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 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]

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

Mahmoud Assran[1,2,3*] Quentin Duval[1] Ishan Misra[1] Piotr Bojanowski[1]
Pascal Vincent[1] Michael Rabbat[1,3] Yann LeCun[1,4] Nicolas Ballas[1]

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

GPT-4 Technical Report

GPT-4.

DINOv2: Learning Robust Visual Features without Supervision

laxime 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,
iarma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal[1],
Patrick Labatut*, Armand Joulin*, Piotr Bojanowski*

Meta AI Research [1]*Inria*

*core team **equal contribution

deepseek

DeepSeek-R1: Incentivizing Reasoning Capabilit Reinforcement Learning

DeepSeek-AI

research@deepseek.com

Meta

The Llama 3 Herd of Models

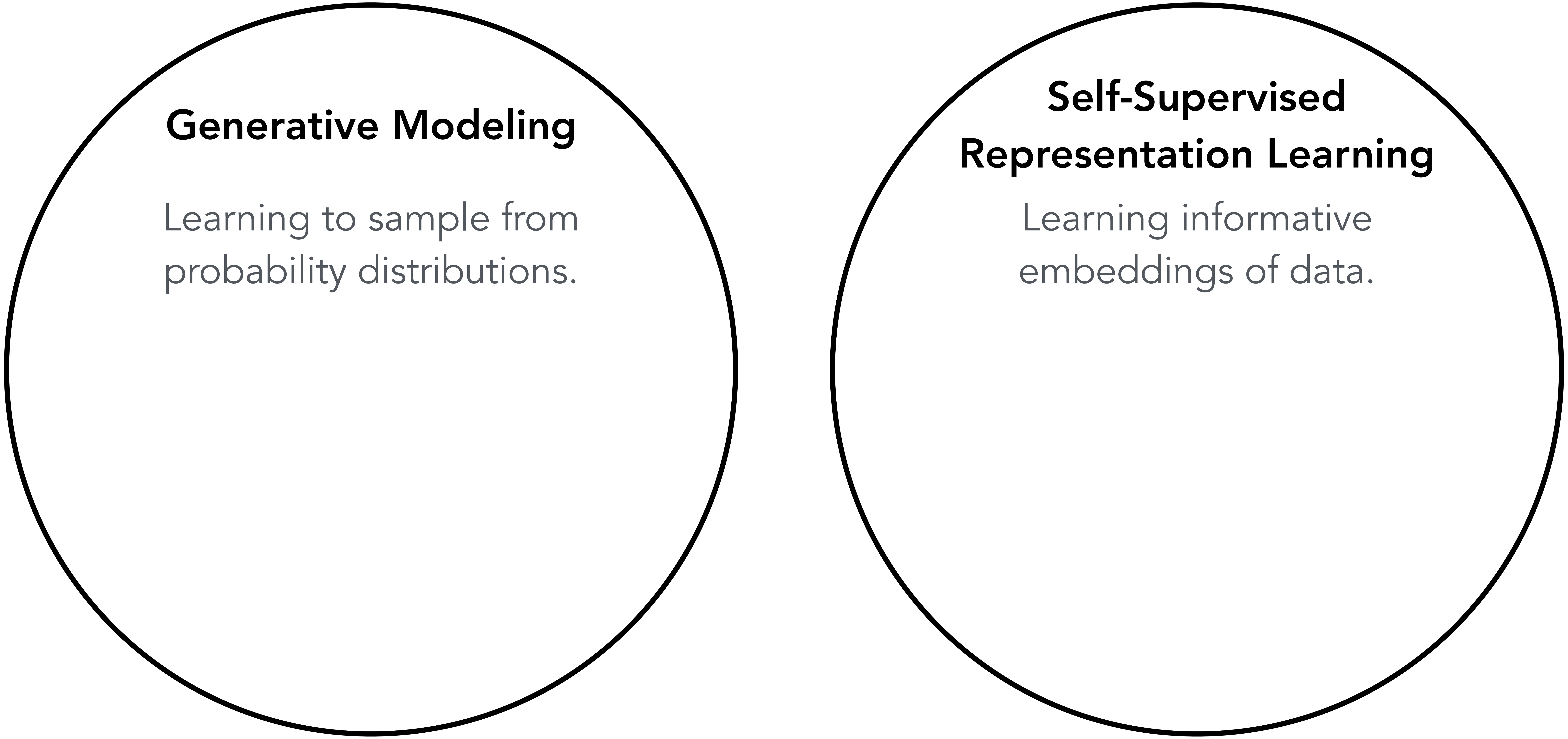
Llama Team, AI @ Meta[1]

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

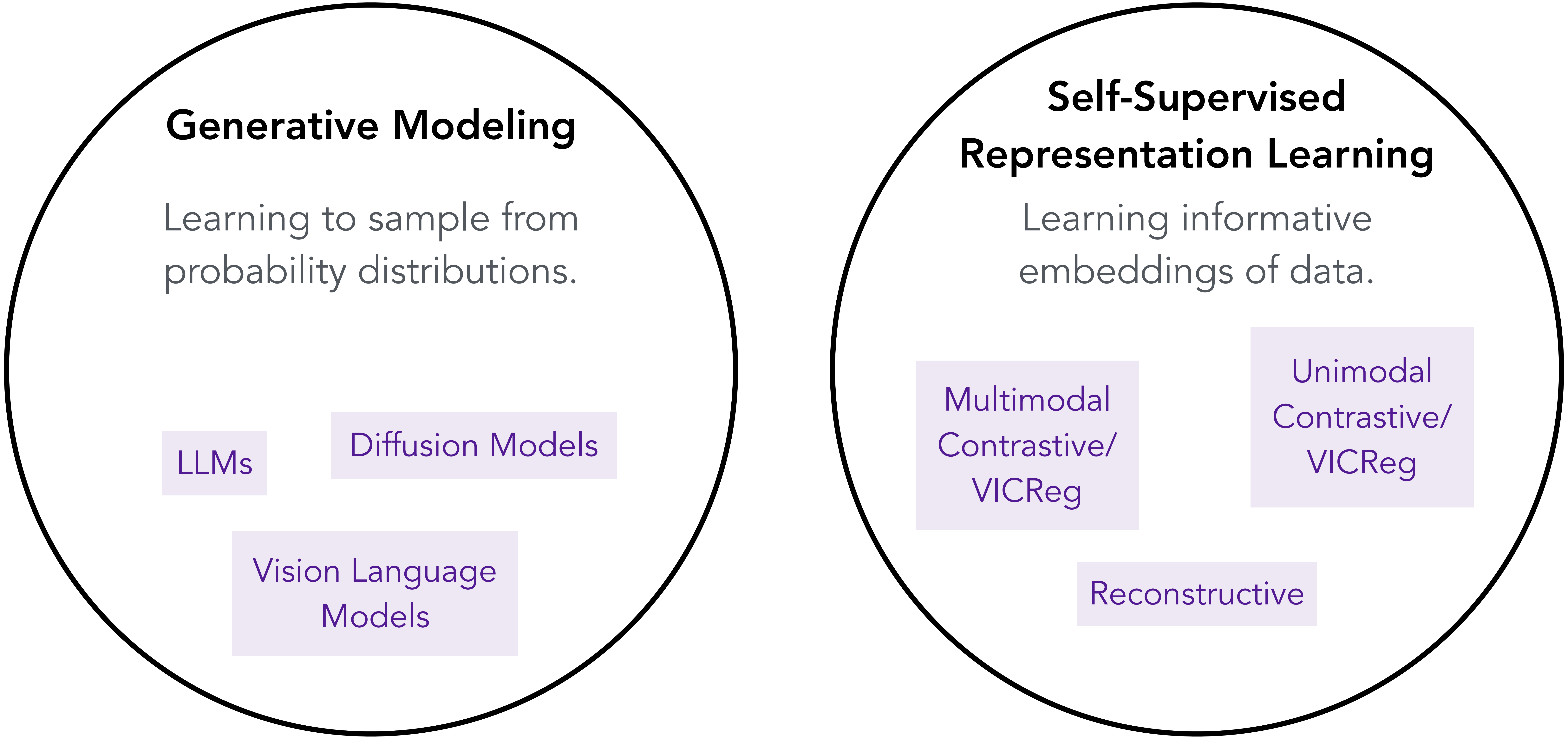
Modern artificial intelligence (AI) systems are powered by foundation models. This paper presents a new set of foundation models, called Llama 3. It is a herd of language models that natively support multilinguality, coding, reasoning, and tool usage. Our largest model is a dense Transformer with 405B 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/

