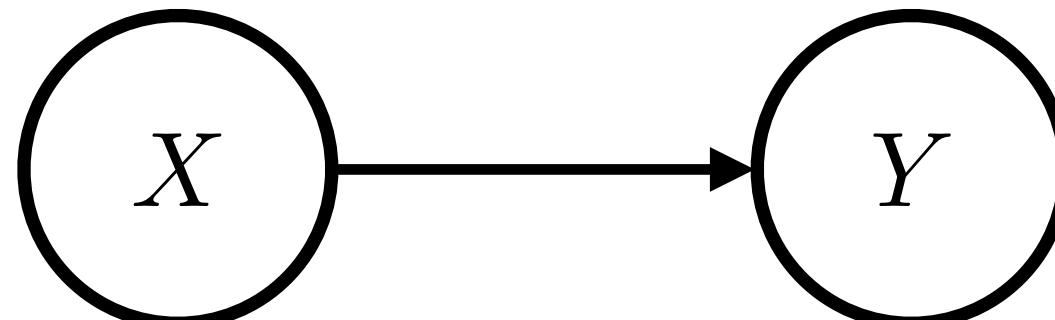
Y = label

Z = caption

$$P_{X,Y}$$

Evaluation

$$f_\star(\mathbf{x}) = \mathbb{E}_{P_{X,Y}} [Y | X = \mathbf{x}]$$



$$\mathbb{E}_{X \sim P_X} [(f_\star(X) - \hat{f}(X))^2] \leq 2\mathbb{E}_{X \sim P_X} [(f_\star(X) - \bar{f}(X))^2] + 2\mathbb{E}_{X \sim P_X} [(\bar{f}(X) - \hat{f}(X))^2]$$

direct predictor ZSP procedure

information-theoretic error

learning error

population version of ZSP
(based on distributions instead of samples)

Roadmap of Theoretical Analysis

1. Define \bar{f} in terms of pre-training, evaluation, and prompting distribution.
 2. Upper bound information-theoretic error using dependence relationships between images, captions, and labels.
 3. Define class of estimators \hat{f} , and bound learning error using tools from statistical learning theory.

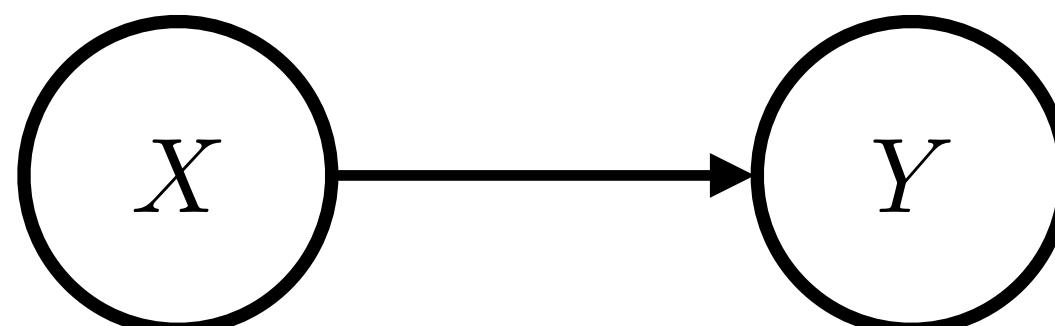
$X = \text{image}$

$Y = \text{label}$

Z = caption



$$f_\star(x) = \mathbb{E}_{P_{X,Y}}[Y|X=x]$$



$$\mathbb{E}_{X \sim P_X} [(f_\star(X) - \hat{f}(X))^2] \leq 2\mathbb{E}_{X \sim P_X} [(f_\star(X) - \bar{f}(X))^2] + 2\mathbb{E}_{X \sim P_X} [(\bar{f}(X) - \hat{f}(X))^2]$$

direct predictor
 ZSP procedure
 information-theoretic error
 learning error
 population version of ZSP
 (based on distributions instead of samples)

Roadmap of Theoretical Analysis

1. Define \bar{f} in terms of pre-training, evaluation, and prompting distribution.
2. Upper bound **information-theoretic error** using dependence relationships between images, captions, and labels.
3. Define class of estimators \hat{f} , and bound **learning error** using tools from statistical learning theory.

X = image

Y = label

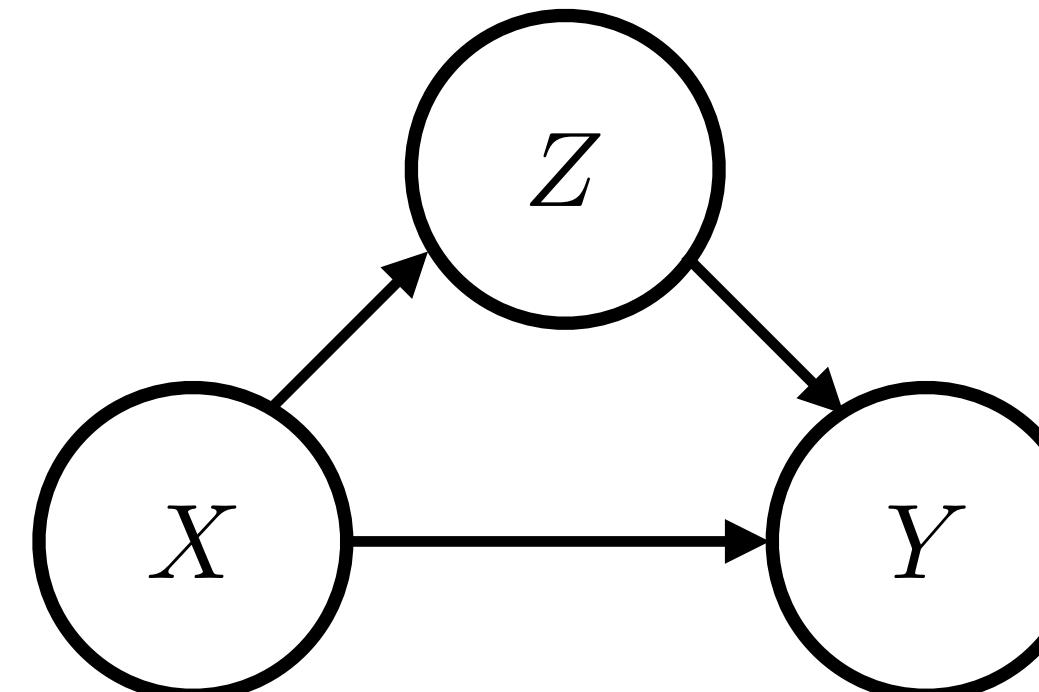
Z = caption

$P_{X,Y}$
Evaluation

$Q_{X,Z}$
Pre-Training

$\rho_{Y,Z}$
Prompting

$$f_\star(\mathbf{x}) = \mathbb{E}_{P_{X,Y}} [Y|X = \mathbf{x}] \quad \bar{f}(\mathbf{x}) = \mathbb{E}_{Q_{X,Z}} [\mathbb{E}_{\rho_{Y,Z}} [Y|Z] | X = \mathbf{x}]$$



$$\mathbb{E}_{X \sim P_X} [(f_\star(X) - \hat{f}(X))^2] \leq 2\mathbb{E}_{X \sim P_X} [(f_\star(X) - \bar{f}(X))^2] + 2\mathbb{E}_{X \sim P_X} [(\bar{f}(X) - \hat{f}(X))^2]$$

direct predictor
ZSP procedure

information-theoretic error

learning error

population version of ZSP
 (based on distributions instead of samples)

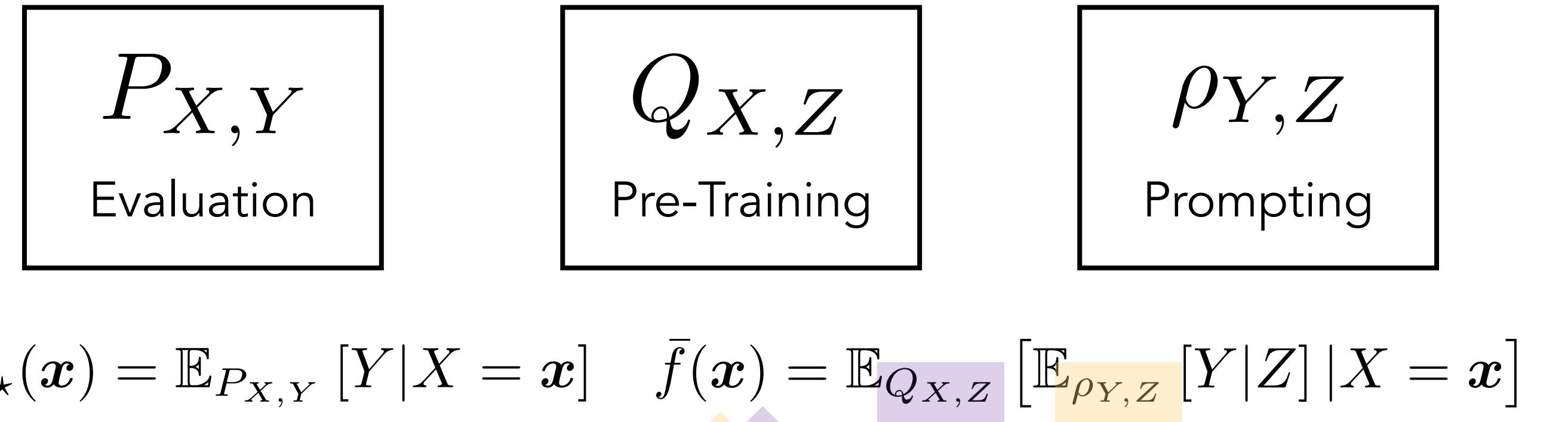
Roadmap of Theoretical Analysis

1. Define \bar{f} in terms of pre-training, evaluation, and prompting distribution.
2. Upper bound **information-theoretic error** using dependence relationships between images, captions, and labels.
3. Define class of estimators \hat{f} , and bound **learning error** using tools from statistical learning theory.

X = image

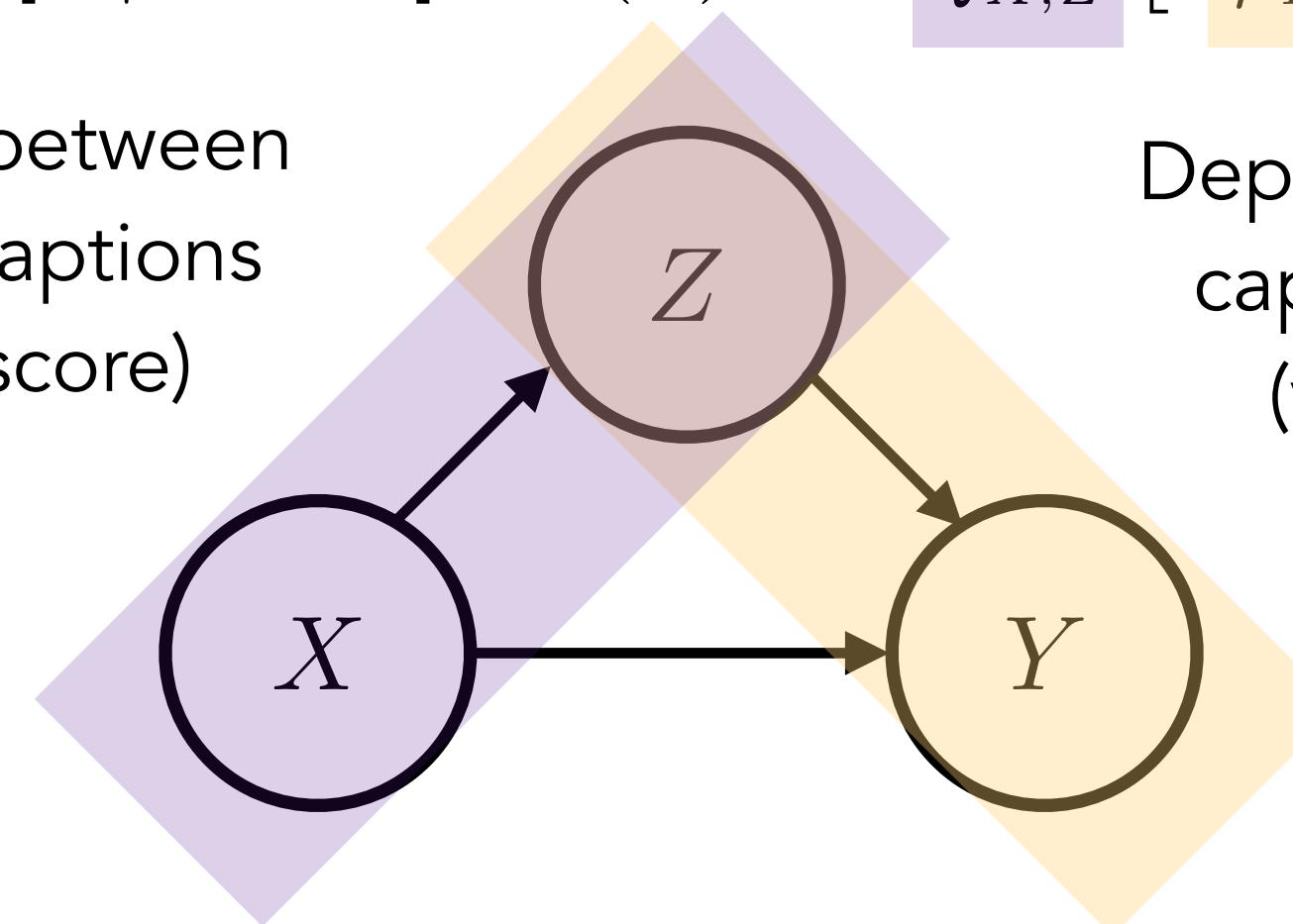
Y = label

Z = caption



$$f_\star(\mathbf{x}) = \mathbb{E}_{P_{X,Y}} [Y|X = \mathbf{x}] \quad \bar{f}(\mathbf{x}) = \mathbb{E}_{Q_{X,Z}} [\mathbb{E}_{\rho_{Y,Z}} [Y|Z] |X = \mathbf{x}]$$

Dependence between
images and captions
(e.g., CLIP score)



Dependence between
captions and labels
(via prompting)

$$\mathbb{E}_{X \sim P_X} [(f_\star(X) - \hat{f}(X))^2] \leq 2\mathbb{E}_{X \sim P_X} [(f_\star(X) - \bar{f}(X))^2] + 2\mathbb{E}_{X \sim P_X} [(\bar{f}(X) - \hat{f}(X))^2]$$

direct predictor ZSP procedure

information-theoretic error

learning error

population version of ZSP
(based on distributions instead of samples)

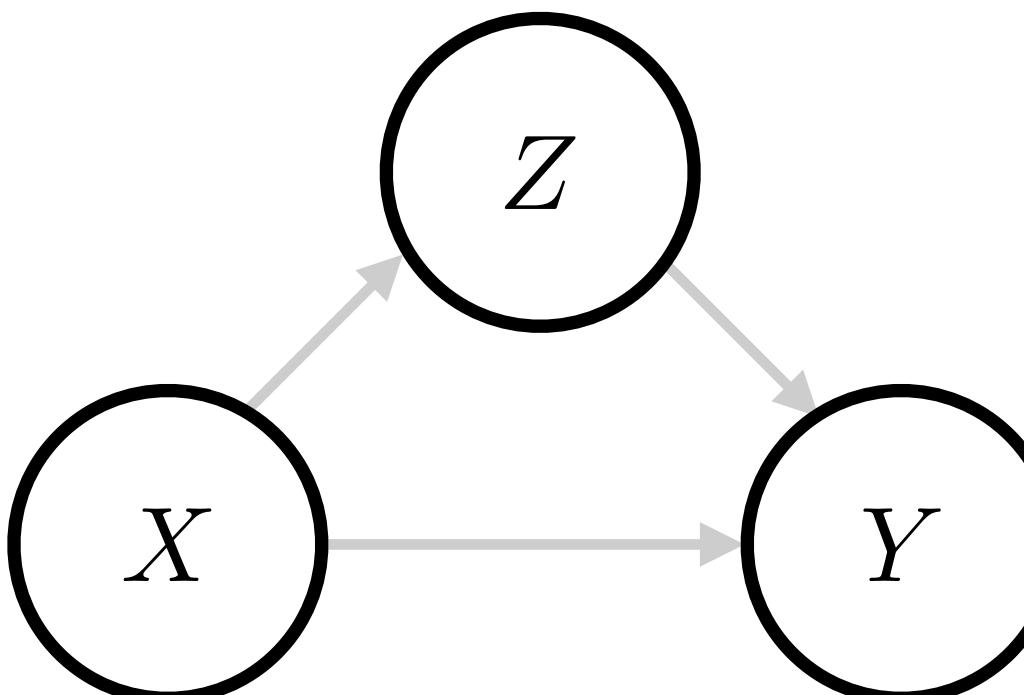
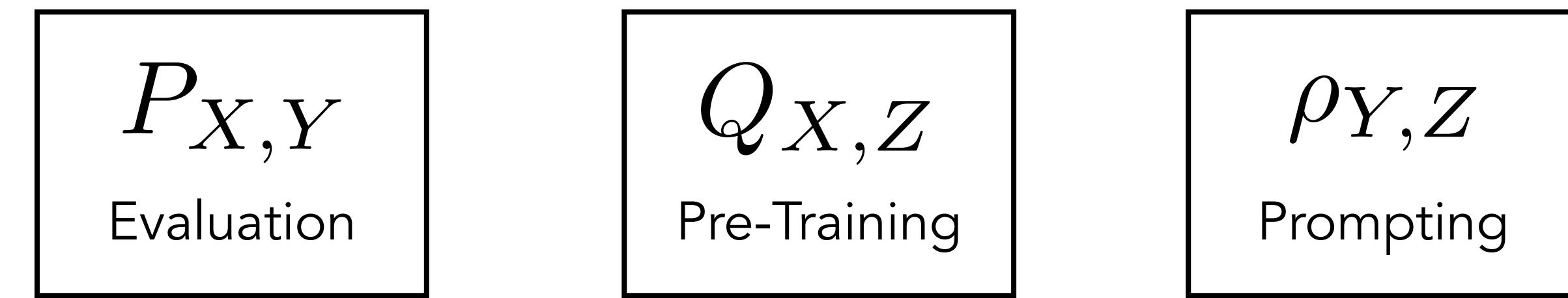
Roadmap of Theoretical Analysis

1. Define \bar{f} in terms of pre-training, evaluation, and prompting distribution.
 2. Upper bound **information-theoretic error** using dependence relationships between images, captions, and labels.
 3. Define class of estimators \hat{f} , and bound **learning error** using tools from statistical learning theory.