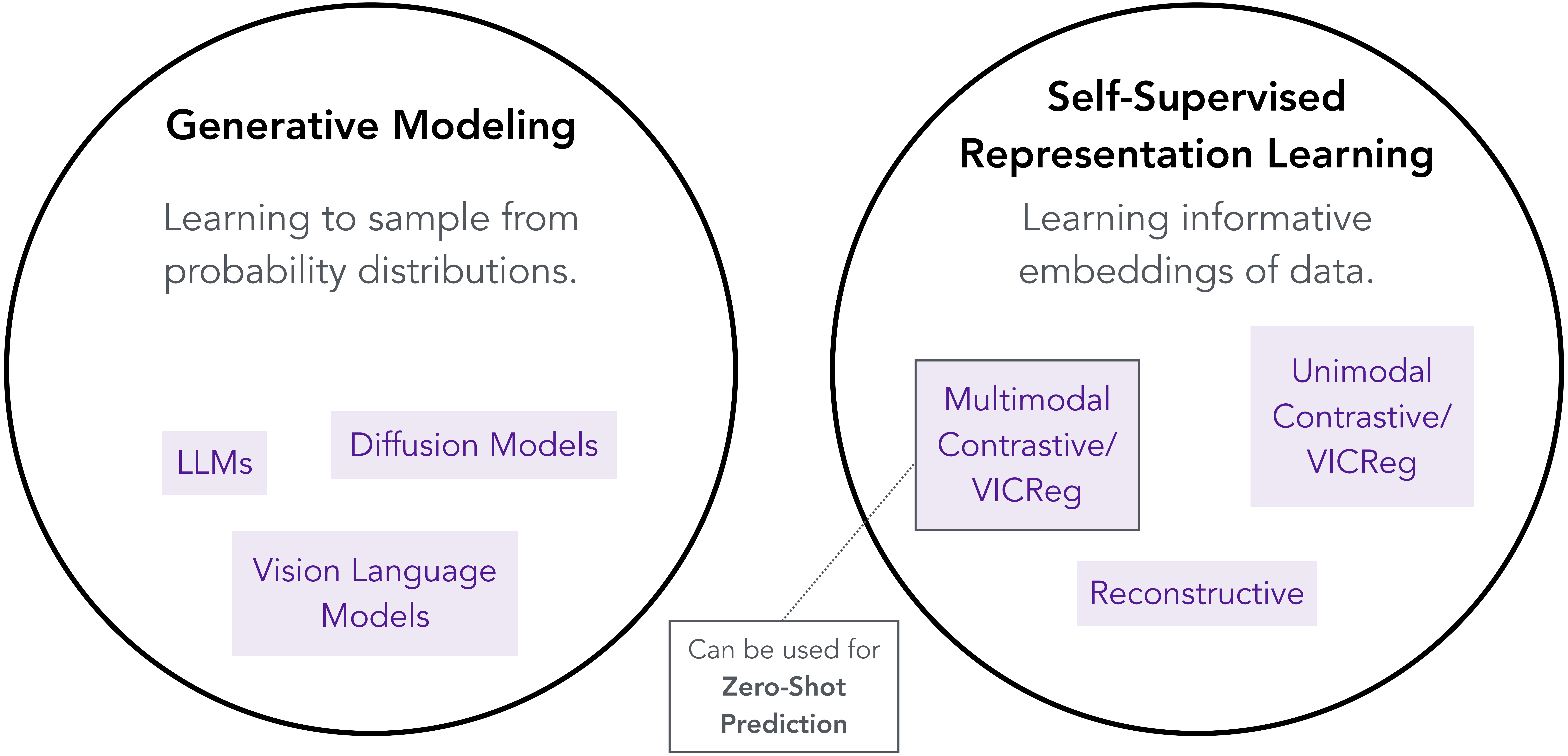
The Mystery of Foundation Models



The Mystery of Foundation Models

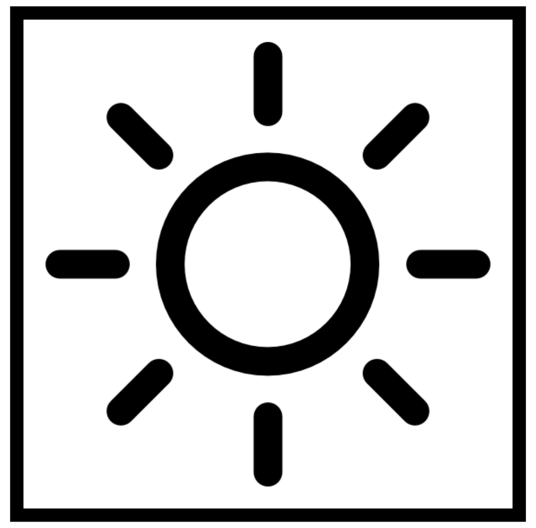


The Mystery of Foundation Models



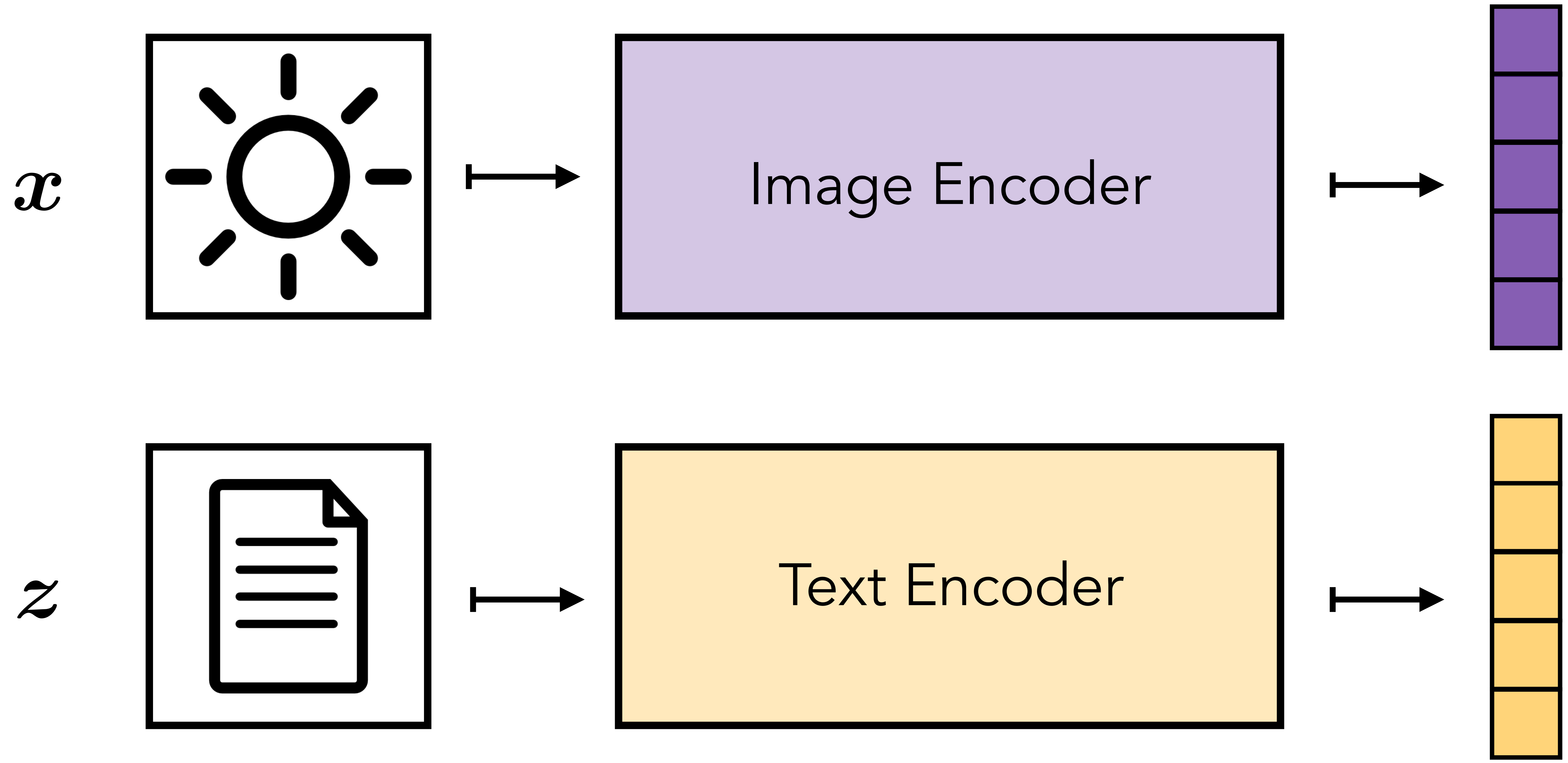
Foundation Models and Zero-Shot Prediction