$X = \text{image}$

$Y = \text{label}$

Z = caption



$$\mathbb{E}_{X \sim P_X} [(f_\star(X) - \hat{f}(X))^2] \leq 2\mathbb{E}_{X \sim P_X} [(f_\star(X) - \bar{f}(X))^2] + 2\mathbb{E}_{X \sim P_X} [(\bar{f}(X) - \hat{f}(X))^2]$$

direct predictor ZSP procedure

information-theoretic error

learning error

population version of ZSP
(based on distributions instead of samples)

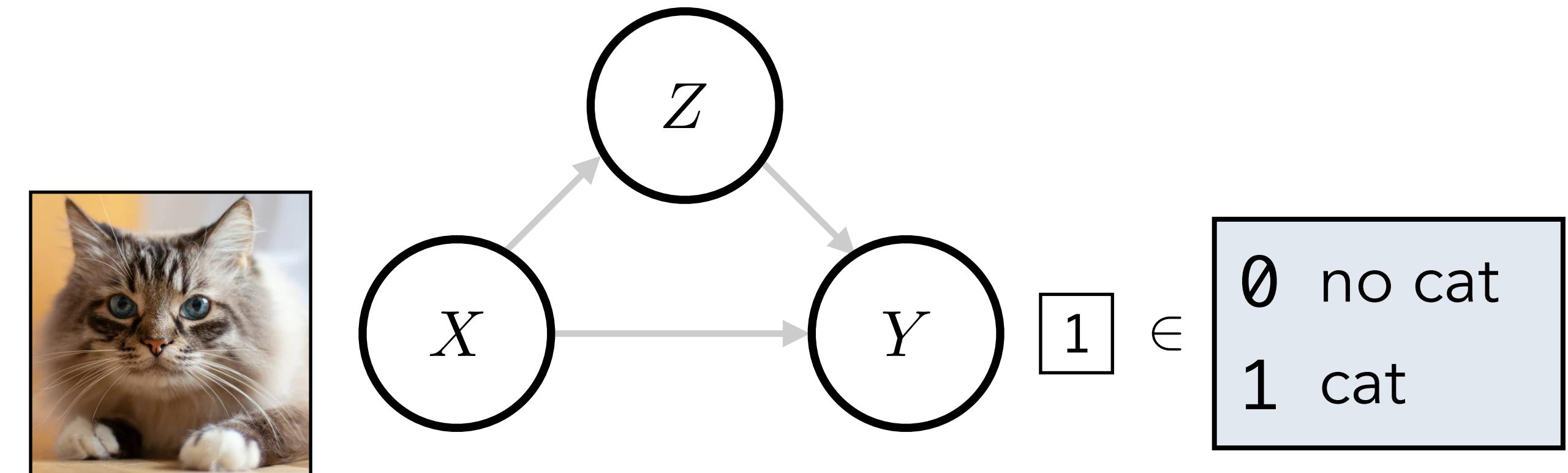
Roadmap of Theoretical Analysis

1. Define \bar{f} in terms of pre-training, evaluation, and prompting distribution.
 2. Upper bound information-theoretic error using dependence relationships between images, captions, and labels.
 3. Define class of estimators \hat{f} , and bound learning error using tools from statistical learning theory.

$X = \text{image}$

$Y = \text{label}$

Z = caption



$$\mathbb{E}_{X \sim P_X} [(f_\star(X) - \hat{f}(X))^2] \leq 2\mathbb{E}_{X \sim P_X} [(f_\star(X) - \bar{f}(X))^2] + 2\mathbb{E}_{X \sim P_X} [(\bar{f}(X) - \hat{f}(X))^2]$$

direct predictor ZSP procedure

information-theoretic error

learning error

population version of ZSP
(based on distributions instead of samples)

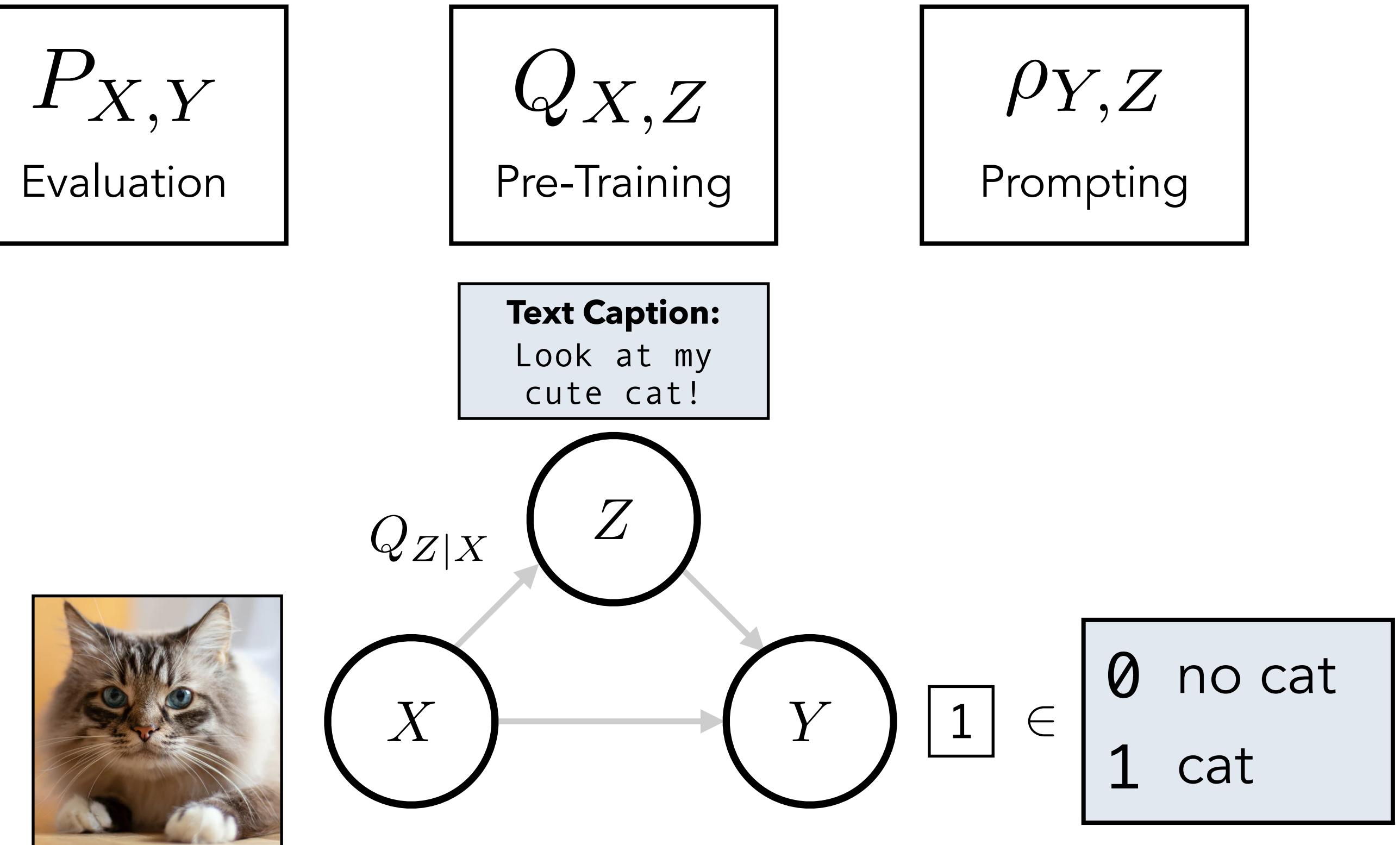
Roadmap of Theoretical Analysis

1. Define \bar{f} in terms of pre-training, evaluation, and prompting distribution.
 2. Upper bound information-theoretic error using dependence relationships between images, captions, and labels.
 3. Define class of estimators \hat{f} , and bound learning error using tools from statistical learning theory.

$X = \text{image}$

$Y = \text{label}$

Z = caption



$$\mathbb{E}_{X \sim P_X} [(f_\star(X) - \hat{f}(X))^2] \leq 2\mathbb{E}_{X \sim P_X} [(f_\star(X) - \bar{f}(X))^2] + 2\mathbb{E}_{X \sim P_X} [(\bar{f}(X) - \hat{f}(X))^2]$$

direct predictor ZSP procedure

information-theoretic error

learning error

population version of ZSP
(based on distributions instead of samples)

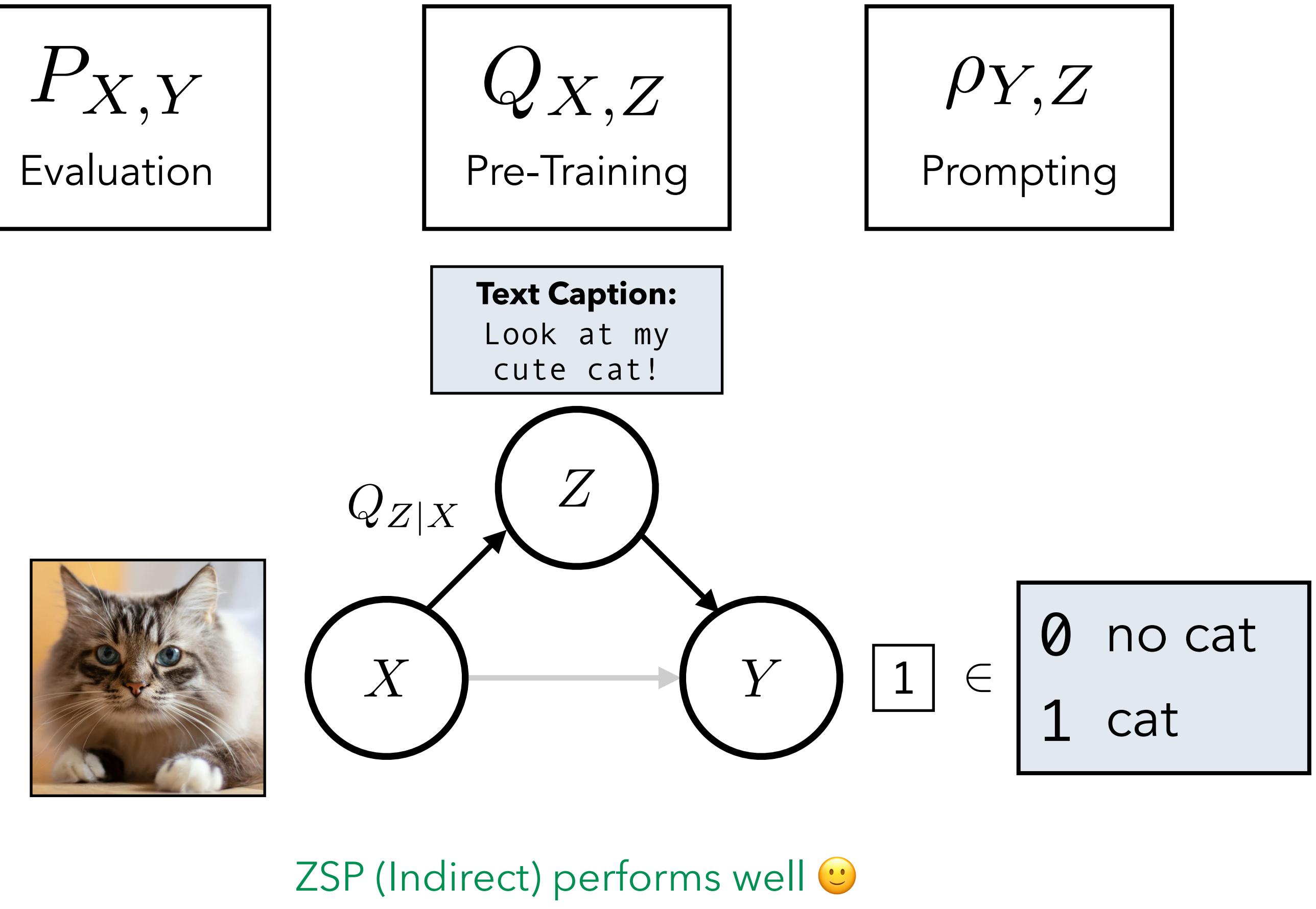
Roadmap of Theoretical Analysis

1. Define \bar{f} in terms of pre-training, evaluation, and prompting distribution.
 2. Upper bound information-theoretic error using dependence relationships between images, captions, and labels.
 3. Define class of estimators \hat{f} , and bound learning error using tools from statistical learning theory.

$X = \text{image}$

$Y = \text{label}$

Z = caption



$$\mathbb{E}_{X \sim P_X} [(f_\star(X) - \hat{f}(X))^2] \leq 2\mathbb{E}_{X \sim P_X} [(f_\star(X) - \bar{f}(X))^2] + 2\mathbb{E}_{X \sim P_X} [(\bar{f}(X) - \hat{f}(X))^2]$$

direct predictor ZSP procedure

information-theoretic error

learning error

population version of ZSP
(based on distributions instead of samples)

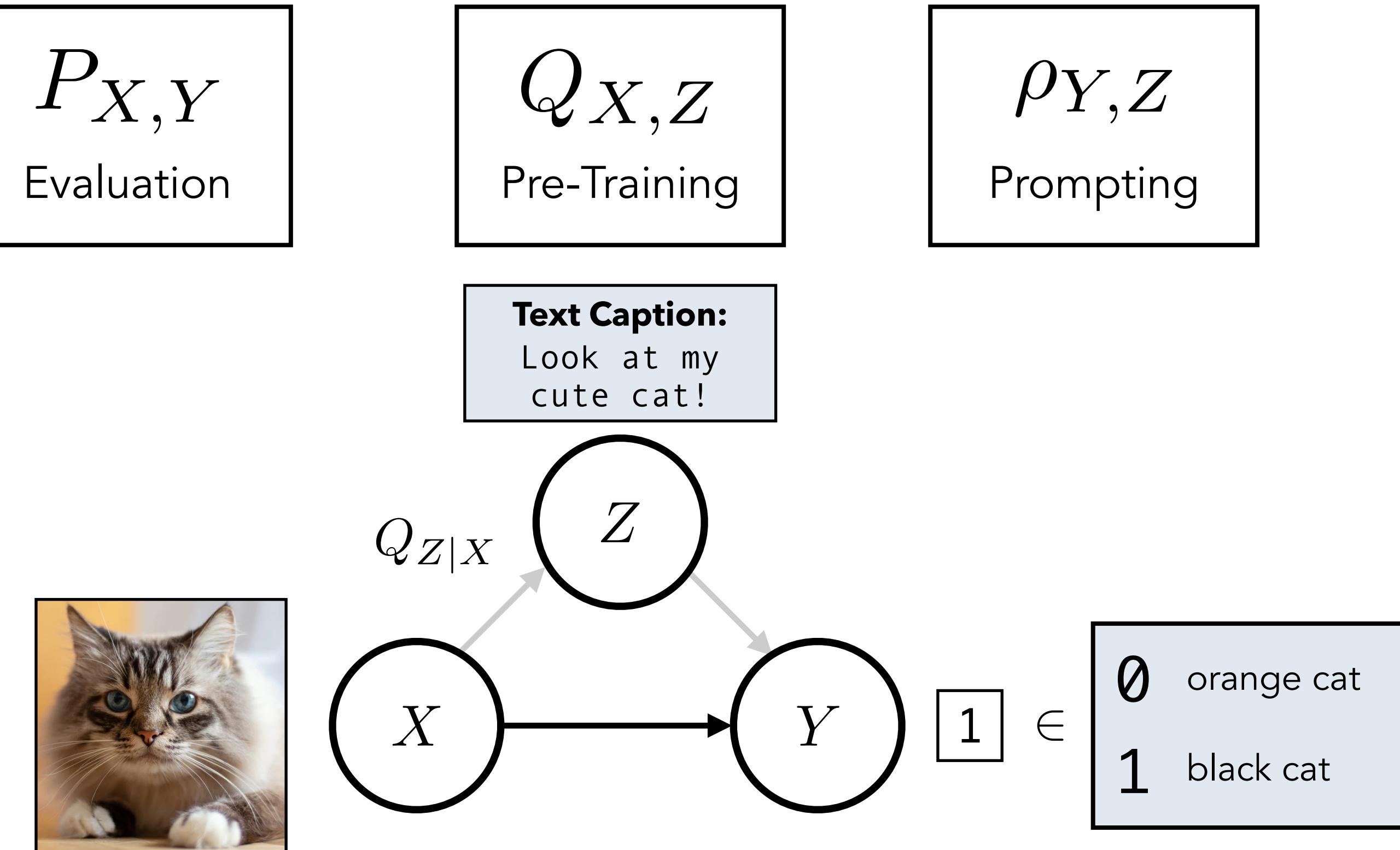
Roadmap of Theoretical Analysis

1. Define \bar{f} in terms of pre-training, evaluation, and prompting distribution.
 2. Upper bound information-theoretic error using dependence relationships between images, captions, and labels.
 3. Define class of estimators \hat{f} , and bound learning error using tools from statistical learning theory.