x

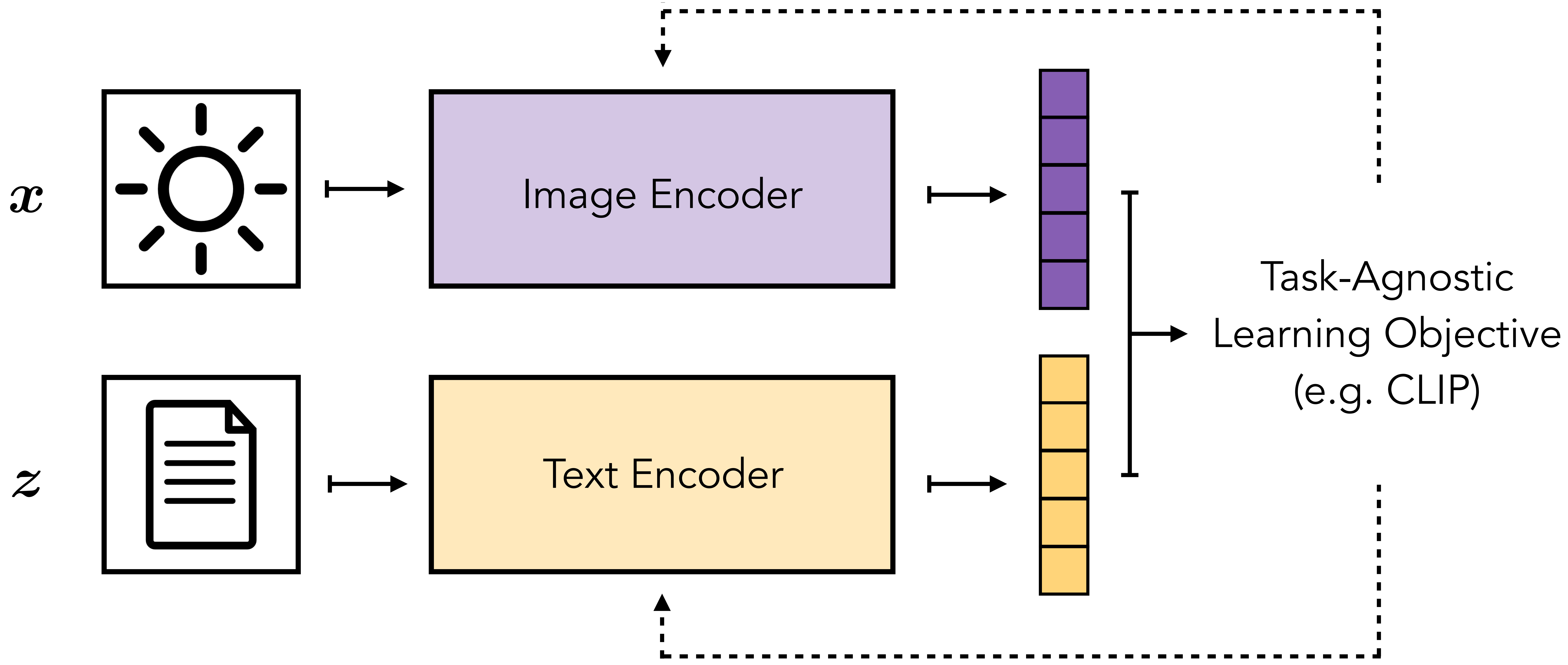


z

Foundation Models and Zero-Shot Prediction

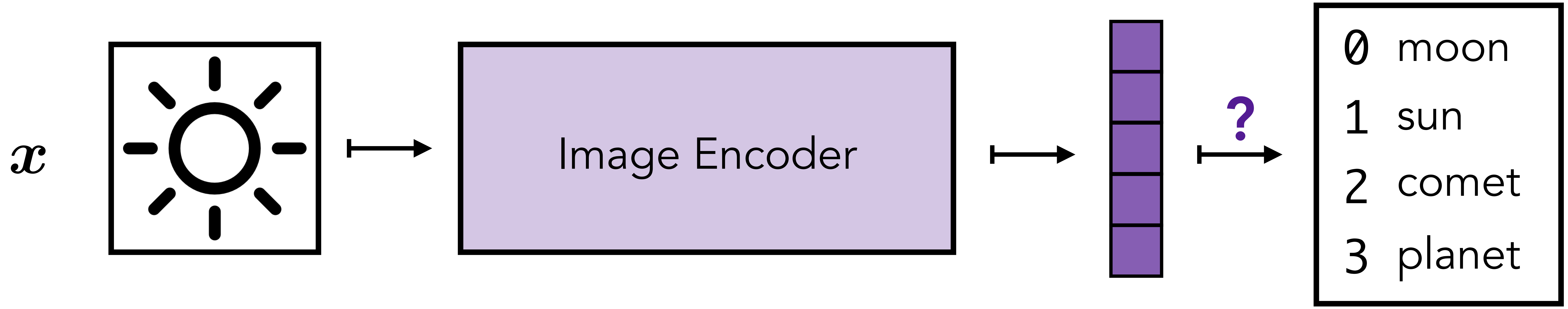


Foundation Models and Zero-Shot Prediction



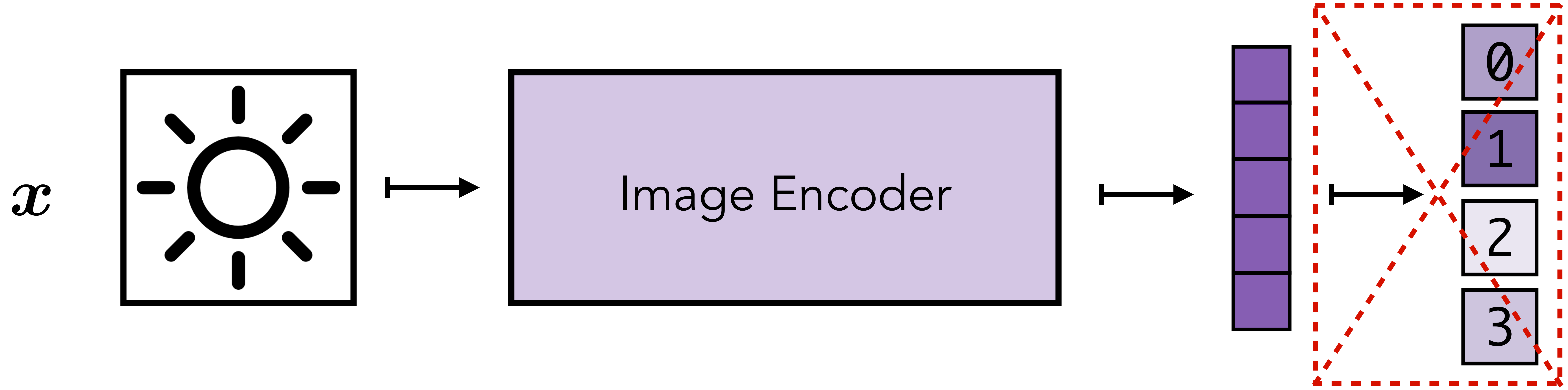
Pre-Training

Foundation Models and Zero-Shot Prediction



Evaluation

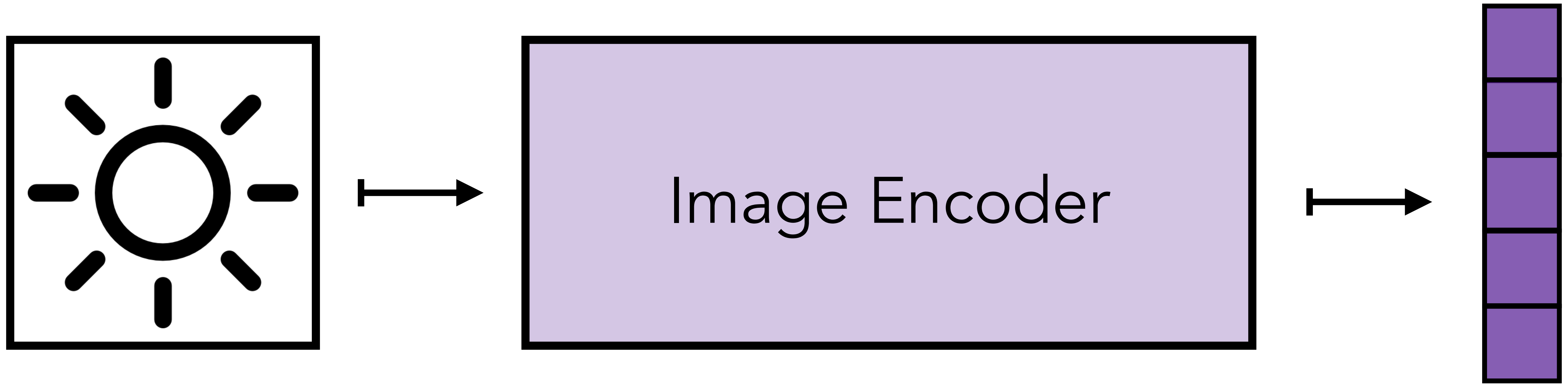
Foundation Models and Zero-Shot Prediction



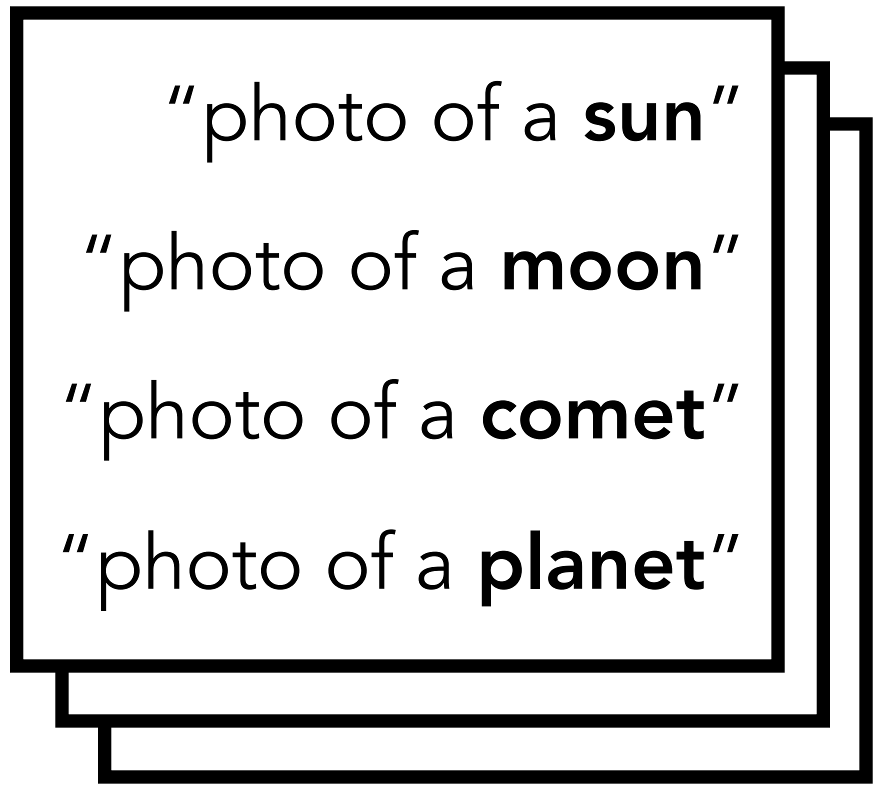
Evaluation

Foundation Models and Zero-Shot Prediction

$\boldsymbol{x}$



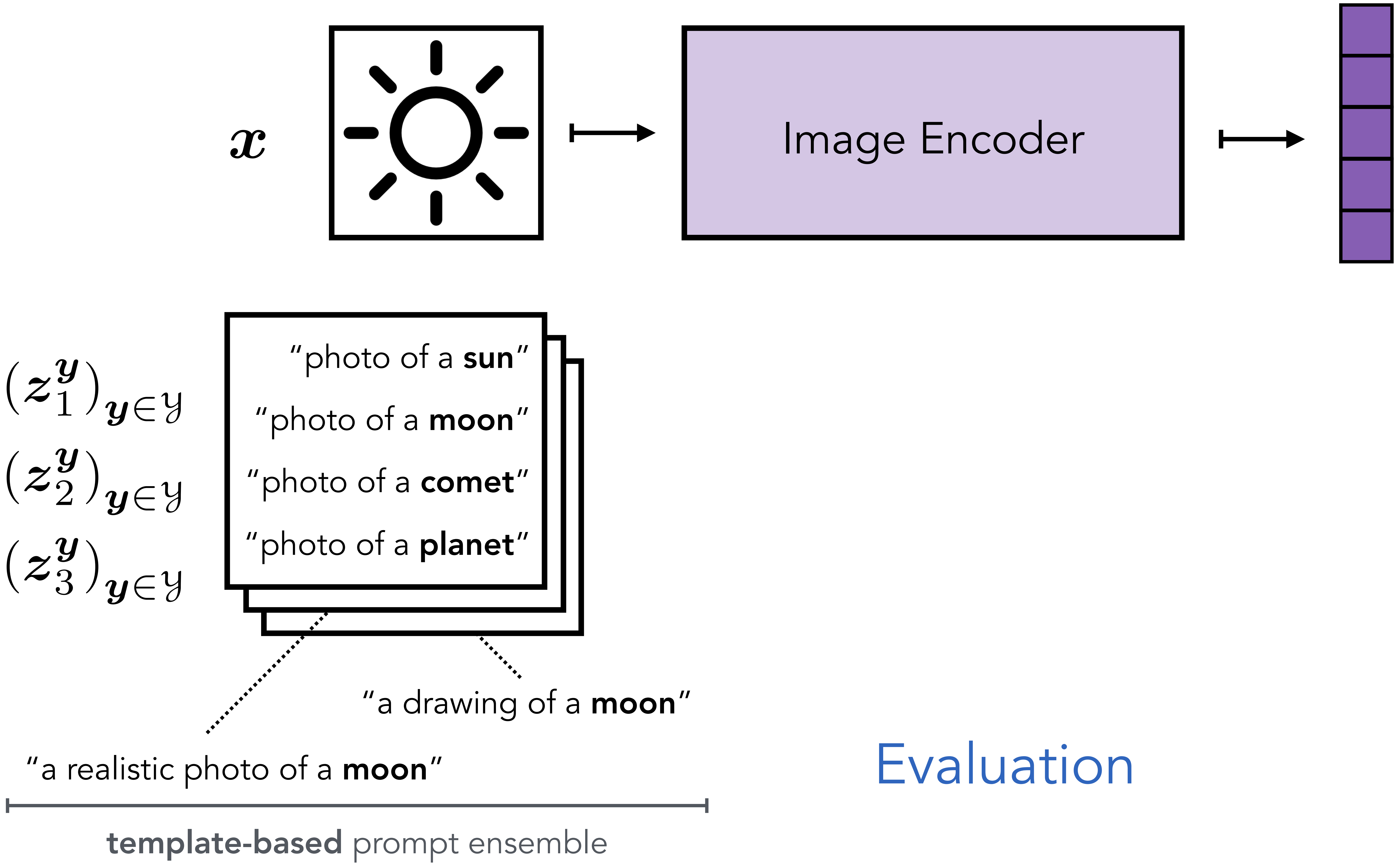
$(\boldsymbol{z}_1^{\boldsymbol{y}})_{\boldsymbol{y}\in\mathcal{Y}}$
$(\boldsymbol{z}_2^{\boldsymbol{y}})_{\boldsymbol{y}\in\mathcal{Y}}$
$(\boldsymbol{z}_3^{\boldsymbol{y}})_{\boldsymbol{y}\in\mathcal{Y}}$