$X = \text{image}$

$Y = \text{label}$

Z = caption



ZSP (Indirect) performs poorly 😞

$$\mathbb{E}_{X \sim P_X} [(f_\star(X) - \hat{f}(X))^2] \leq 2\mathbb{E}_{X \sim P_X} [(f_\star(X) - \bar{f}(X))^2] + 2\mathbb{E}_{X \sim P_X} [(\bar{f}(X) - \hat{f}(X))^2]$$

direct predictor ZSP procedure

information-theoretic error

learning error

population version of ZSP
(based on distributions instead of samples)

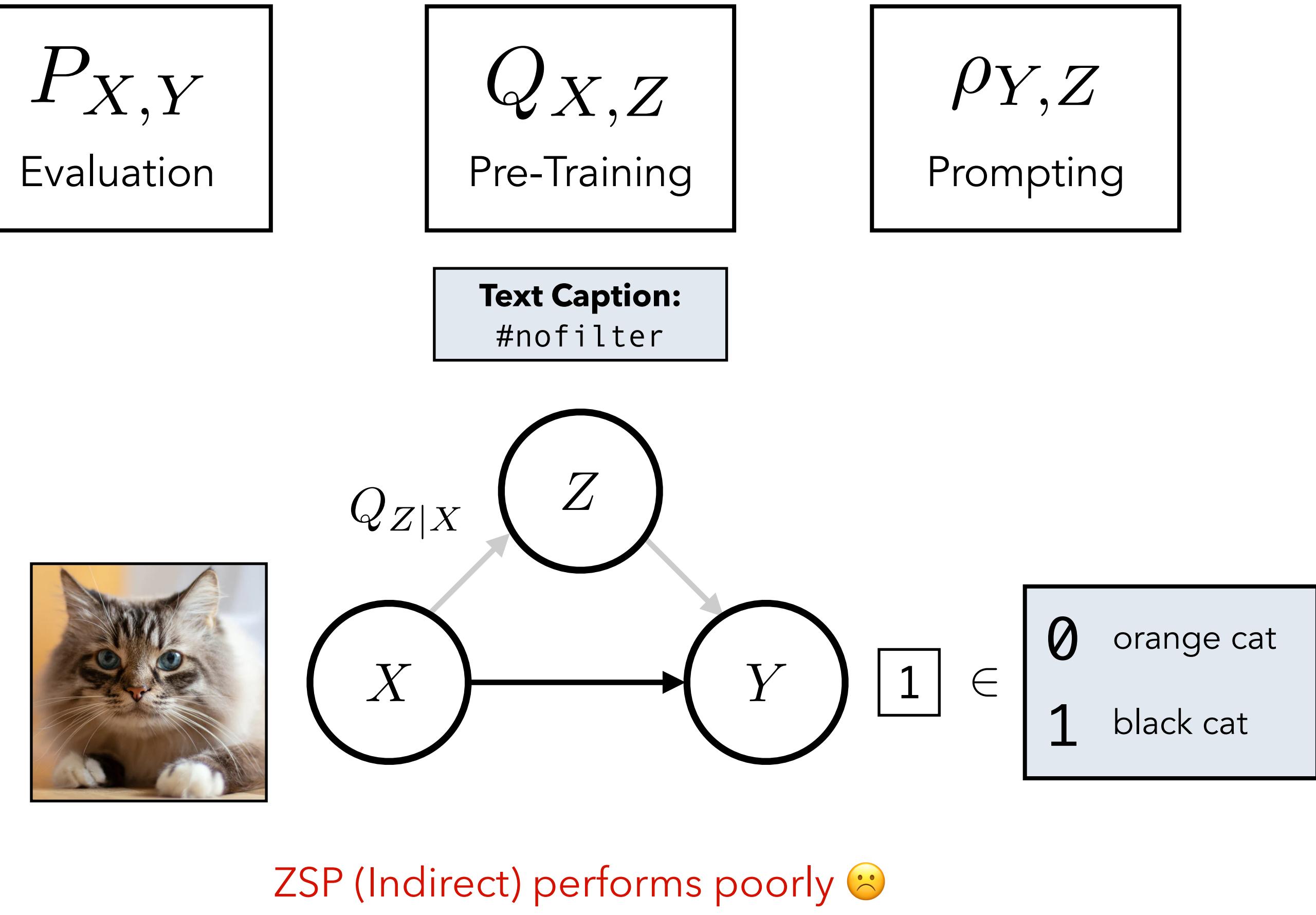
Roadmap of Theoretical Analysis

1. Define \bar{f} in terms of pre-training, evaluation, and prompting distribution.
 2. Upper bound information-theoretic error using dependence relationships between images, captions, and labels.
 3. Define class of estimators \hat{f} , and bound learning error using tools from statistical learning theory.

$X = \text{image}$

$Y = \text{label}$

Z = caption



$$\mathbb{E}_{X \sim P_X} [(f_\star(X) - \hat{f}(X))^2] \leq 2\mathbb{E}_{X \sim P_X} [(f_\star(X) - \bar{f}(X))^2] + 2\mathbb{E}_{X \sim P_X} [(\bar{f}(X) - \hat{f}(X))^2]$$

direct predictor
ZSP procedure

information-theoretic error

learning error

population version of ZSP
(based on distributions instead of samples)

Roadmap of Theoretical Analysis

1. Define \bar{f} in terms of pre-training, evaluation, and prompting distribution.
2. Upper bound **information-theoretic error** using dependence relationships between images, captions, and labels.
3. Define class of estimators \hat{f} , and bound **learning error** using tools from statistical learning theory.

X = image

Y = label

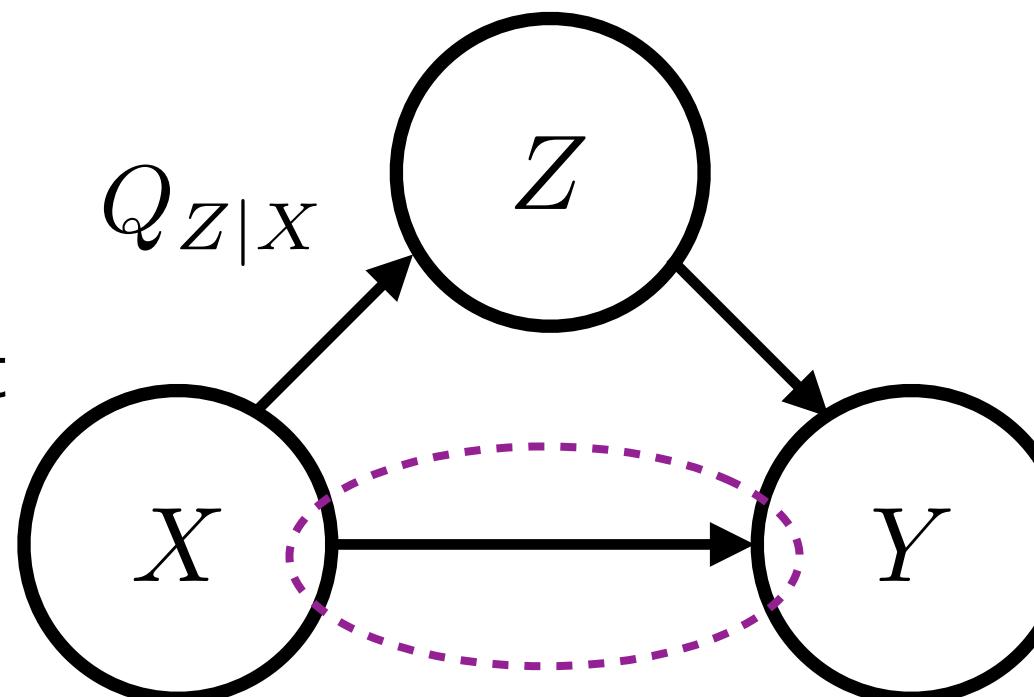
Z = caption

Theorem. (Mehta & Harchaoui, ICML '25)

$$\mathbb{E}_{X \sim P_X} [(f_\star(X) - \bar{f}(X))^2] \lesssim$$

$P_{X,Y,Z}$ denotes any joint distribution such that

$$P_{Z|X} = Q_{Z|X}.$$



$$\mathbb{E}_{X \sim P_X} [(f_\star(X) - \hat{f}(X))^2] \leq 2\mathbb{E}_{X \sim P_X} [(f_\star(X) - \bar{f}(X))^2] + 2\mathbb{E}_{X \sim P_X} [(\bar{f}(X) - \hat{f}(X))^2]$$

direct predictor
 ZSP procedure
 information-theoretic error
 learning error
 population version of ZSP
 (based on distributions instead of samples)

Roadmap of Theoretical Analysis

1. Define \bar{f} in terms of pre-training, evaluation, and prompting distribution.
2. Upper bound **information-theoretic error** using dependence relationships between images, captions, and labels.
3. Define class of estimators \hat{f} , and bound **learning error** using tools from statistical learning theory.

X = image

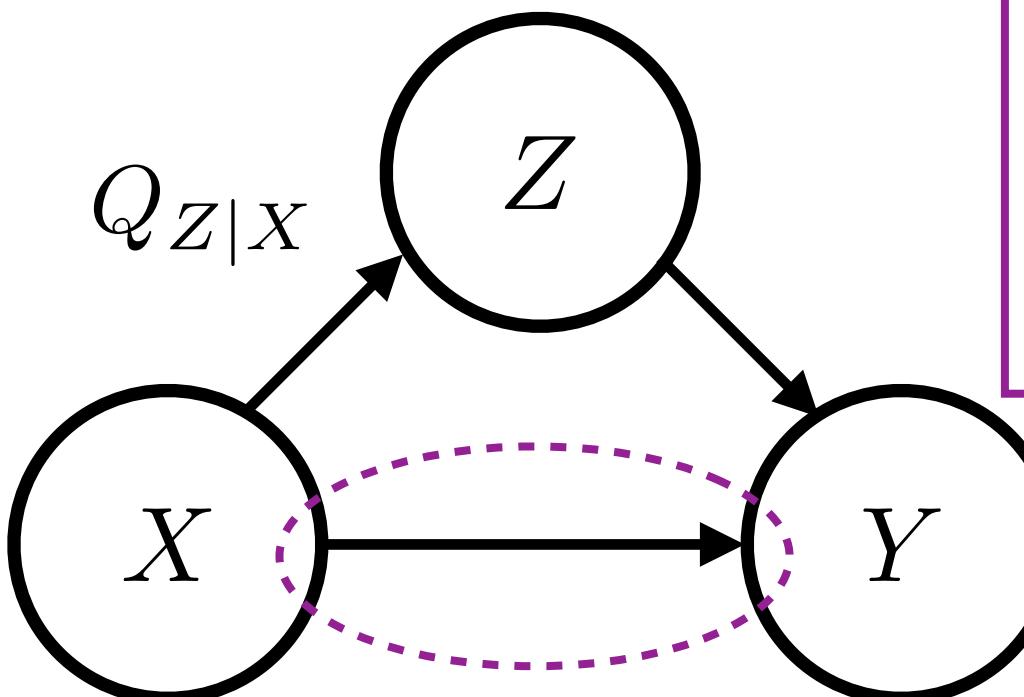
Y = label

Z = caption

Theorem. (Mehta & Harchaoui, ICML '25)

$$\mathbb{E}_{X \sim P_X} [(f_\star(X) - \bar{f}(X))^2] \lesssim I(X, Y|Z) + \text{err}(P_{Y|Z}, \rho_{Y|Z})$$

$P_{X,Y,Z}$ denotes any joint distribution such that
 $P_{Z|X} = Q_{Z|X}$.



Conditional dependence of X and Y given Z , or cost of taking the indirect path through Z .

$$\mathbb{E}_{X \sim P_X} [(f_\star(X) - \hat{f}(X))^2] \leq 2\mathbb{E}_{X \sim P_X} [(f_\star(X) - \bar{f}(X))^2] + 2\mathbb{E}_{X \sim P_X} [(\bar{f}(X) - \hat{f}(X))^2]$$

direct predictor
 ZSP procedure

information-theoretic error

learning error

population version of ZSP
 (based on distributions instead of samples)

Roadmap of Theoretical Analysis

1. Define \bar{f} in terms of pre-training, evaluation, and prompting distribution.
2. Upper bound **information-theoretic error** using dependence relationships between images, captions, and labels.
3. Define class of estimators \hat{f} , and bound **learning error** using tools from statistical learning theory.

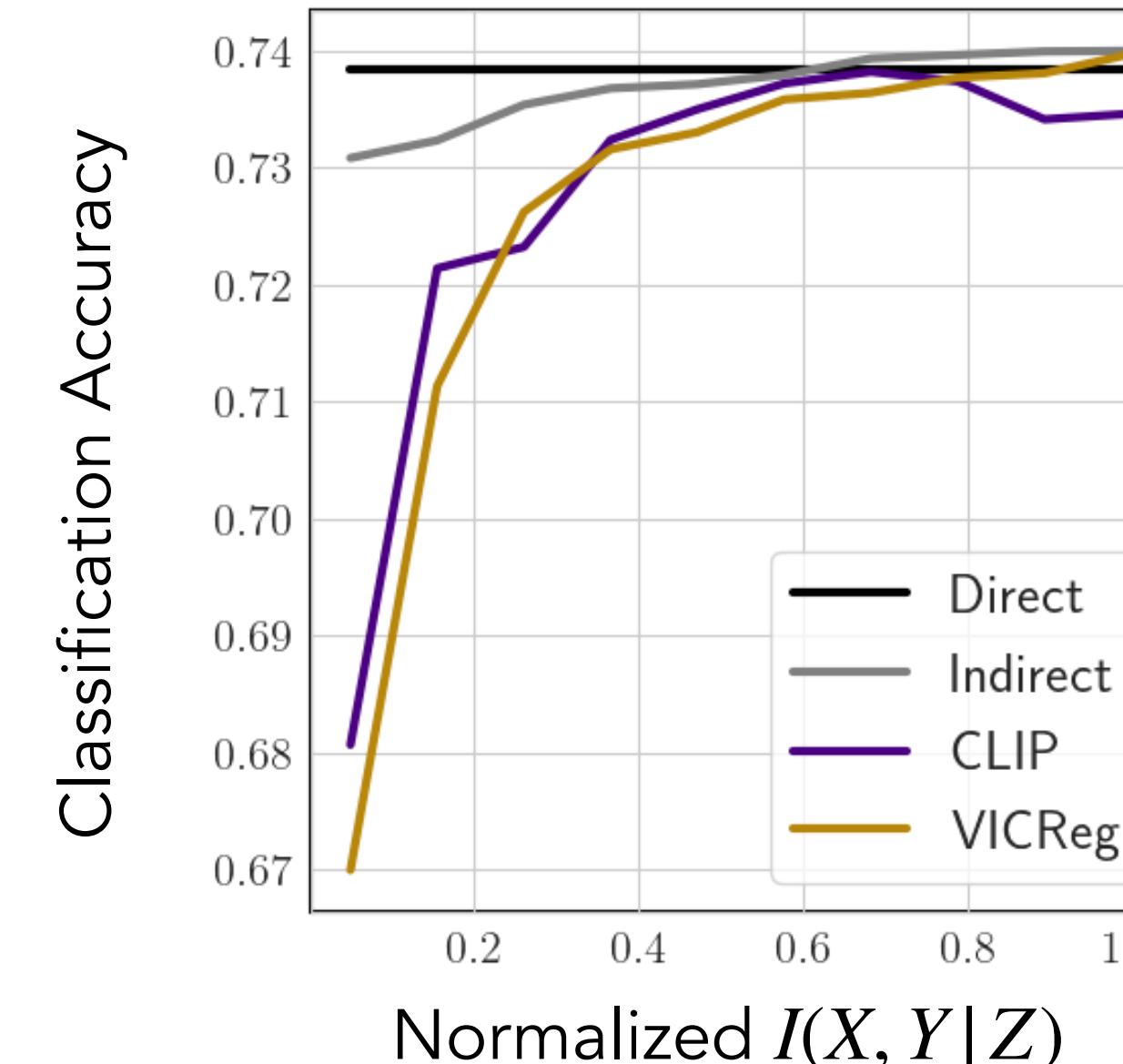
X = image

Y = label

Z = caption

Theorem. (Mehta & Harchaoui, ICML '25)

$$\mathbb{E}_{X \sim P_X} [(f_\star(X) - \bar{f}(X))^2] \lesssim I(X, Y|Z) + \text{err}(P_{Y,Z}, \rho_{Y,Z})$$



Conditional dependence of X and Y given Z , or cost of taking the indirect path through Z .

$$\mathbb{E}_{X \sim P_X} [(f_\star(X) - \hat{f}(X))^2] \leq 2\mathbb{E}_{X \sim P_X} [(f_\star(X) - \bar{f}(X))^2] + 2\mathbb{E}_{X \sim P_X} [(\bar{f}(X) - \hat{f}(X))^2]$$

direct predictor
ZSP procedure

information-theoretic error
learning error

population version of ZSP
(based on distributions instead of samples)

Roadmap of Theoretical Analysis

1. Define \bar{f} in terms of pre-training, evaluation, and prompting distribution.
2. Upper bound **information-theoretic error** using dependence relationships between images, captions, and labels.
3. Define class of estimators \hat{f} , and bound **learning error** using tools from statistical learning theory.

X = image

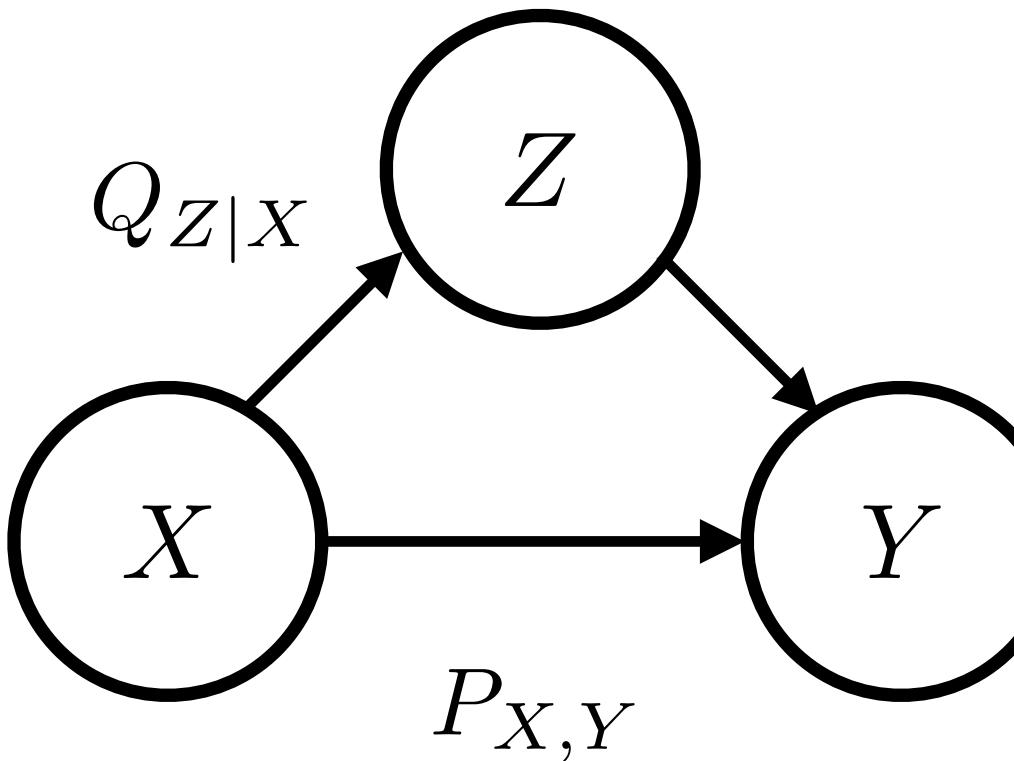
Y = label

Z = caption

Theorem. (Mehta & Harchaoui, ICML '25)

$$\mathbb{E}_{X \sim P_X} [(f_\star(X) - \bar{f}(X))^2] \lesssim I(X, Y|Z) + \text{err}(P_{Y|Z}, \rho_{Y|Z})$$

$P_{X,Y,Z}$ denotes any joint distribution such that
 $P_{Z|X} = Q_{Z|X}$.



$$\mathbb{E}_{X \sim P_X} [(f_\star(X) - \hat{f}(X))^2] \leq 2\mathbb{E}_{X \sim P_X} [(f_\star(X) - \bar{f}(X))^2] + 2\mathbb{E}_{X \sim P_X} [(\bar{f}(X) - \hat{f}(X))^2]$$

direct predictor
ZSP procedure

information-theoretic error
learning error

population version of ZSP
(based on distributions instead of samples)

Roadmap of Theoretical Analysis

1. Define \bar{f} in terms of pre-training, evaluation, and prompting distribution.
2. Upper bound **information-theoretic error** using dependence relationships between images, captions, and labels.
3. Define class of estimators \hat{f} , and bound **learning error** using tools from statistical learning theory.