Idea: Convert labels into **prompts** (pseudo-captions)

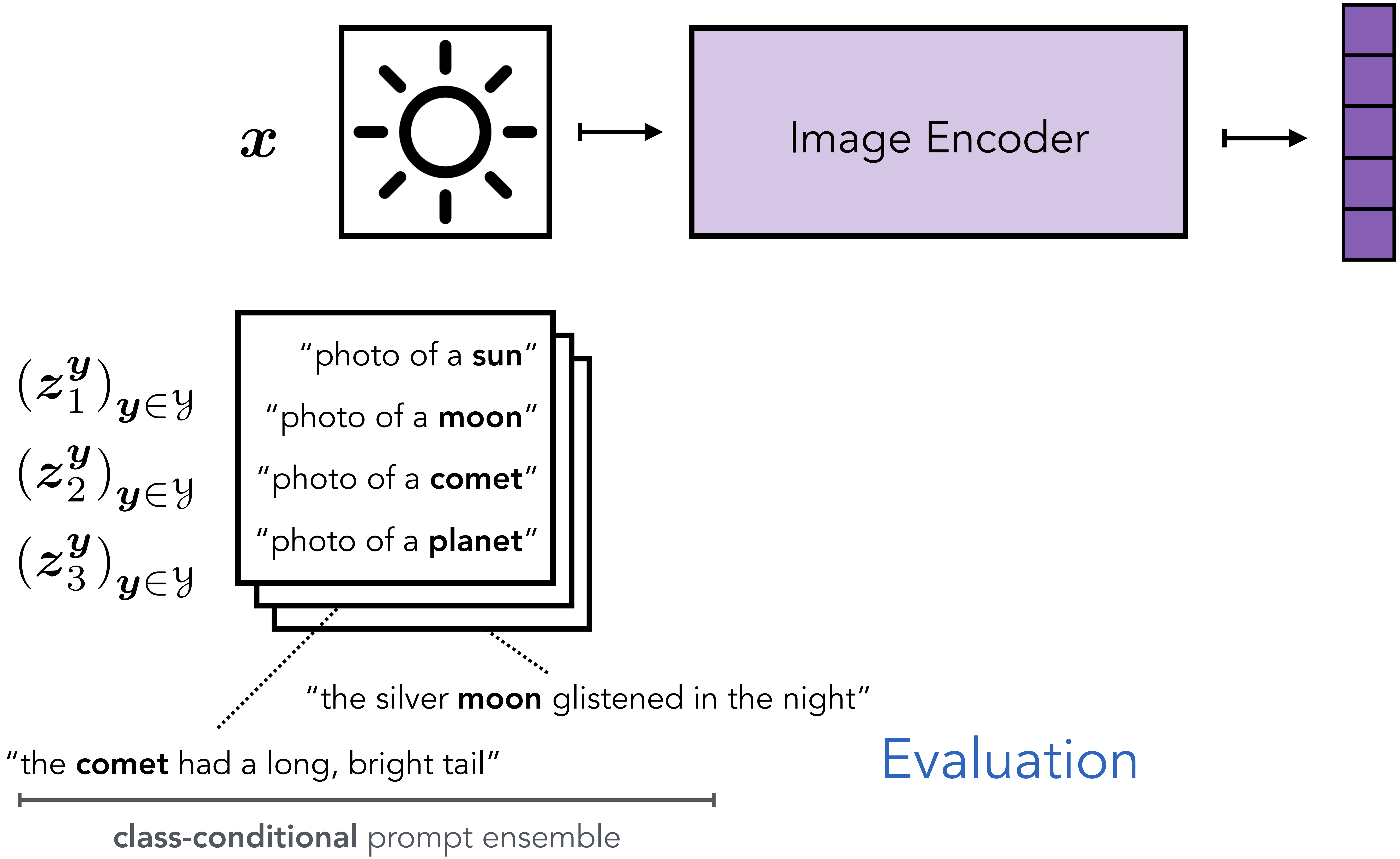
Evaluation

Foundation Models and Zero-Shot Prediction



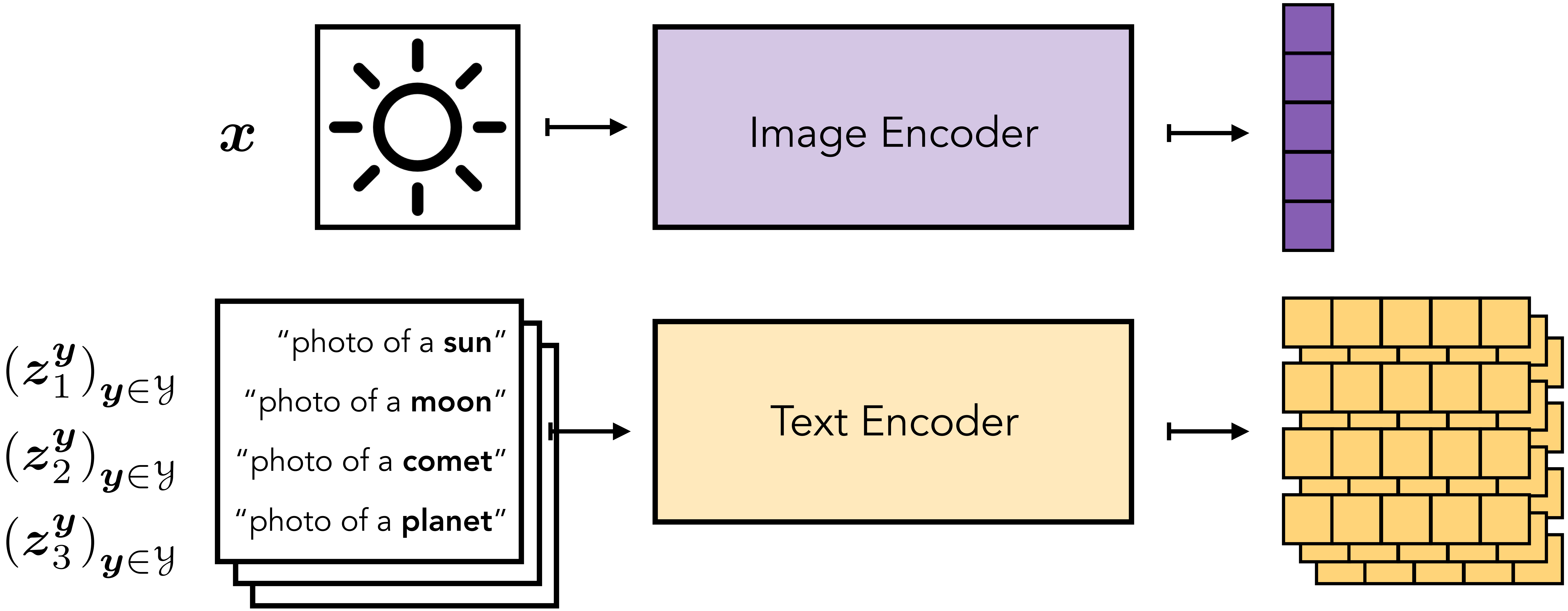
Evaluation

Foundation Models and Zero-Shot Prediction



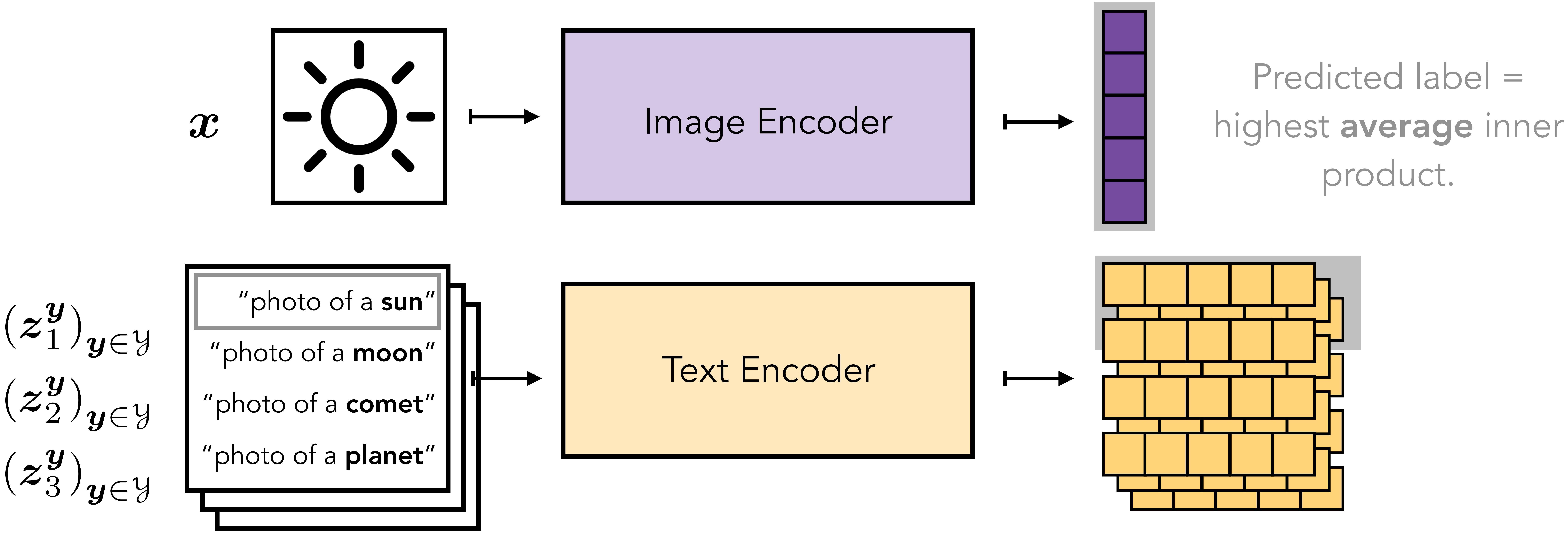
Evaluation

Foundation Models and Zero-Shot Prediction



Evaluation

Foundation Models and Zero-Shot Prediction



Evaluation

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 Covert[1] Rosanne Liu[2, 3] Ali Farhadi[1]
[1]University of Washington [2]Google DeepMind [3]ML 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 Covert[1] Rosanne Liu[2, 3] Ali Farhadi[1]
[1]University of Washington [2]Google DeepMind [3]ML Collective

Generalization Analysis (Supervised Learning)

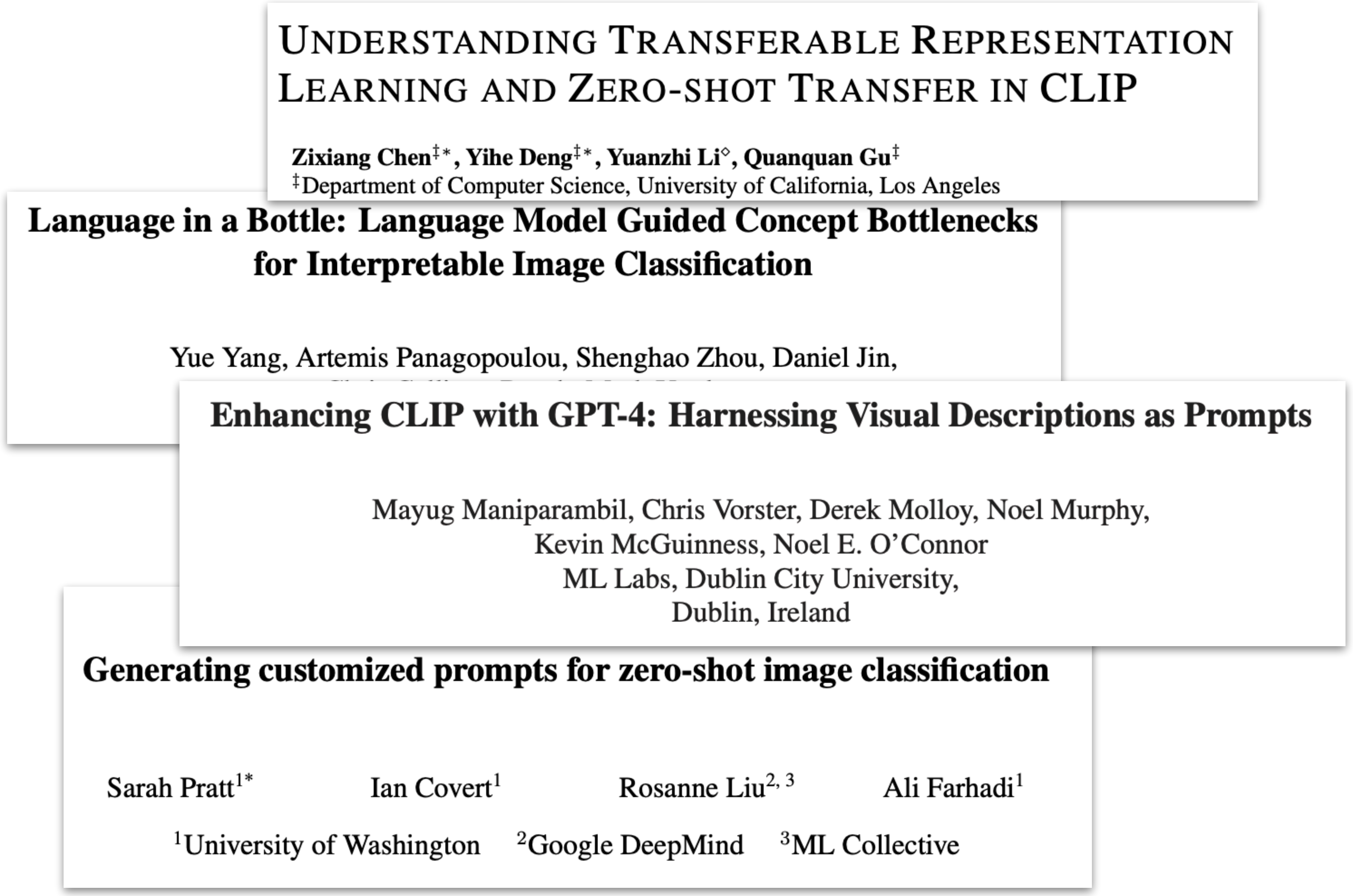
target function

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

predictor trained on
$(X_1, Y_1), \ldots, (X_N, Y_N) \sim P_{X,Y}$

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 Covert[1] Rosanne Liu[2, 3] Ali Farhadi[1]
[1]University of Washington [2]Google DeepMind [3]ML Collective