X = image

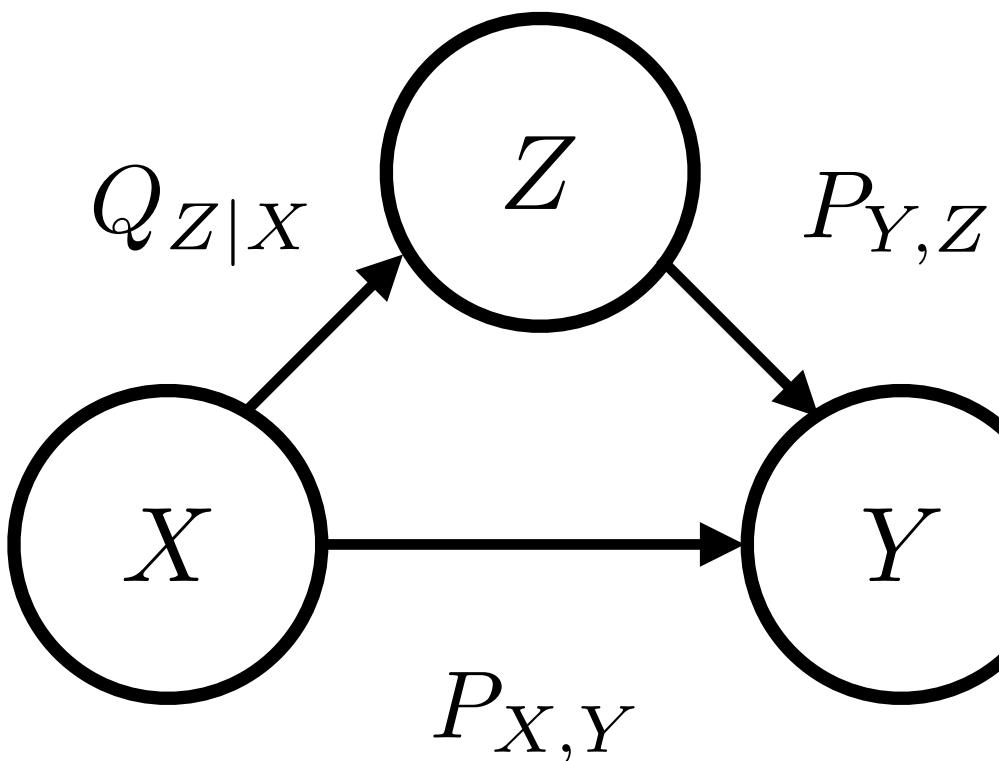
Y = label

Z = caption

Theorem. (Mehta & Harchaoui, ICML '25)

$$\mathbb{E}_{X \sim P_X} [(f_\star(X) - \bar{f}(X))^2] \lesssim I(X, Y|Z) + \text{err}(P_{Y,Z}, \rho_{Y,Z})$$

$P_{X,Y,Z}$ denotes any joint distribution such that
 $P_{Z|X} = Q_{Z|X}$.



$$\mathbb{E}_{X \sim P_X} [(f_\star(X) - \hat{f}(X))^2] \leq 2\mathbb{E}_{X \sim P_X} [(f_\star(X) - \bar{f}(X))^2] + 2\mathbb{E}_{X \sim P_X} [(\bar{f}(X) - \hat{f}(X))^2]$$

The diagram illustrates the decomposition of the expected squared error. A horizontal pink line segment connects two vertical dotted lines. The left vertical dotted line is labeled "direct predictor" below it and "ZSP procedure" above it. The right vertical dotted line is labeled "population version of ZSP" below it and "(based on distributions instead of samples)" above it. The pink line segment is labeled "information-theoretic error" in pink text below it. A green line segment extends from the right vertical dotted line to the right, labeled "learning error" in green text below it.

Roadmap of Theoretical Analysis

1. Define \bar{f} in terms of pre-training, evaluation, and prompting distribution.
 2. Upper bound information-theoretic error using dependence relationships between images, captions, and labels.
 3. Define class of estimators \hat{f} , and bound learning error using tools from statistical learning theory.

$X = \text{image}$

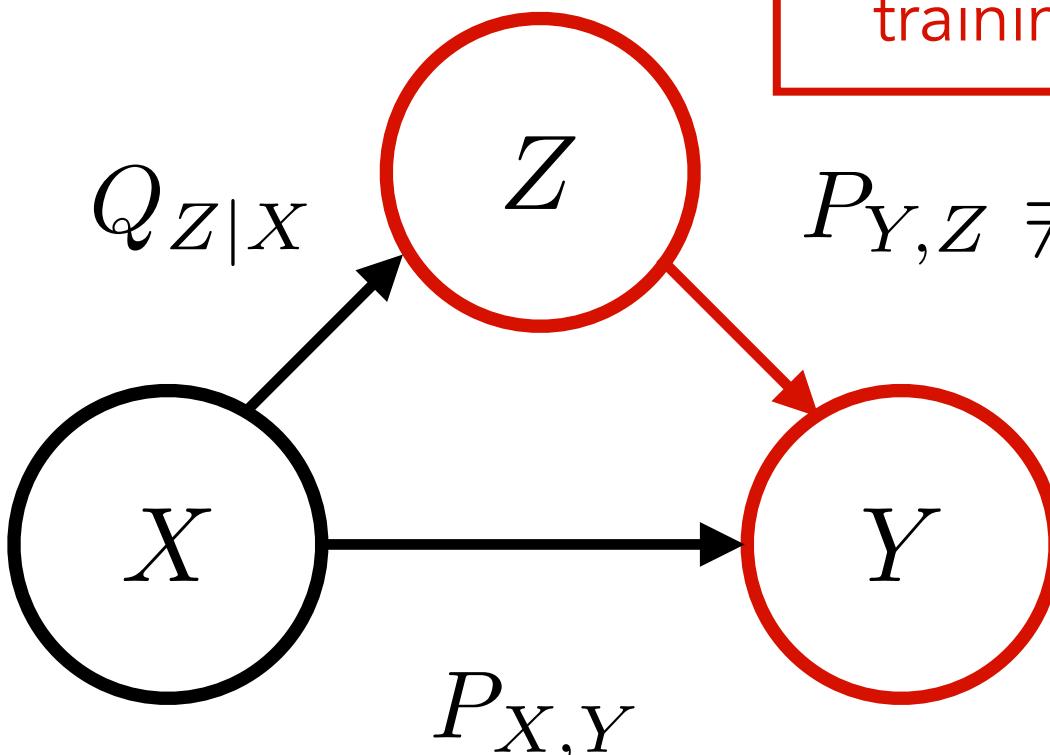
$Y = \text{label}$

Z = caption

Theorem. (Mehta & Harchaoui, ICML '25)

$$\mathbb{E}_{X \sim P_X} \left[(f_\star(X) - \bar{f}(X))^2 \right] \lesssim I(X, Y | Z) + \text{err}(P_{Y,Z}, \rho_{Y,Z})$$

$P_{X,Y,Z}$ denotes any joint distribution such that



Prompt “bias”, or incompatibility of the prompt distribution with pre-training/evaluation distributions.

Prompts	Captions
photo of a ship	Cruise ship in the Bahamas
photo of a car	Selling car for cheap
photo of a horse	I love horses

$$\mathbb{E}_{X \sim P_X} [(f_\star(X) - \hat{f}(X))^2] \leq 2\mathbb{E}_{X \sim P_X} [(f_\star(X) - \bar{f}(X))^2] + 2\mathbb{E}_{X \sim P_X} [(\bar{f}(X) - \hat{f}(X))^2]$$

The diagram illustrates the decomposition of the total error into two components. A horizontal dotted line is divided into two segments: a pink segment labeled "information-theoretic error" and a green segment labeled "learning error". Dotted vertical lines from the endpoints of these segments point down to the terms in the inequality. The pink segment corresponds to the term $2\mathbb{E}_{X \sim P_X} [(f_\star(X) - \bar{f}(X))^2]$, and the green segment corresponds to the term $2\mathbb{E}_{X \sim P_X} [(\bar{f}(X) - \hat{f}(X))^2]$. Below the pink segment, the text "population version of ZSP (based on distributions instead of samples)" is written.

information-theoretic error learning error

direct predictor ZSP procedure

population version of ZSP
(based on distributions instead of samples)

Roadmap of Theoretical Analysis

1. Define \bar{f} in terms of pre-training, evaluation, and prompting distribution.
2. Upper bound **information-theoretic error** using dependence relationships between images, captions, and labels.
3. Define class of estimators \hat{f} , and bound **learning error** using tools from statistical learning theory.

$X =$ image

$Y =$ label

$Z =$ caption

$$\mathbb{E}_{X \sim P_X} [(f_\star(X) - \hat{f}(X))^2] \leq 2\mathbb{E}_{X \sim P_X} [(f_\star(X) - \bar{f}(X))^2] + 2\mathbb{E}_{X \sim P_X} [(\bar{f}(X) - \hat{f}(X))^2]$$

direct predictor ZSP procedure

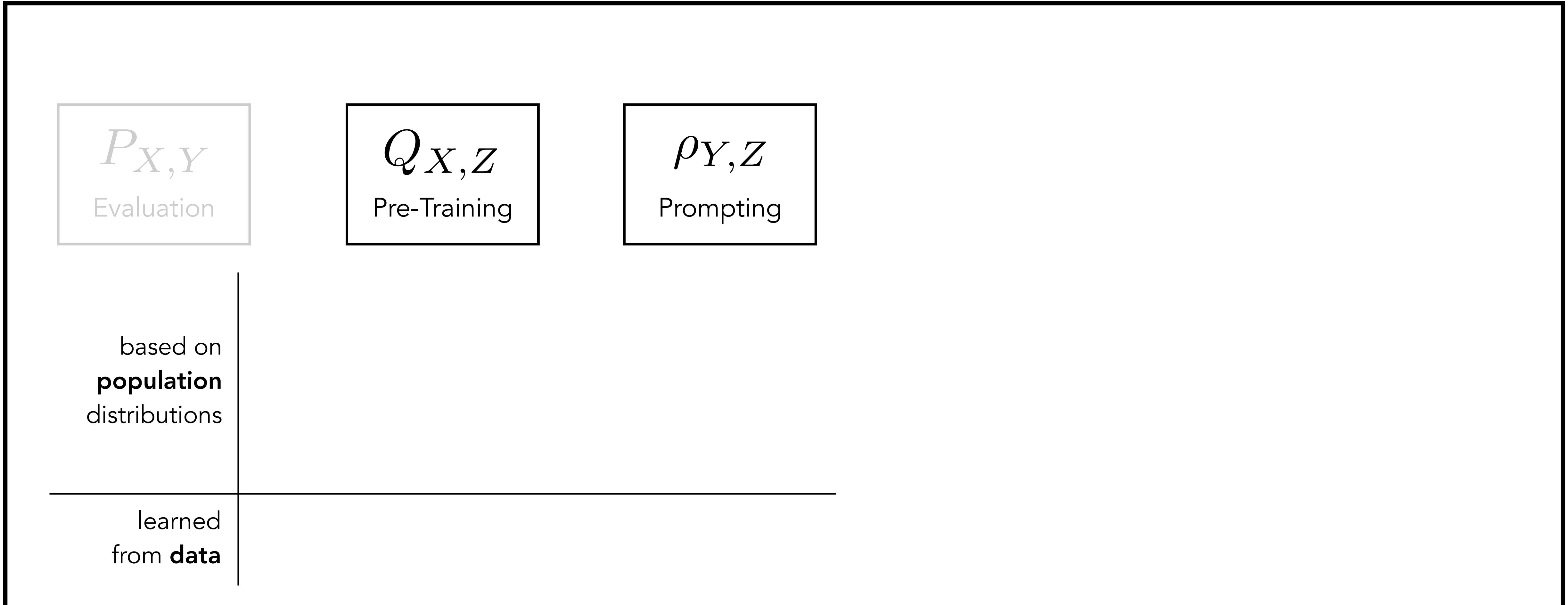
information-theoretic error

learning error

population version of ZSP
(based on distributions instead of samples)

$$\mathbb{E}_{X \sim P_X} [(f_\star(X) - \hat{f}(X))^2] \leq 2\mathbb{E}_{X \sim P_X} [(f_\star(X) - \bar{f}(X))^2] + 2\mathbb{E}_{X \sim P_X} [(\bar{f}(X) - \hat{f}(X))^2]$$

direct predictor
 ZSP procedure
 information-theoretic error
 learning error
 population version of ZSP
 (based on distributions instead of samples)



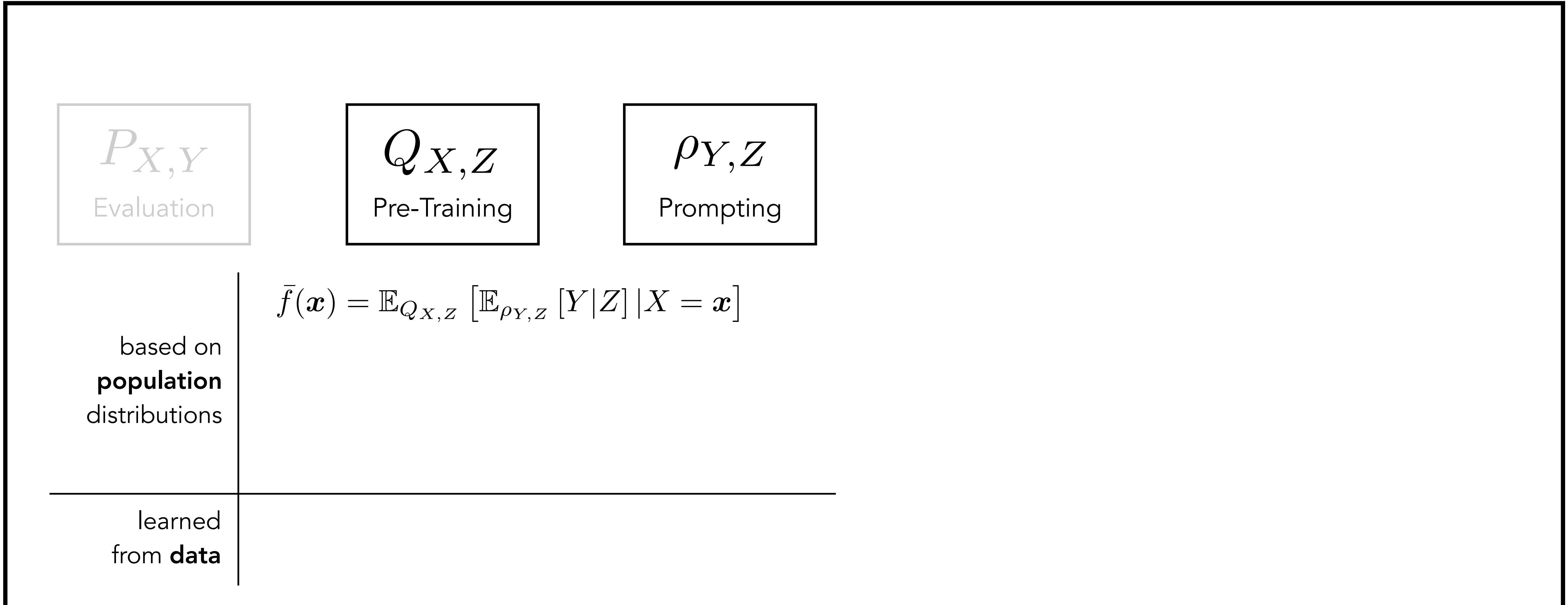
$$\mathbb{E}_{X \sim P_X} [(f_\star(X) - \hat{f}(X))^2] \leq 2\mathbb{E}_{X \sim P_X} [(f_\star(X) - \bar{f}(X))^2] + 2\mathbb{E}_{X \sim P_X} [(\bar{f}(X) - \hat{f}(X))^2]$$

A diagram illustrating the decomposition of the error. It shows three components: a yellow bar at the top labeled "direct predictor" and "ZSP procedure", a pink bar below it labeled "information-theoretic error", and a green bar below the pink one labeled "learning error". Dotted lines connect the start of the pink bar to the end of the yellow bar, and the start of the green bar to the end of the pink bar.

information-theoretic error

learning error

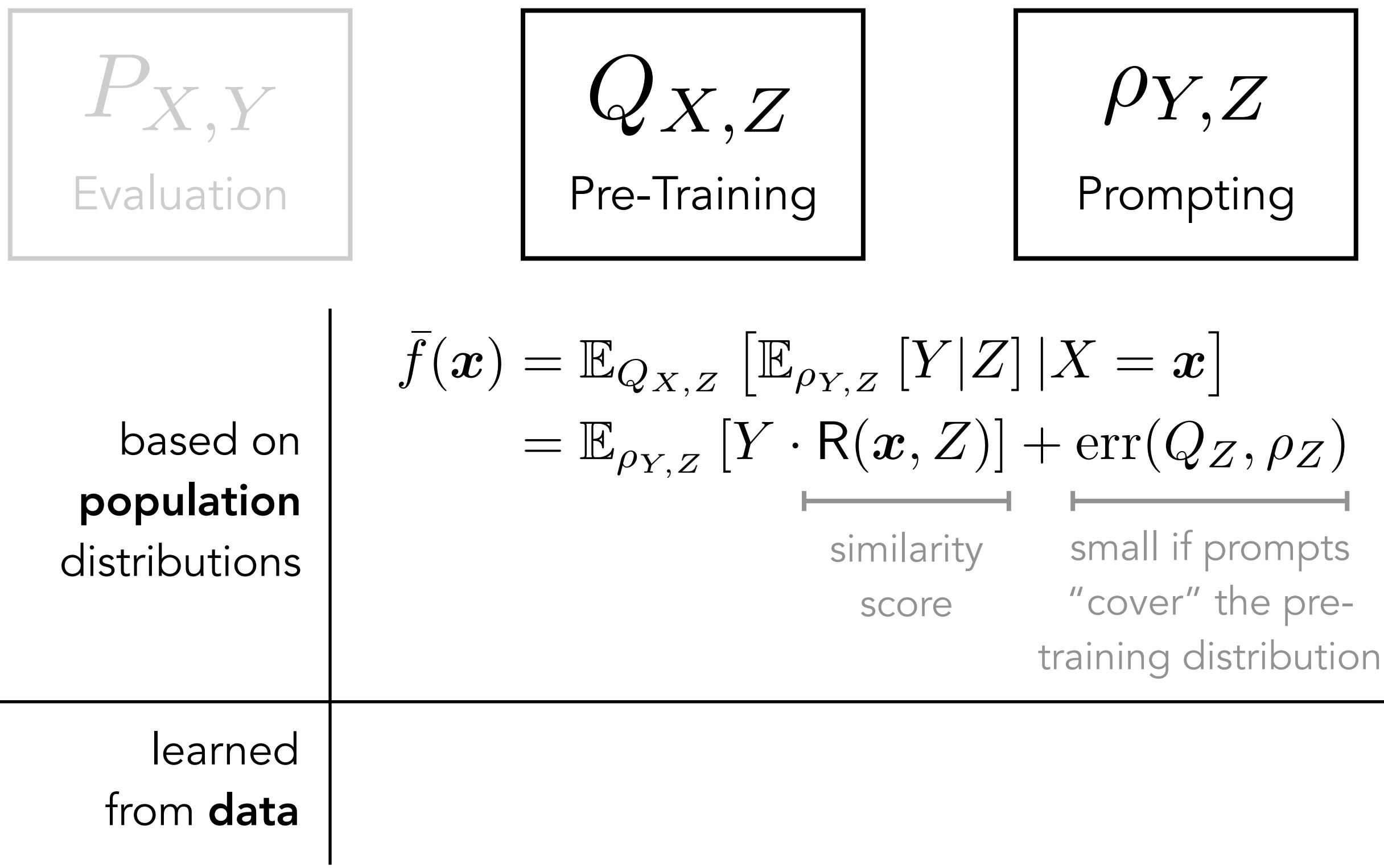
population version of ZSP
(based on distributions instead of samples)



$$\mathbb{E}_{X \sim P_X} [(f_\star(X) - \hat{f}(X))^2] \leq 2\mathbb{E}_{X \sim P_X} [(f_\star(X) - \bar{f}(X))^2] + 2\mathbb{E}_{X \sim P_X} [(\bar{f}(X) - \hat{f}(X))^2]$$

The diagram illustrates the decomposition of the expected squared error. A horizontal pink line segment connects two vertical dotted lines. The left vertical dotted line is labeled "direct predictor" below it and "ZSP procedure" above it. The right vertical dotted line is labeled "learning error" below it and "population version of ZSP" above it, followed by the note "(based on distributions instead of samples)". The pink line segment is labeled "information-theoretic error" in pink text.