Generalization Analysis (Supervised Learning)

target function

$$\mathbb{E}_{X \sim P_X}\left[\left(\hat{f}(X) - f_\star(X)\right)^2\right] \leq O\left(\frac{1}{N^{\square}}\right)$$

predictor trained on $(X_1, Y_1), \ldots, (X_N, Y_N) \sim P_{X,Y}$

convergence rate w.r.t. number of training examples

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 Covert[1] Rosanne Liu[2, 3] Ali Farhadi[1]
[1]University of Washington [2]Google DeepMind [3]ML Collective

Generalization Analysis (Zero-Shot Prediction)

target function

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

predictor based on N pre-training examples and M prompts

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 Covert[1] Rosanne Liu[2, 3] Ali Farhadi[1]
[1]University of Washington [2]Google DeepMind [3]ML Collective

Generalization Analysis (Zero-Shot Prediction)

target function

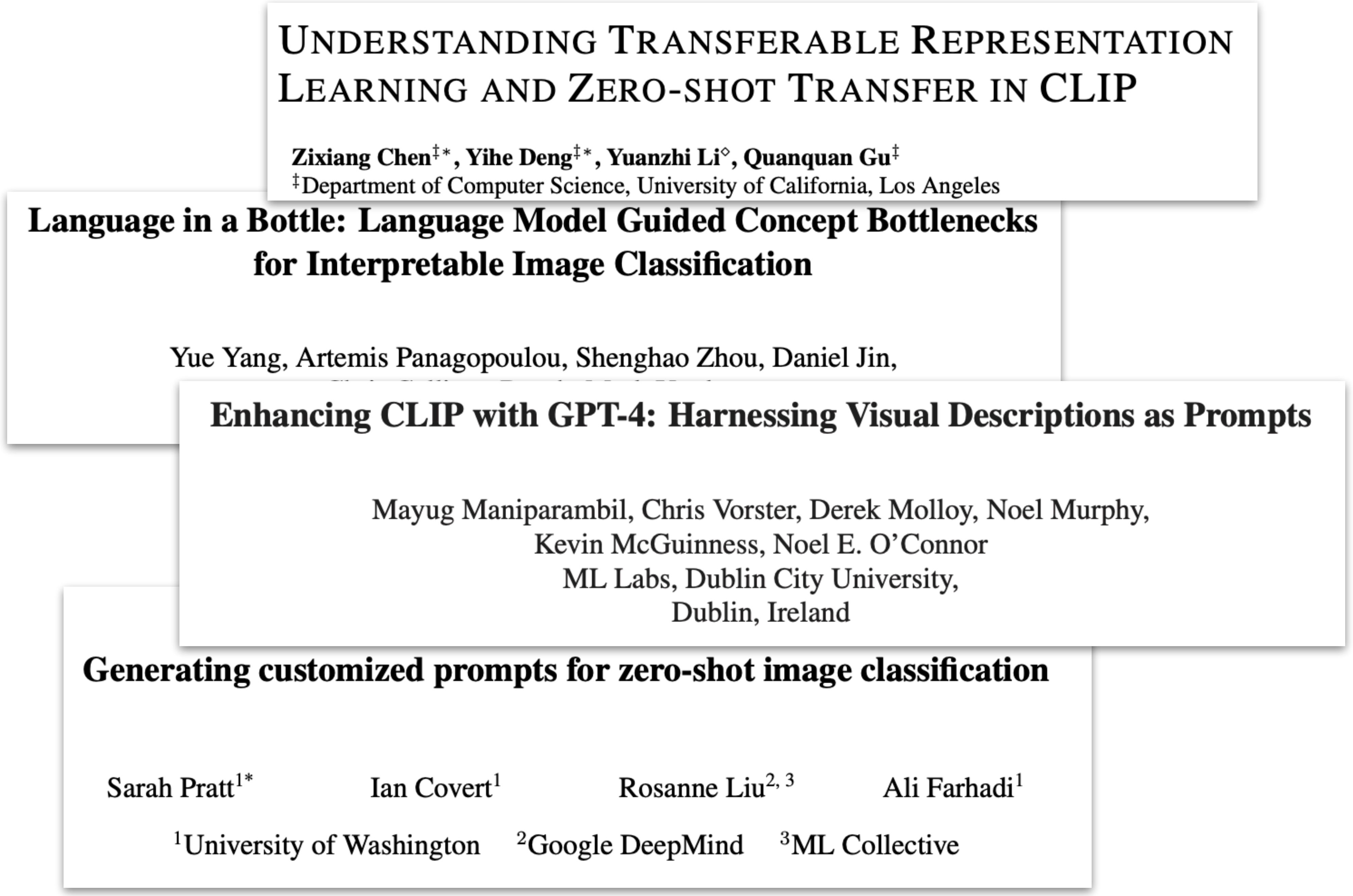
$$\mathbb{E}_{X\sim P_X}\left[\left(\hat{f}(X) - f_\star(X)\right)^2\right] \lesssim \frac{1}{N^{\square}} + \frac{1}{M^{\square}} + ?$$

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 Covert[1] Rosanne Liu[2, 3] Ali Farhadi[1]
[1]University of Washington [2]Google DeepMind [3]ML 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.

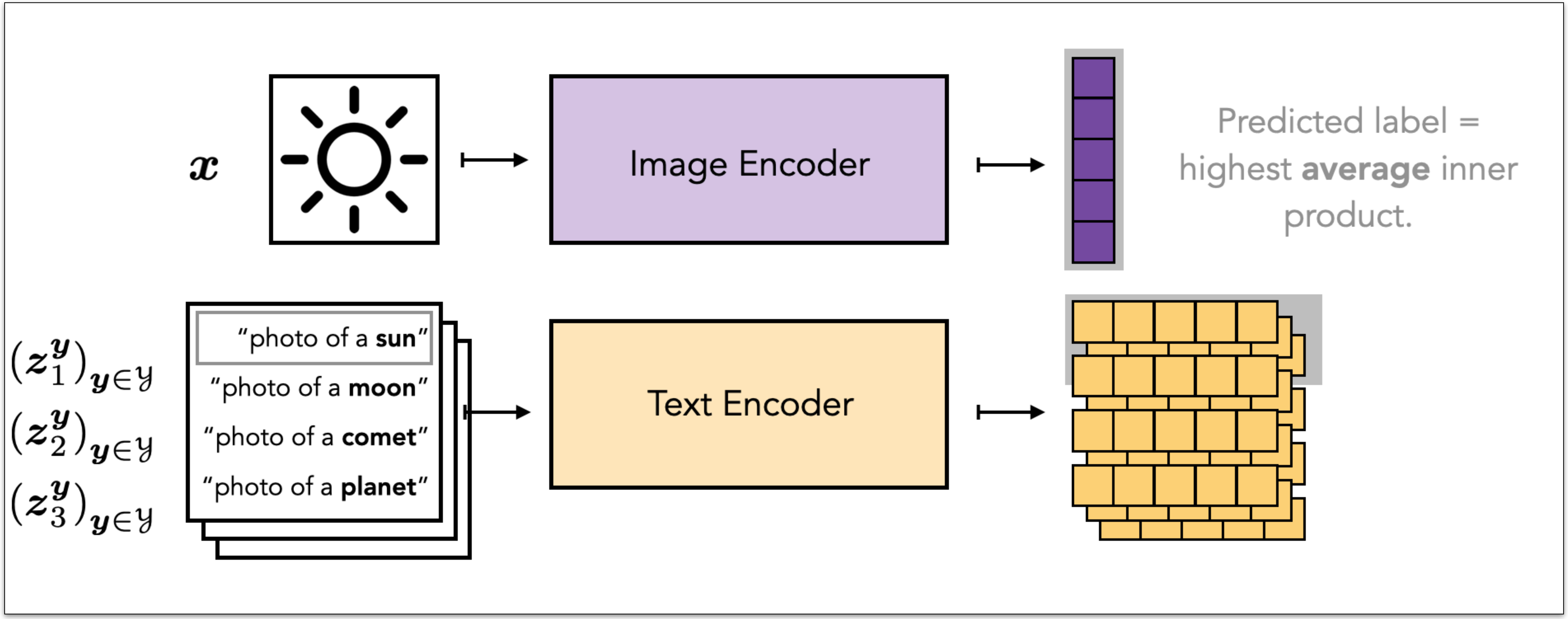
$$\mathbb{E}_{X \sim P_X}\left[(f_\star(X) - \hat{f}(X))^2\right] \leq$$