Approach 1: Similarity Score Learning



$$\mathbb{E}_{X \sim P_X} [(f_\star(X) - \hat{f}(X))^2] \leq 2\mathbb{E}_{X \sim P_X} [(f_\star(X) - \bar{f}(X))^2] + 2\mathbb{E}_{X \sim P_X} [(\bar{f}(X) - \hat{f}(X))^2]$$

A diagram illustrating the decomposition of the error. It shows three components: "direct predictor" (yellow), "ZSP procedure" (orange), and "population version of ZSP (based on distributions instead of samples)" (yellow). A pink horizontal bar labeled "information-theoretic error" spans from the direct predictor to the population version of ZSP. A green horizontal bar labeled "learning error" spans from the ZSP procedure to the population version of ZSP. Dotted lines connect the endpoints of these bars to the respective components.

information-theoretic error

learning error

direct predictor ZSP procedure population version of ZSP
(based on distributions instead of samples)

Approach 1: Similarity Score Learning

$P_{X,Y}$
Evaluation

$Q_{X,Z}$
Pre-Training

$\rho_{Y,Z}$
Prompting

based on **population** distributions

$$\begin{aligned}
 \bar{f}(\mathbf{x}) &= \mathbb{E}_{Q_{X,Z}} [\mathbb{E}_{\rho_{Y,Z}} [Y|Z] | X = \mathbf{x}] \\
 &= \mathbb{E}_{\rho_{Y,Z}} [Y \cdot R(\mathbf{x}, Z)] + \text{err}(Q_Z, \rho_Z)
 \end{aligned}$$

A diagram illustrating the decomposition of the population version of \bar{f} . It shows two terms: "similarity score" (grey) and "small if prompts 'cover' the pre-training distribution" (grey).

learned from **data**

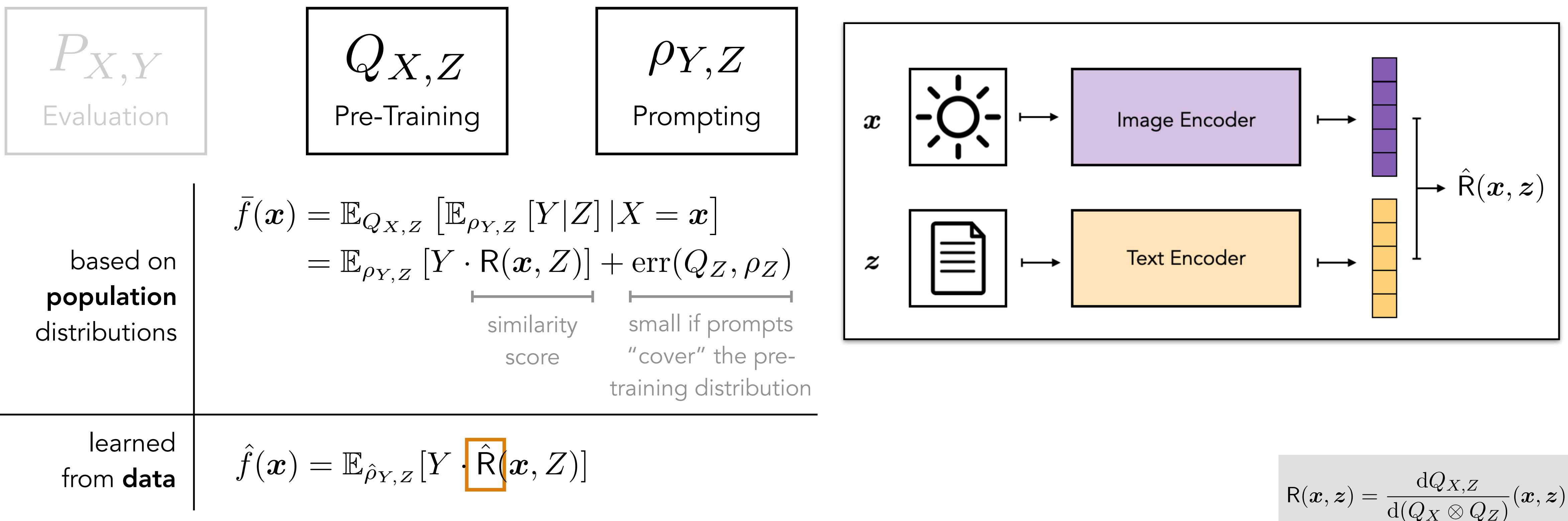
$$\hat{f}(\mathbf{x}) = \mathbb{E}_{\hat{\rho}_{Y,Z}} [Y \cdot \hat{R}(\mathbf{x}, Z)]$$

$$R(\mathbf{x}, \mathbf{z}) = \frac{dQ_{X,Z}}{d(Q_X \otimes Q_Z)}(\mathbf{x}, \mathbf{z})$$

$$\mathbb{E}_{X \sim P_X} [(f_\star(X) - \hat{f}(X))^2] \leq 2\mathbb{E}_{X \sim P_X} [(f_\star(X) - \bar{f}(X))^2] + 2\mathbb{E}_{X \sim P_X} [(\bar{f}(X) - \hat{f}(X))^2]$$

The diagram illustrates the decomposition of the expected squared error. A horizontal pink line segment connects two vertical dotted lines. The left vertical dotted line is labeled "direct predictor" below it and "ZSP procedure" above it. The right vertical dotted line is labeled "population version of ZSP" below it and "(based on distributions instead of samples)" above it. The pink line segment is labeled "information-theoretic error" in pink text below it. A green line segment extends from the right vertical dotted line to the right, labeled "learning error" in green text below it.

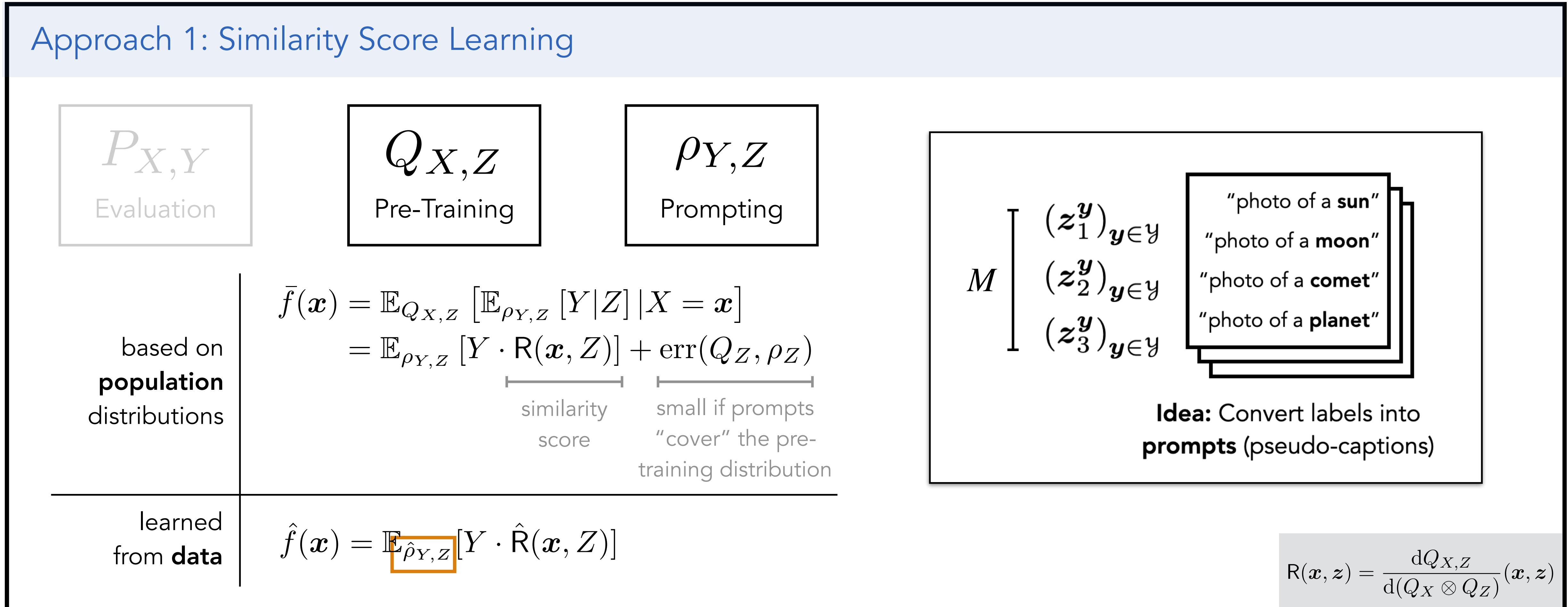
Approach 1: Similarity Score Learning



$$\mathbb{E}_{X \sim P_X} [(f_\star(X) - \hat{f}(X))^2] \leq 2\mathbb{E}_{X \sim P_X} [(f_\star(X) - \bar{f}(X))^2] + 2\mathbb{E}_{X \sim P_X} [(\bar{f}(X) - \hat{f}(X))^2]$$

The diagram illustrates the decomposition of the expected squared error. A horizontal pink line segment connects two vertical dotted lines. The left vertical dotted line is labeled "direct predictor" below it and "ZSP procedure" above it. The right vertical dotted line is labeled "learning error" below it and "population version of ZSP" above it, followed by the note "(based on distributions instead of samples)". The pink line segment is labeled "information-theoretic error" in pink text.

Approach 1: Similarity Score Learning



$$\mathbb{E}_{X \sim P_X} [(f_\star(X) - \hat{f}(X))^2] \leq 2\mathbb{E}_{X \sim P_X} [(f_\star(X) - \bar{f}(X))^2] + 2\mathbb{E}_{X \sim P_X} [(\bar{f}(X) - \hat{f}(X))^2]$$

The diagram illustrates the decomposition of the total error into two components. A horizontal pink line labeled "information-theoretic error" connects the first term $\mathbb{E}_{X \sim P_X} [(f_\star(X) - \bar{f}(X))^2]$ to the second term $\mathbb{E}_{X \sim P_X} [(\bar{f}(X) - \hat{f}(X))^2]$. A green line labeled "learning error" connects the second term to the final expression $\mathbb{E}_{X \sim P_X} [(f_\star(X) - \hat{f}(X))^2]$. Below the terms, vertical dotted lines connect them to their respective labels: "direct predictor" under $f_\star(X)$, "ZSP procedure" under $\hat{f}(X)$, "population version of ZSP (based on distributions instead of samples)" under $\bar{f}(X)$.

Approach 1: Similarity Score Learning

Theorem. (Mehta & Harchaoui, ICML '25)

$$\mathbb{E}_{X \sim P_X} [(\bar{f}(X) - \hat{f}(X))^2] \lesssim$$

based on **population** distributions

$$\begin{aligned} \bar{f}(\mathbf{x}) &= \mathbb{E}_{Q_{X,Z}} [\mathbb{E}_{\rho_{Y,Z}} [Y|Z] | X = \mathbf{x}] \\ &= \mathbb{E}_{\rho_{Y,Z}} [Y \cdot R(\mathbf{x}, Z)] + \text{err}(Q_Z, \rho_Z) \end{aligned}$$

similarity score small if prompts "cover" the pre-training distribution

learned from **data**

$$\hat{f}(\mathbf{x}) = \mathbb{E}_{\hat{\rho}_{Y,Z}} [Y \cdot \hat{R}(\mathbf{x}, Z)]$$

$$R(\mathbf{x}, \mathbf{z}) = \frac{dQ_{X,Z}}{d(Q_X \otimes Q_Z)}(\mathbf{x}, \mathbf{z})$$

$$\mathbb{E}_{X \sim P_X} [(f_\star(X) - \hat{f}(X))^2] \leq 2\mathbb{E}_{X \sim P_X} [(f_\star(X) - \bar{f}(X))^2] + 2\mathbb{E}_{X \sim P_X} [(\bar{f}(X) - \hat{f}(X))^2]$$

The diagram illustrates the decomposition of the total error into two components. A horizontal pink line labeled "information-theoretic error" connects the first term $\mathbb{E}_{X \sim P_X} [(f_\star(X) - \bar{f}(X))^2]$ to the second term $\mathbb{E}_{X \sim P_X} [(\bar{f}(X) - \hat{f}(X))^2]$. A green line labeled "learning error" connects the second term to the final expression $2\mathbb{E}_{X \sim P_X} [(\bar{f}(X) - \hat{f}(X))^2]$. Vertical dotted lines from the terms point to labels below: "direct predictor" under $f_\star(X)$, "ZSP procedure" under $\hat{f}(X)$, "population version of ZSP (based on distributions instead of samples)" under $\bar{f}(X)$, and "learning error" under the final term.

Approach 1: Similarity Score Learning

Theorem. (Mehta & Harchaoui, ICML '25)

$$\mathbb{E}_{X \sim P_X} [(\bar{f}(X) - \hat{f}(X))^2] \lesssim d(\hat{\mathbf{R}}, \mathbf{R}) + d(\hat{\rho}_{Y,Z}, \rho_{Y,Z})$$

based on
population
distributions

$$\begin{aligned} \bar{f}(\mathbf{x}) &= \mathbb{E}_{Q_{X,Z}} [\mathbb{E}_{\rho_{Y,Z}} [Y|Z] | X = \mathbf{x}] \\ &= \mathbb{E}_{\rho_{Y,Z}} [Y \cdot \mathbf{R}(\mathbf{x}, Z)] + \text{err}(Q_Z, \rho_Z) \end{aligned}$$

similarity score small if prompts "cover" the pre-training distribution

learned
from **data**

$$\hat{f}(\mathbf{x}) = \mathbb{E}_{\hat{\rho}_{Y,Z}} [Y \cdot \hat{\mathbf{R}}(\mathbf{x}, Z)]$$

$$\mathbf{R}(\mathbf{x}, \mathbf{z}) = \frac{dQ_{X,Z}}{d(Q_X \otimes Q_Z)}(\mathbf{x}, \mathbf{z})$$

$$\mathbb{E}_{X \sim P_X} [(f_\star(X) - \hat{f}(X))^2] \leq 2\mathbb{E}_{X \sim P_X} [(f_\star(X) - \bar{f}(X))^2] + 2\mathbb{E}_{X \sim P_X} [(\bar{f}(X) - \hat{f}(X))^2]$$



 direct predictor ZSP procedure population version of ZSP
 (based on distributions instead of samples)

Approach 1: Similarity Score Learning

Theorem. (Mehta & Harchaoui, ICML '25)

$$\mathbb{E}_{X \sim P_X} [(\bar{f}(X) - \hat{f}(X))^2] \lesssim d(\hat{\mathbf{R}}, \mathbf{R}) + d(\hat{\rho}_{Y,Z}, \rho_{Y,Z})$$

sample complexity

$$\left(\frac{1}{N^\square} \right)$$

prompt complexity

$$\left(\frac{1}{M^\square} \right)$$

based on **population** distributions

$$\begin{aligned} \bar{f}(\mathbf{x}) &= \mathbb{E}_{Q_{X,Z}} [\mathbb{E}_{\rho_{Y,Z}} [Y|Z] | X = \mathbf{x}] \\ &= \mathbb{E}_{\rho_{Y,Z}} [Y \cdot \mathbf{R}(\mathbf{x}, Z)] + \text{err}(Q_Z, \rho_Z) \end{aligned}$$

similarity score small if prompts "cover" the pre-training distribution

learned from **data**

$$\hat{f}(\mathbf{x}) = \mathbb{E}_{\hat{\rho}_{Y,Z}} [Y \cdot \hat{\mathbf{R}}(\mathbf{x}, Z)]$$

$$\mathbf{R}(\mathbf{x}, \mathbf{z}) = \frac{dQ_{X,Z}}{d(Q_X \otimes Q_Z)}(\mathbf{x}, \mathbf{z})$$

$$\mathbb{E}_{X \sim P_X} [(f_\star(X) - \hat{f}(X))^2] \leq 2\mathbb{E}_{X \sim P_X} [(f_\star(X) - \bar{f}(X))^2] + 2\mathbb{E}_{X \sim P_X} [(\bar{f}(X) - \hat{f}(X))^2]$$

The diagram illustrates the decomposition of the expected squared error. A horizontal pink line segment connects two vertical dotted lines. The left vertical dotted line is labeled "direct predictor" below it and "ZSP procedure" above it. The right vertical dotted line is labeled "population version of ZSP" below it and "(based on distributions instead of samples)" above it. The pink line segment is labeled "information-theoretic error" in pink text below it. A green line segment extends from the right vertical dotted line to the right, labeled "learning error" in green text below it.

Approach 1: Similarity Score Learning

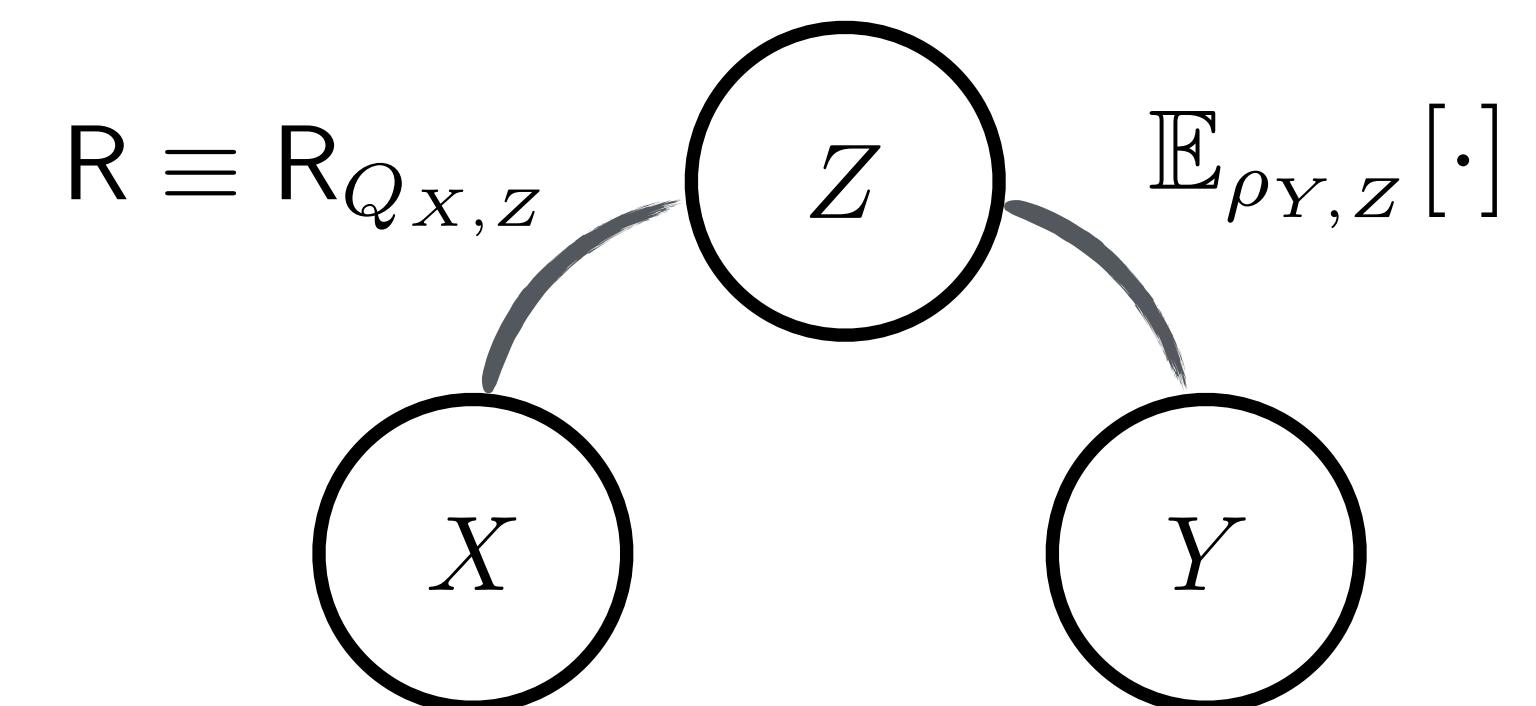
Theorem. (Mehta & Harchaoui, ICML '25)

$$\mathbb{E}_{X \sim P_X} \left[(\bar{f}(X) - \hat{f}(X))^2 \right] \lesssim d(\hat{\mathsf{R}}, \mathsf{R}) + d(\hat{\rho}_{Y,Z}, \rho_{Y,Z})$$

based on
population
distributions

learned from **data**

$$\hat{f}(\mathbf{x}) = \mathbb{E}_{\hat{\rho}_{Y,Z}}[Y \cdot \hat{\mathsf{R}}(\mathbf{x}, Z)]$$



$$\mathsf{R}(x, z) = \frac{\mathrm{d}Q_{X,Z}}{\mathrm{d}(Q_X \otimes Q_Z)}(x, z)$$

$$\mathbb{E}_{X \sim P_X} [(f_\star(X) - \hat{f}(X))^2] \leq 2\mathbb{E}_{X \sim P_X} [(f_\star(X) - \bar{f}(X))^2] + 2\mathbb{E}_{X \sim P_X} [(\bar{f}(X) - \hat{f}(X))^2]$$

direct predictor ZSP procedure

information-theoretic error

learning error

population version of ZSP
(based on distributions instead of samples)

Approach 2: Two-Stage Prediction

Theorem. (Mehta & Harchaoui, ICML '25)

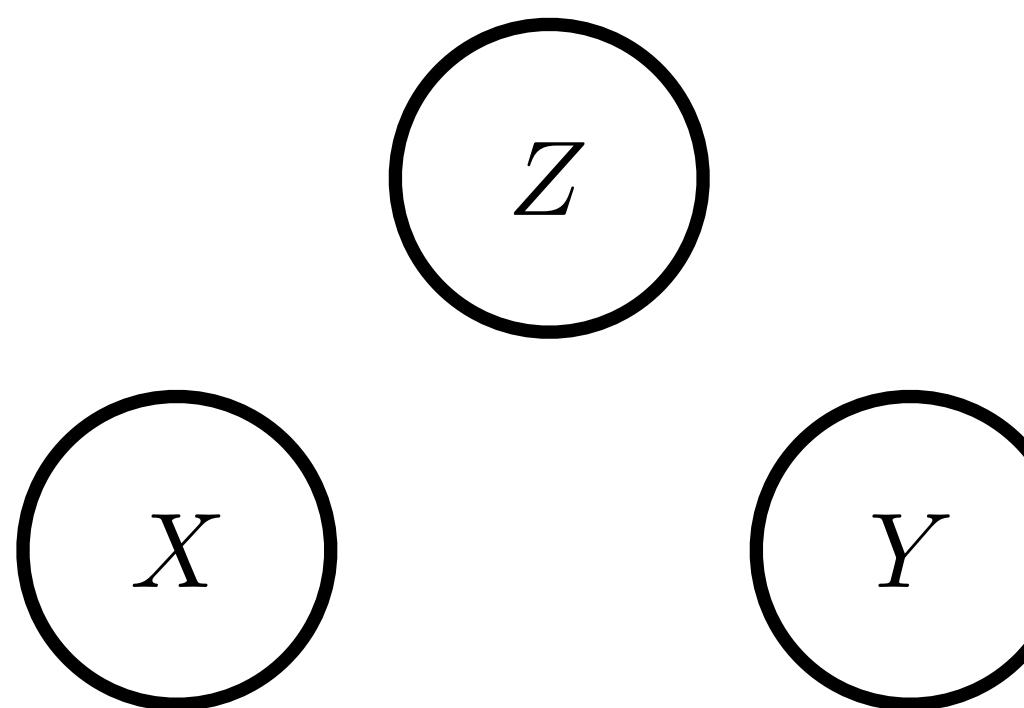
$$\mathbb{E}_{X \sim P_X} \left[(\bar{f}(X) - \hat{f}(X))^2 \right] \lesssim$$

based on
population
distributions

$$\bar{f}(x) = \mathbb{E}_{Q_{X,Z}} [\mathbb{E}_{\rho_{Y,Z}} [Y|Z] | X = x]$$

learned from **data**

$$\hat{f}(x) = \hat{g}_M(\hat{h}_N(x))$$



$$\mathbb{E}_{X \sim P_X} [(f_\star(X) - \hat{f}(X))^2] \leq 2\mathbb{E}_{X \sim P_X} [(f_\star(X) - \bar{f}(X))^2] + 2\mathbb{E}_{X \sim P_X} [(\bar{f}(X) - \hat{f}(X))^2]$$

direct predictor ZSP procedure

information-theoretic error

learning error

population version of ZSP
(based on distributions instead of samples)

Approach 2: Two-Stage Prediction

Theorem. (Mehta & Harchaoui, ICML '25)

$$\mathbb{E}_{X \sim P_X} \left[(\bar{f}(X) - \hat{f}(X))^2 \right] \lesssim$$

based on
population
distributions

$$\bar{f}(x) = \mathbb{E}_{Q_{X,Z}} \left[\mathbb{E}_{\rho_{Y,Z}} [Y|Z] | X = x \right]$$


1st Stage

learned
from **data**

$$\hat{f}(x) = \hat{g}_M(\hat{h}_N(x))$$

1st Stage Prediction from N Examples



Text Caption:

Look at my
cute cat!

$$\mathbb{E}_{X \sim P_X} [(f_\star(X) - \hat{f}(X))^2] \leq 2\mathbb{E}_{X \sim P_X} [(f_\star(X) - \bar{f}(X))^2] + 2\mathbb{E}_{X \sim P_X} [(\bar{f}(X) - \hat{f}(X))^2]$$

The diagram illustrates the decomposition of the expected squared error. A horizontal pink line segment connects two vertical dotted lines. The left vertical dotted line is labeled "direct predictor" below it and "ZSP procedure" above it. The right vertical dotted line is labeled "learning error" below it and "population version of ZSP" above it, followed by the note "(based on distributions instead of samples)". The pink line segment is labeled "information-theoretic error" in pink text.

Approach 2: Two-Stage Prediction

Theorem. (Mehta & Harchaoui, ICML '25)

$$\mathbb{E}_{X \sim P_X} \left[(\bar{f}(X) - \hat{f}(X))^2 \right] \lesssim$$

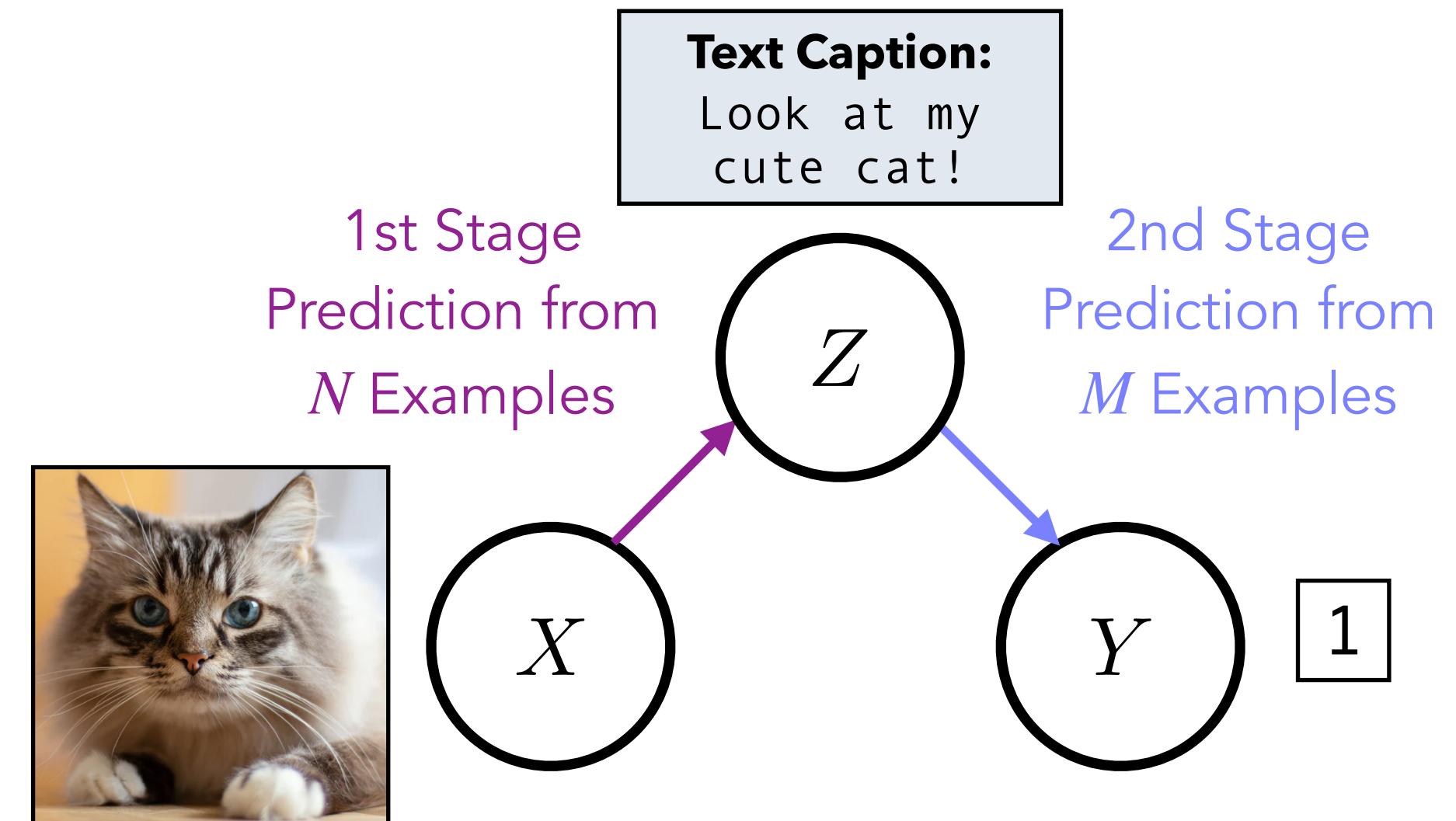
based on
population
distributions

$$\bar{f}(x) = \mathbb{E}_{Q_{X,Z}} \left[\mathbb{E}_{\rho_{Y,Z}} [Y|Z] | X = x \right]$$


The diagram illustrates the two-stage nature of the function $\bar{f}(x)$. A purple bracket underlines the inner expectation term $\mathbb{E}_{\rho_{Y,Z}} [Y|Z]$, labeled "1st Stage". A blue bracket underlines the entire expression $\mathbb{E}_{Q_{X,Z}} \left[\cdot \right]$, labeled "2nd Stage".

learned
from **data**

$$\hat{f}(x) = \hat{g}_M(\hat{h}_N(x))$$



$$\mathbb{E}_{X \sim P_X} [(f_\star(X) - \hat{f}(X))^2] \leq 2\mathbb{E}_{X \sim P_X} [(f_\star(X) - \bar{f}(X))^2] + 2\mathbb{E}_{X \sim P_X} [(\bar{f}(X) - \hat{f}(X))^2]$$

The diagram illustrates the decomposition of the expected squared error. A horizontal pink line segment connects two vertical dotted lines. The left vertical dotted line is labeled "direct predictor" below it and "ZSP procedure" above it. The right vertical dotted line is labeled "learning error" below it and "population version of ZSP" above it, followed by the note "(based on distributions instead of samples)". The pink line segment is labeled "information-theoretic error" in pink text.

Approach 2: Two-Stage Prediction

Theorem. (Mehta & Harchaoui, ICML '25)

$$\mathbb{E}_{X \sim P_X} \left[(\bar{f}(X) - \hat{f}(X))^2 \right] \lesssim \frac{1}{N^\square} + \frac{1}{M^\square}$$

based on
population
distributions

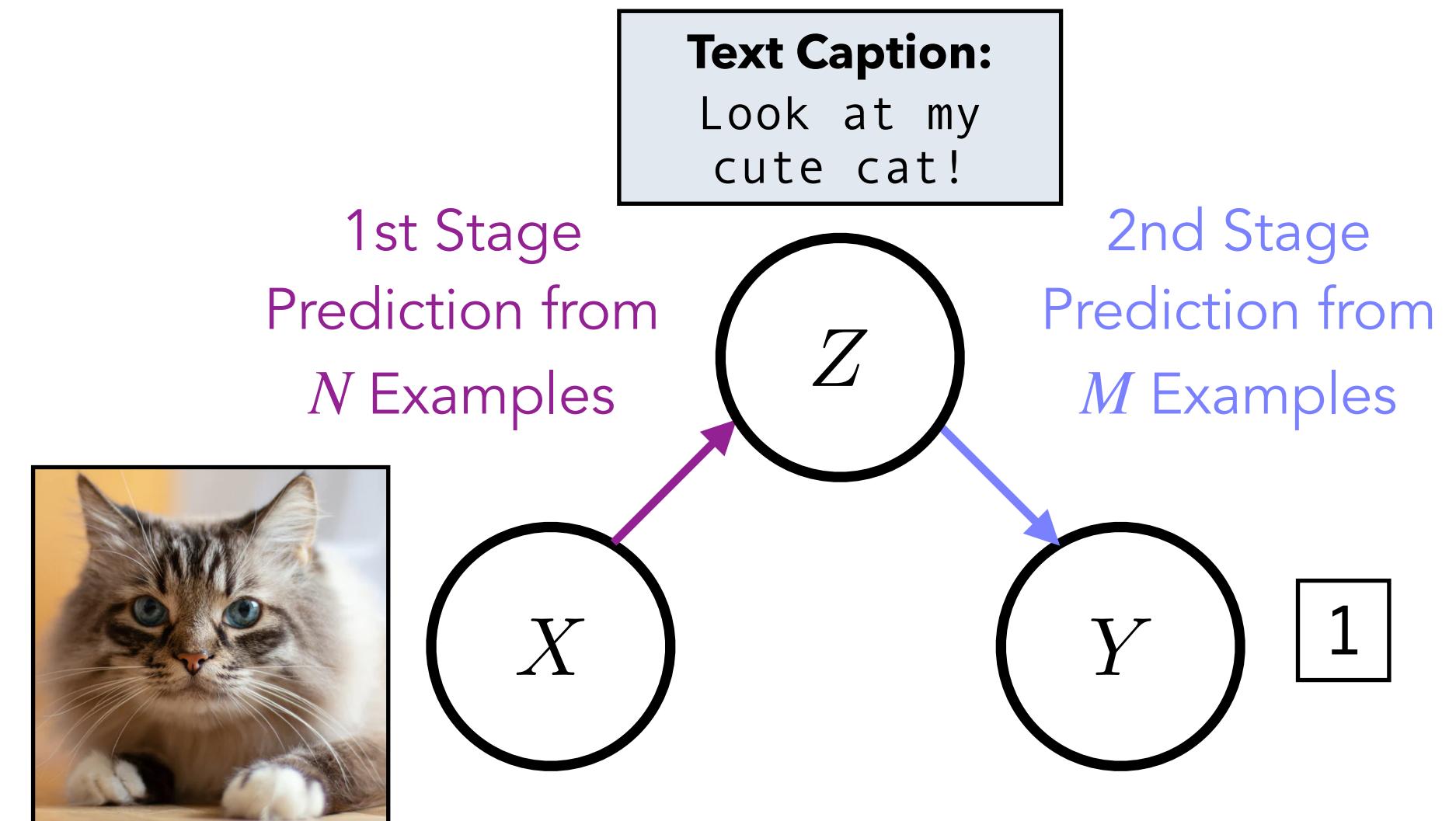
$$\bar{f}(x) = \mathbb{E}_{Q_{X,Z}} \left[\mathbb{E}_{\rho_{Y,Z}} [Y|Z] | X = x \right]$$


1st Stage

2nd Stage

learned from **data**

$$\hat{f}(x) = \hat{g}_M(\hat{h}_N(x))$$



$$\mathbb{E}_{X \sim P_X} \left[(f_\star(X) - \hat{f}(X))^2 \right] \leq 2\mathbb{E}_{X \sim P_X} \left[(f_\star(X) - \bar{f}(X))^2 \right] + 2\mathbb{E}_{X \sim P_X} \left[(\bar{f}(X) - \hat{f}(X))^2 \right]$$

information-theoretic error

learning error

direct predictor ZSP procedure

population version of ZSP
(based on distributions instead of samples)

$$\mathbb{E}_{X \sim P_X} [(f_\star(X) - \hat{f}(X))^2] \leq 2\mathbb{E}_{X \sim P_X} [(f_\star(X) - \bar{f}(X))^2] + 2\mathbb{E}_{X \sim P_X} [(\bar{f}(X) - \hat{f}(X))^2]$$

direct predictor
 ZSP procedure
 information-theoretic error
 learning error
 population version of ZSP
 (based on distributions instead of samples)

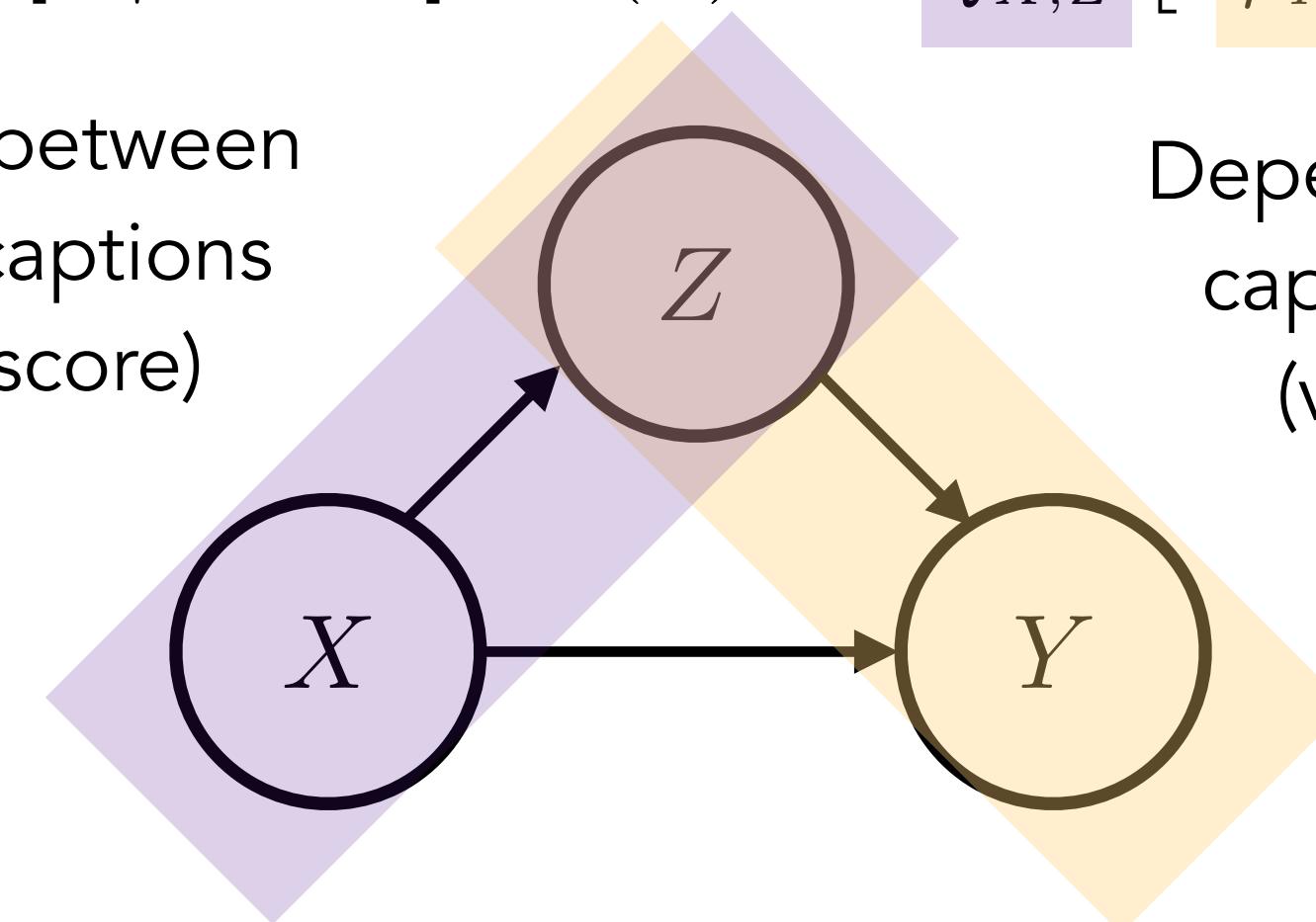
Contributions

1. Theoretical framework to formalize zero-shot prediction (ZSP) and obtain its generalization analysis.
2. Two proof strategies which apply to different classes of methods.
3. Key quantities for success of ZSP: **residual dependence**, **prompt bias**, **sample complexity**, and **prompt complexity**.

$P_{X,Y}$	$Q_{X,Z}$	$\rho_{Y,Z}$
Evaluation	Pre-Training	Prompting

$$f_\star(\mathbf{x}) = \mathbb{E}_{P_{X,Y}} [Y|X = \mathbf{x}] \quad \bar{f}(\mathbf{x}) = \mathbb{E}_{Q_{X,Z}} [\mathbb{E}_{\rho_{Y,Z}} [Y|Z] |X = \mathbf{x}]$$

Dependence between images and captions (e.g., CLIP score)



Dependence between captions and labels (via prompting)

$$\mathbb{E}_{X \sim P_X} [(f_\star(X) - \hat{f}(X))^2] \leq 2\mathbb{E}_{X \sim P_X} [(f_\star(X) - \bar{f}(X))^2] + 2\mathbb{E}_{X \sim P_X} [(\bar{f}(X) - \hat{f}(X))^2]$$

direct predictor ZSP procedure

information-theoretic error

learning error

residual dependence

prompt bias

population version of ZSP

(based on distributions instead of samples)

sample complexity

prompt complexity

Contributions

1. Theoretical framework to formalize zero-shot prediction (ZSP) and obtain its generalization analysis.
2. Two proof strategies which apply to different classes of methods.
3. Key quantities for success of ZSP: **residual dependence**, **prompt bias**, **sample complexity**, and **prompt complexity**.

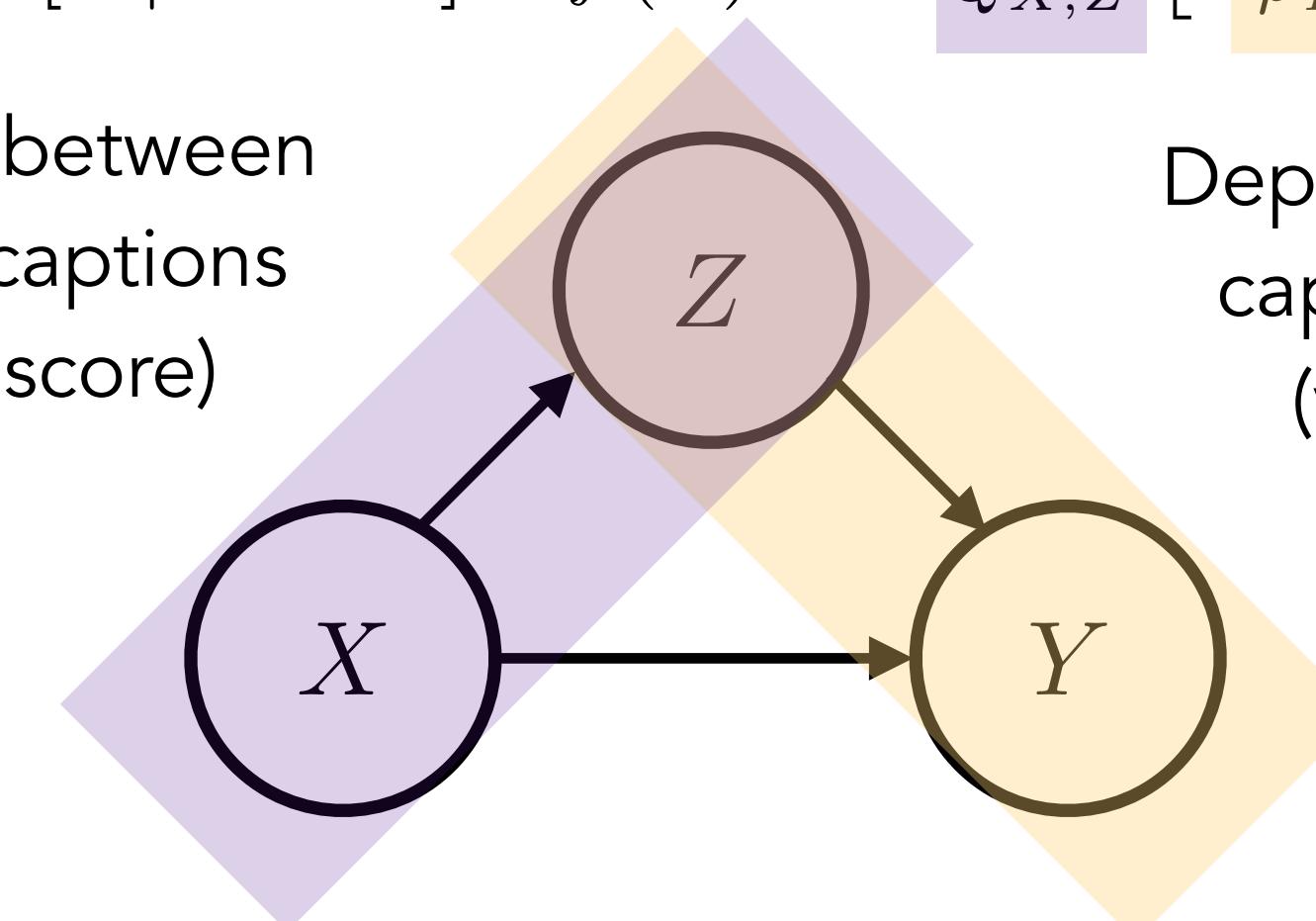
$P_{X,Y}$
Evaluation

$Q_{X,Z}$
Pre-Training

$\rho_{Y,Z}$
Prompting

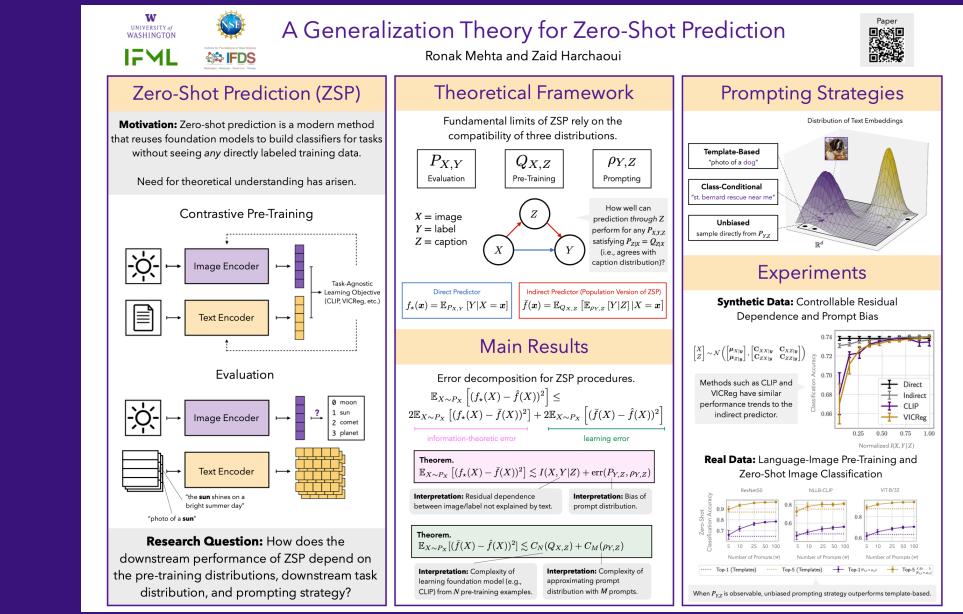
$$f_\star(\mathbf{x}) = \mathbb{E}_{P_{X,Y}} [Y|X = \mathbf{x}] \quad \bar{f}(\mathbf{x}) = \mathbb{E}_{Q_{X,Z}} [\mathbb{E}_{\rho_{Y,Z}} [Y|Z] |X = \mathbf{x}]$$

Dependence between
images and captions
(e.g., CLIP score)



Dependence between
captions and labels
(via prompting)

Thank you!



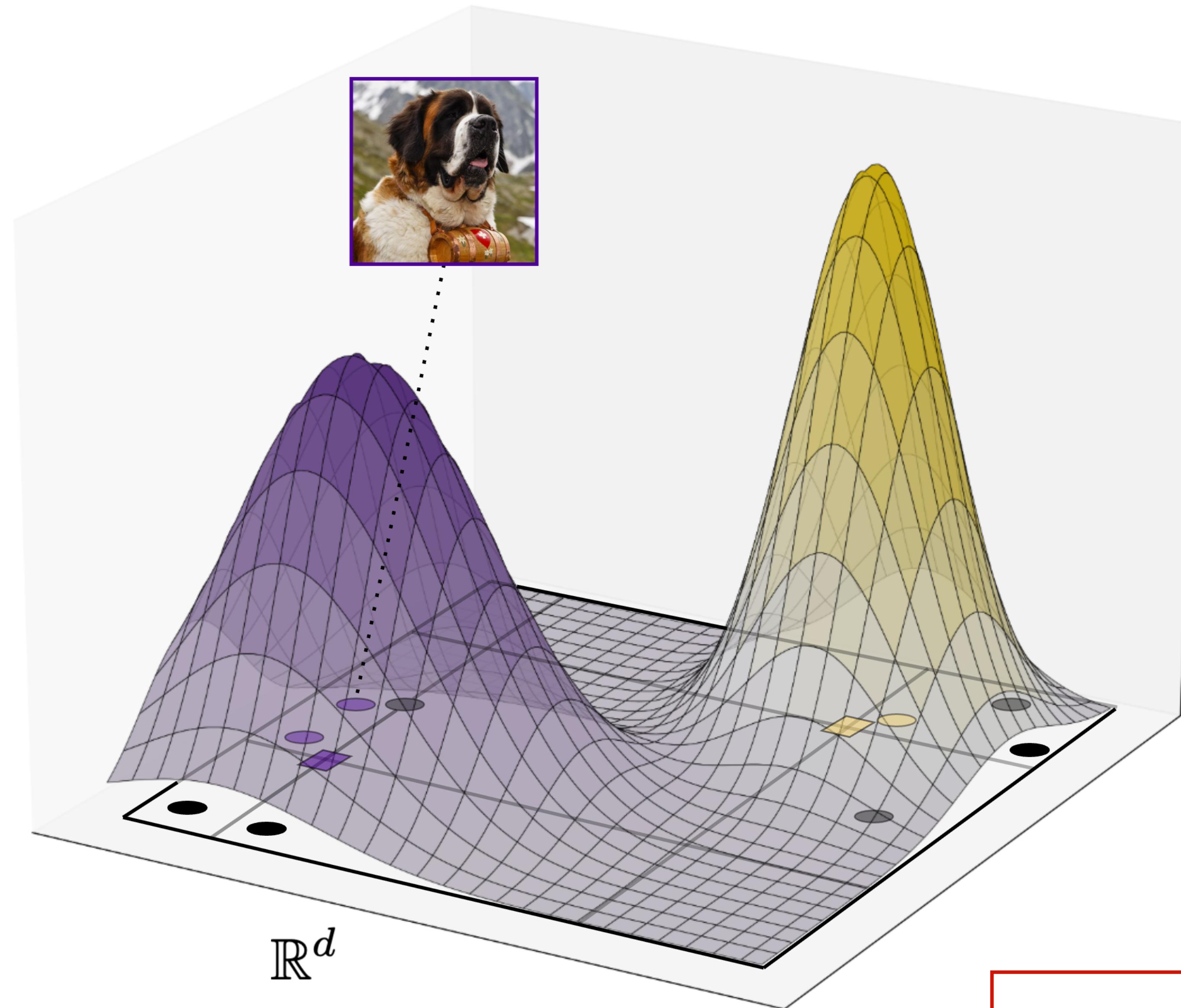
Spotlight Poster Info

Time: Wednesday, July 16, 11:00am PDT - 1:30pm PDT

Place: West Exhibition Hall B2-B3 #W-905



Appendix



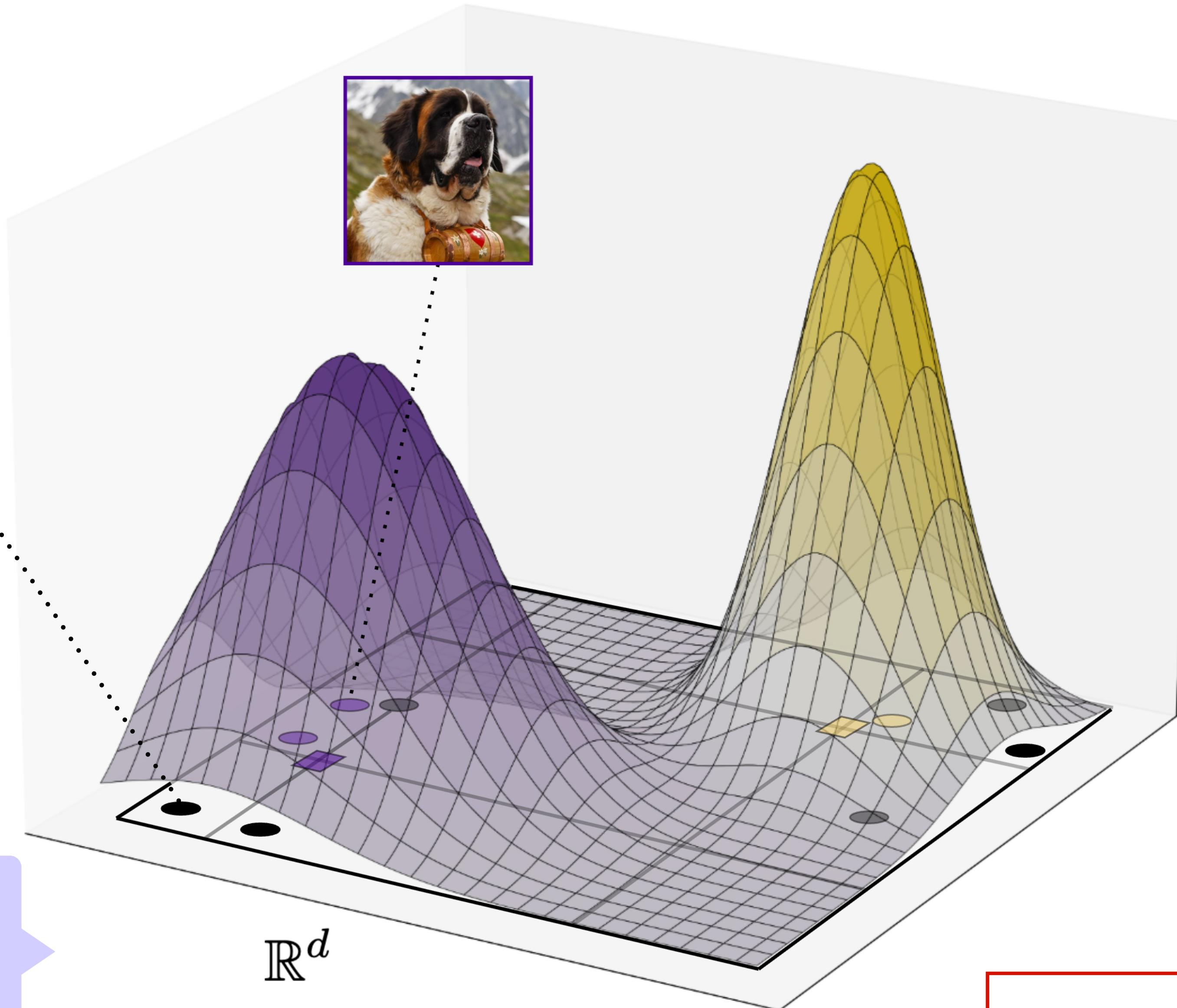
Distribution of Text Embeddings

Prompt Bias

$$\mathbb{E}_{Z \sim P_Z} [(\mathbb{E}_{P_{Y,Z}} [Y|Z] - \mathbb{E}_{\rho_{Y,Z}} [Y|Z])^2]$$

Template-Based
“photo of a dog”

does not easily
separate classes in
embedding space

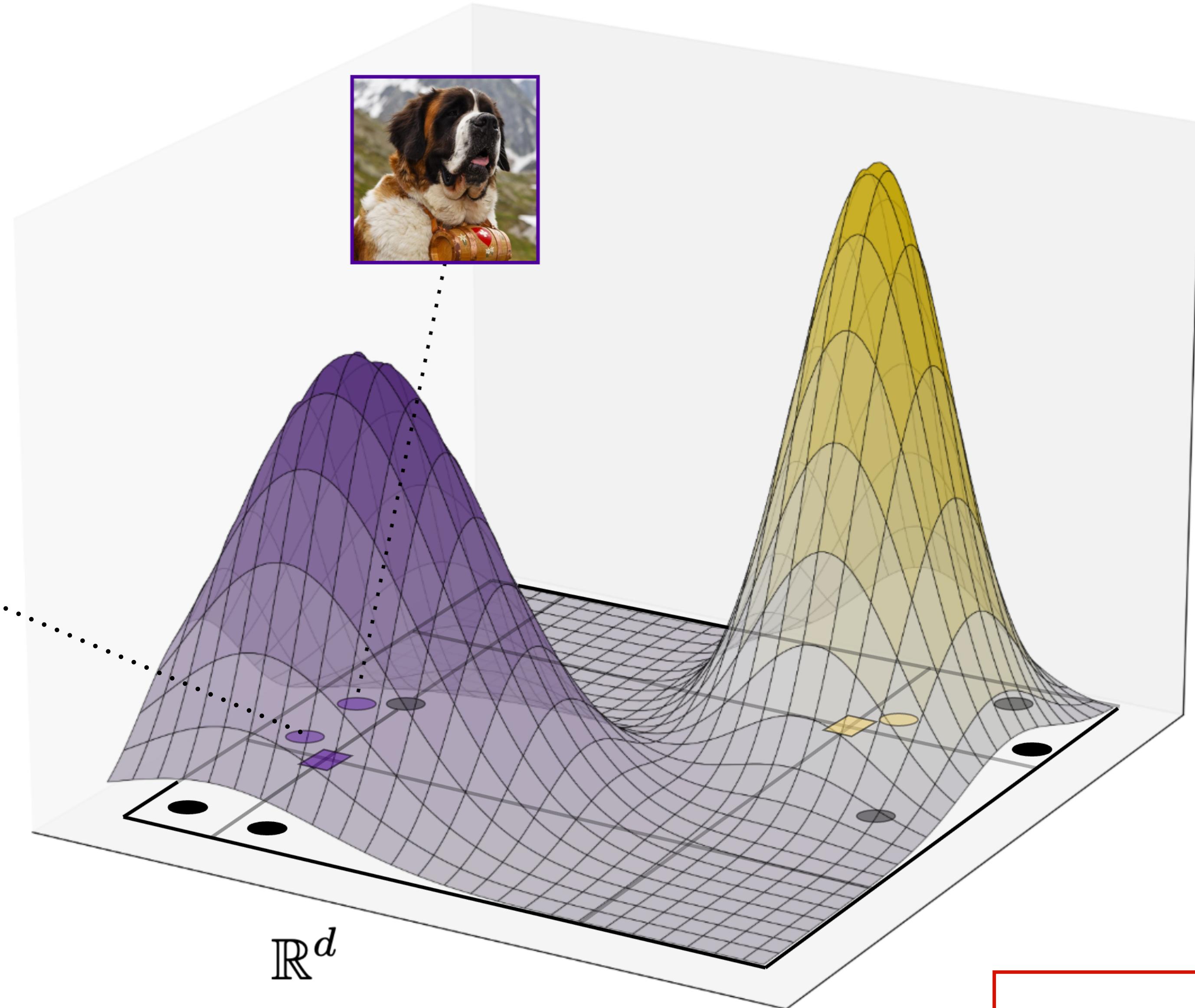


Distribution of Text Embeddings

Prompt Bias

$$\mathbb{E}_{Z \sim P_Z} [(\mathbb{E}_{P_{Y|Z}} [Y|Z] - \mathbb{E}_{\rho_{Y,Z}} [Y|Z])^2]$$

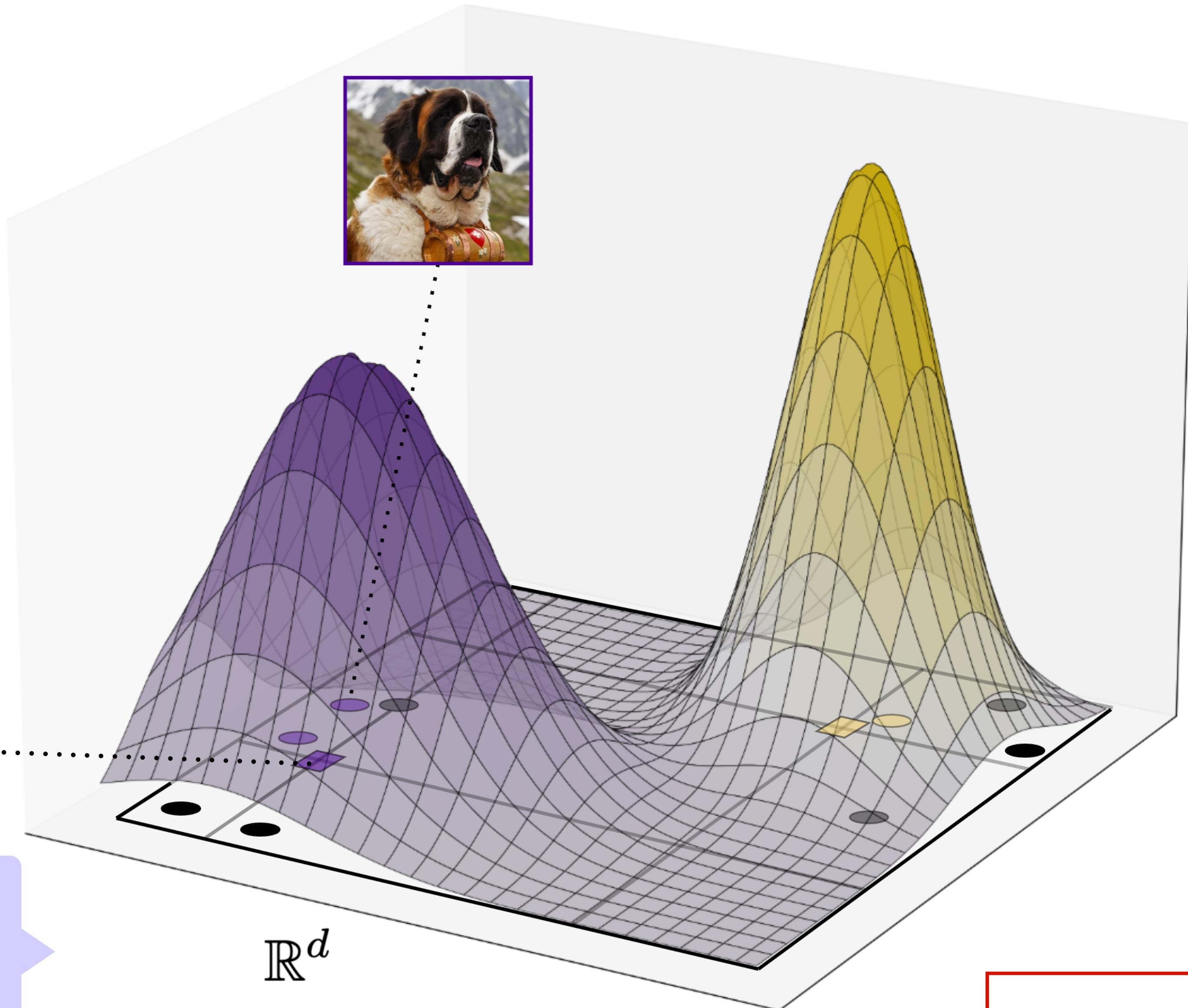
Class-Conditional
"st. bernard rescue near me"



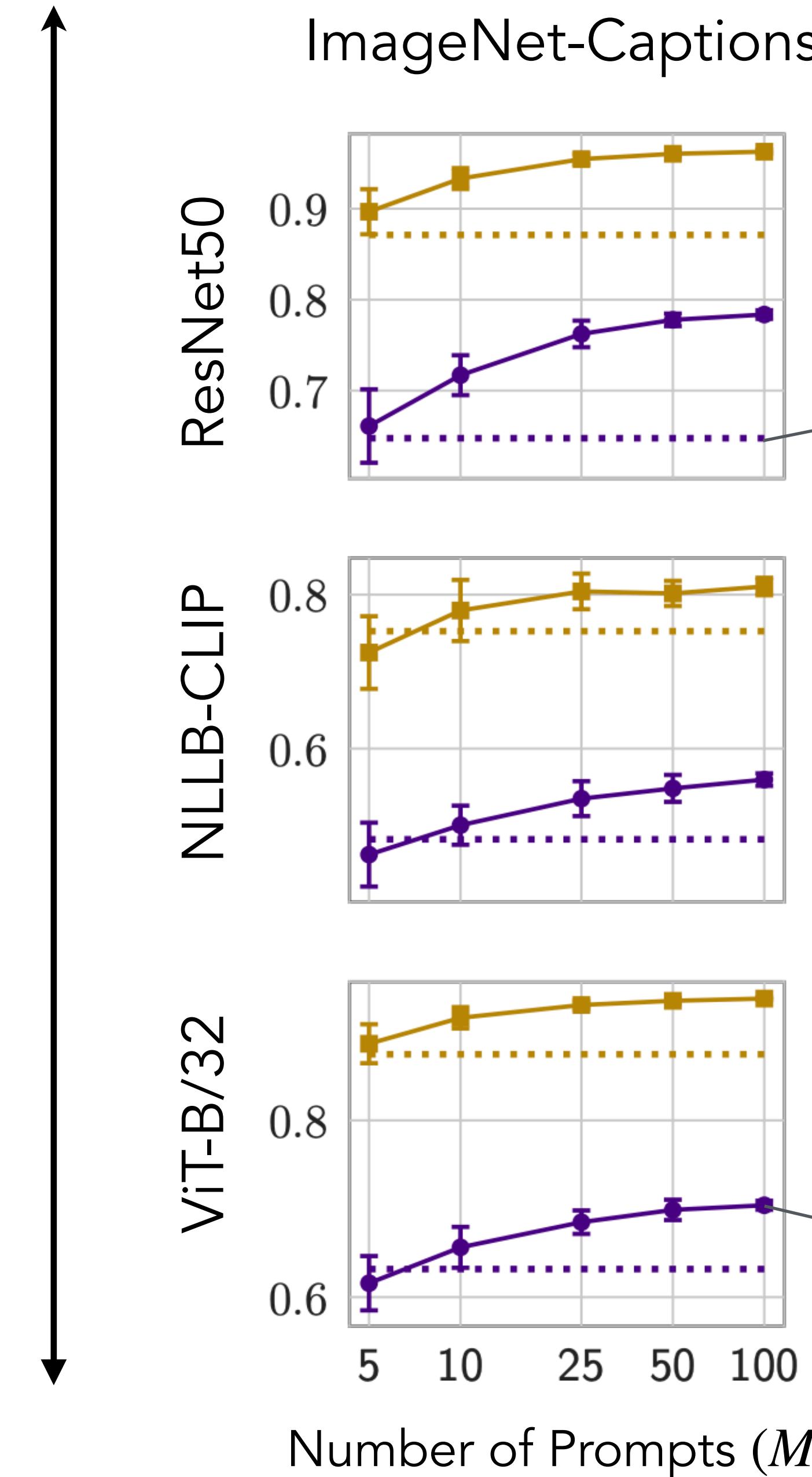
Distribution of Text Embeddings

Prompt Bias

$$\mathbb{E}_{Z \sim P_Z} [(\mathbb{E}_{P_{Y,Z}} [Y|Z] - \mathbb{E}_{\rho_{Y,Z}} [Y|Z])^2]$$



Zero-Shot
Classification
Accuracy
Top-1
Top-5



ImageNet-Captions

"a bad photo of a {c}.",
"a photo of many {c}.",
"a sculpture of a {c}.",
"a photo of the hard to see {c}.",
"a low resolution photo of the {c}.",

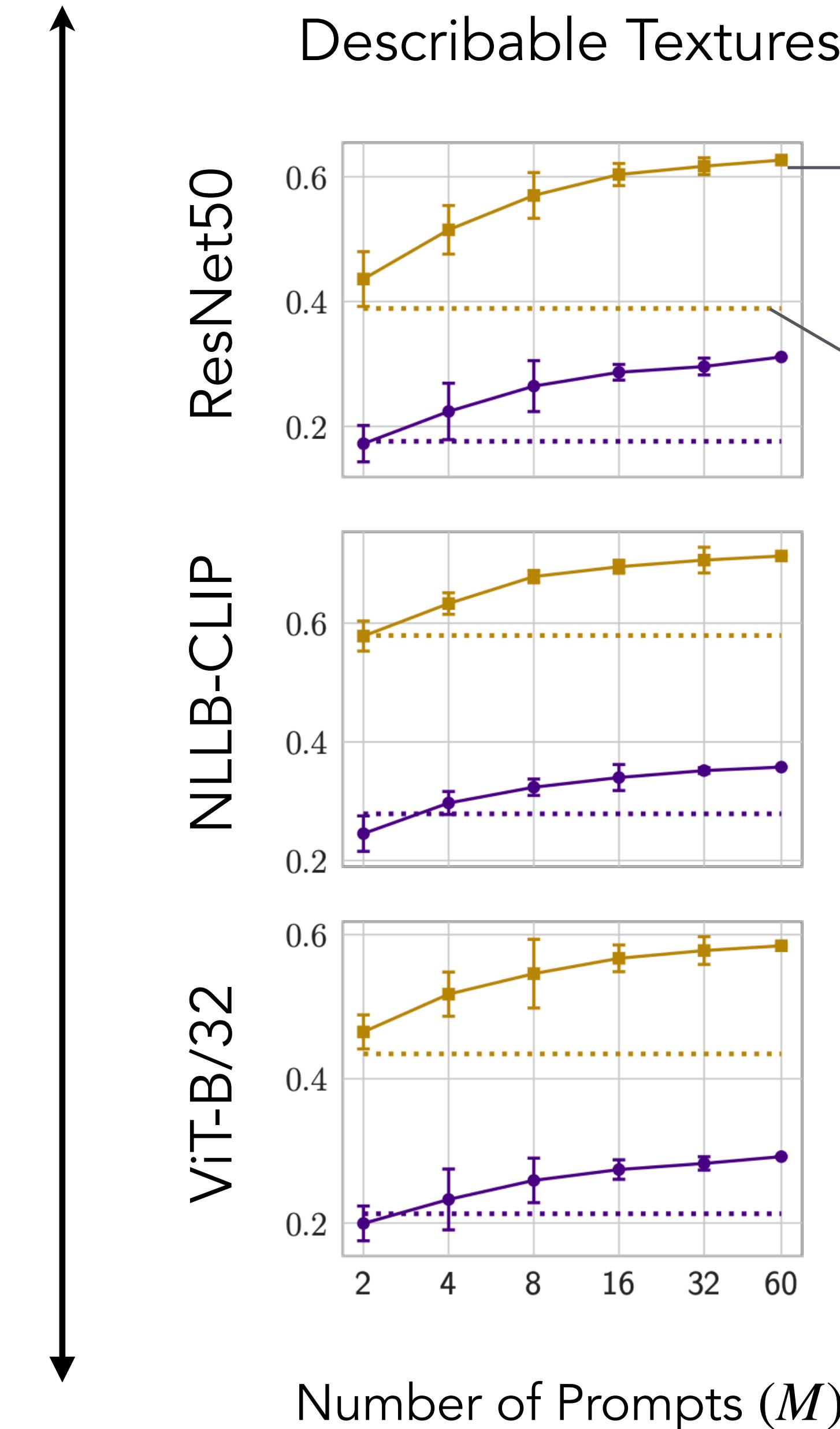
$$\frac{1}{M_y} \sum_{i:y_i=y}$$

Use average embeddings of true captions $P_{Y,Z}$ observed in dataset.

Zero-Shot
Classification
Accuracy

Top-1

Top-5



Describable Textures

gauzy material appears to be a thin and delicate fabric often made of silk or cotton and commonly used in clothing and upholstery.

"a photo of a {texture, pattern, thing, object}"

Multi-View Redundancy

Theorem. Tosh, et al (COLT, 2021)

$$\mathbb{E}[(\mu(X) - \mathbb{E}[Y | X, Z])^2] \leq \varepsilon_X + 2\sqrt{\varepsilon_X \varepsilon_Z} + \varepsilon_Z$$

Similar to our \bar{f} , but no distinction made between pre-training/downstream distributions.

$$\mu(\mathbf{x}) = \mathbb{E} [\mathbb{E} [Y|Z] |X] (\mathbf{x})$$

$$\varepsilon_X := \mathbb{E} [(\mathbb{E}[Y | X] - \mathbb{E}[Y | X, Z])^2] \quad \text{and} \quad \varepsilon_Z := \mathbb{E} [(\mathbb{E}[Y | Z] - \mathbb{E}[Y | X, Z])^2]$$

Both conditional independences satisfied only if $(X, Z) \perp\!\!\!\perp Y$