direct predictor ZSP procedure

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

direct predictor

ZSP procedure



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

direct predictor

ZSP procedure

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

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

direct predictor

ZSP procedure

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

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

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.

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

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

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

information-theoretic error

learning error

direct predictor

ZSP procedure

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

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

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

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

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

direct predictor ZSP procedure

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

Roadmap of Theoretical Analysis

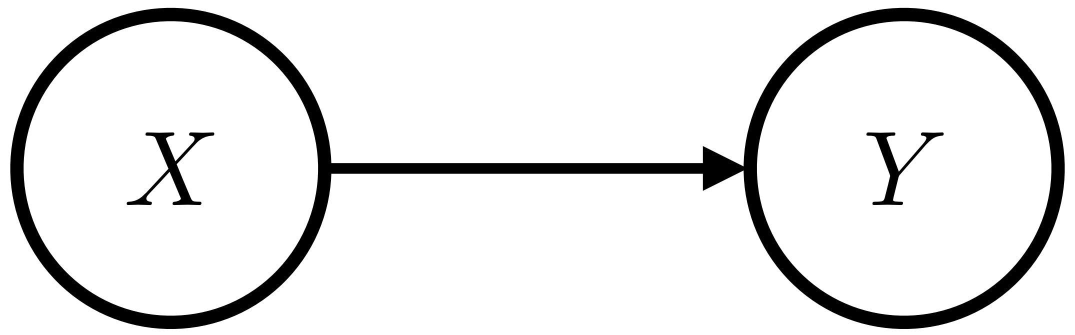
1. Define $\bar{f}$ in terms of pre-training, evaluation, and prompting distribution.
2. Upper bound information-theoretic error using dependence relationships between images, captions, and labels.
3. Define class of estimators $\hat{f}$, and bound learning error using tools from statistical learning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

$X =$ image
$Y =$ label
$Z =$ caption

$P_{X,Y}$

Evaluation

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



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

direct predictor

ZSP procedure

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

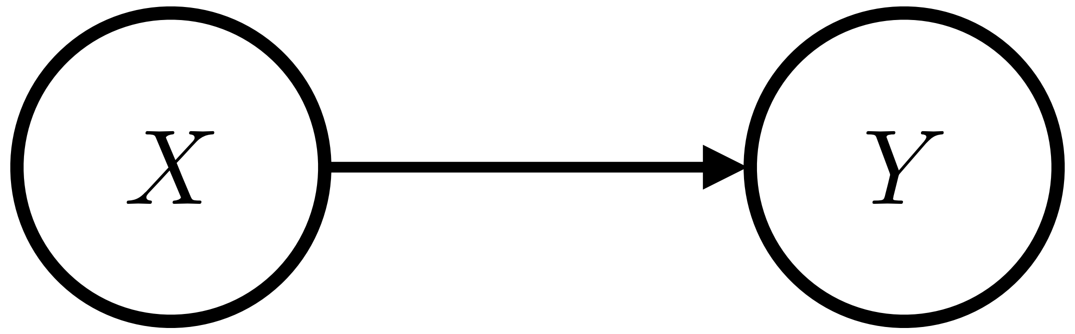
Roadmap of Theoretical Analysis

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

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



$$\mathbb{E}_{X\sim P_X}\left[(f_\star(X) - \hat{f}(X))^2\right] \le \underbrace{2\mathbb{E}_{X\sim P_X}\left[(f_\star(X)-\bar{f}(X))^2\right]}_{\text{information-theoretic error}} + \underbrace{2\mathbb{E}_{X\sim P_X}\left[(\bar{f}(X)-\hat{f}(X))^2\right]}_{\text{learning error}}$$

direct predictor ZSP procedure

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

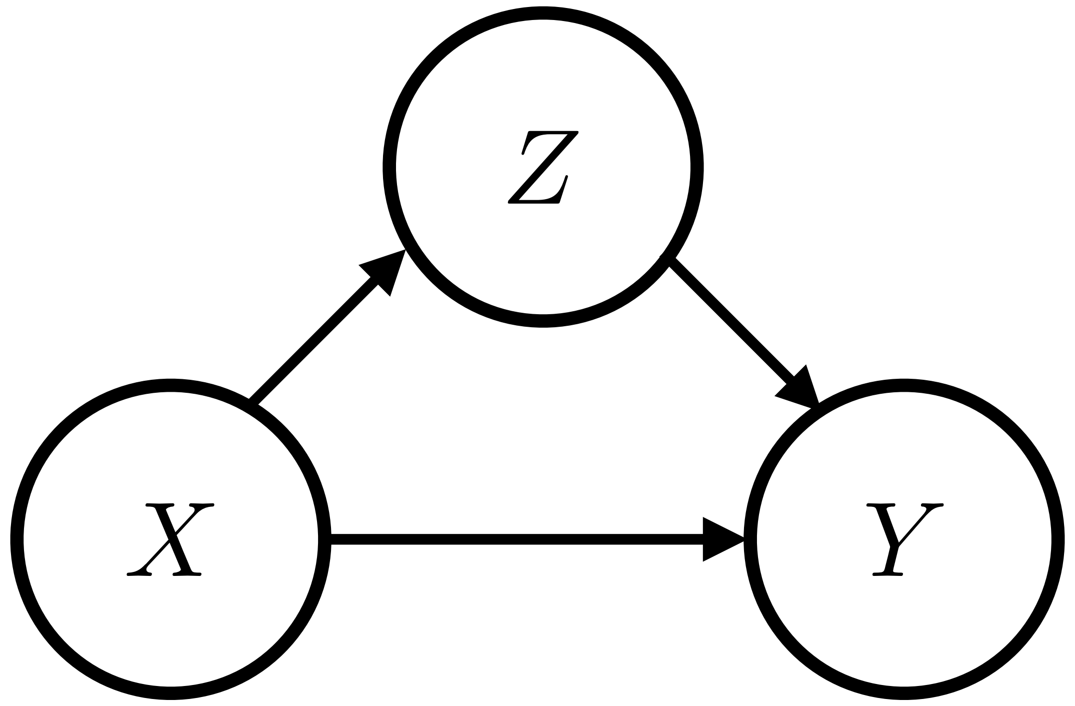
Roadmap of Theoretical Analysis

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

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

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



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

information-theoretic error

learning error

direct predictor

ZSP procedure

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

Roadmap of Theoretical Analysis

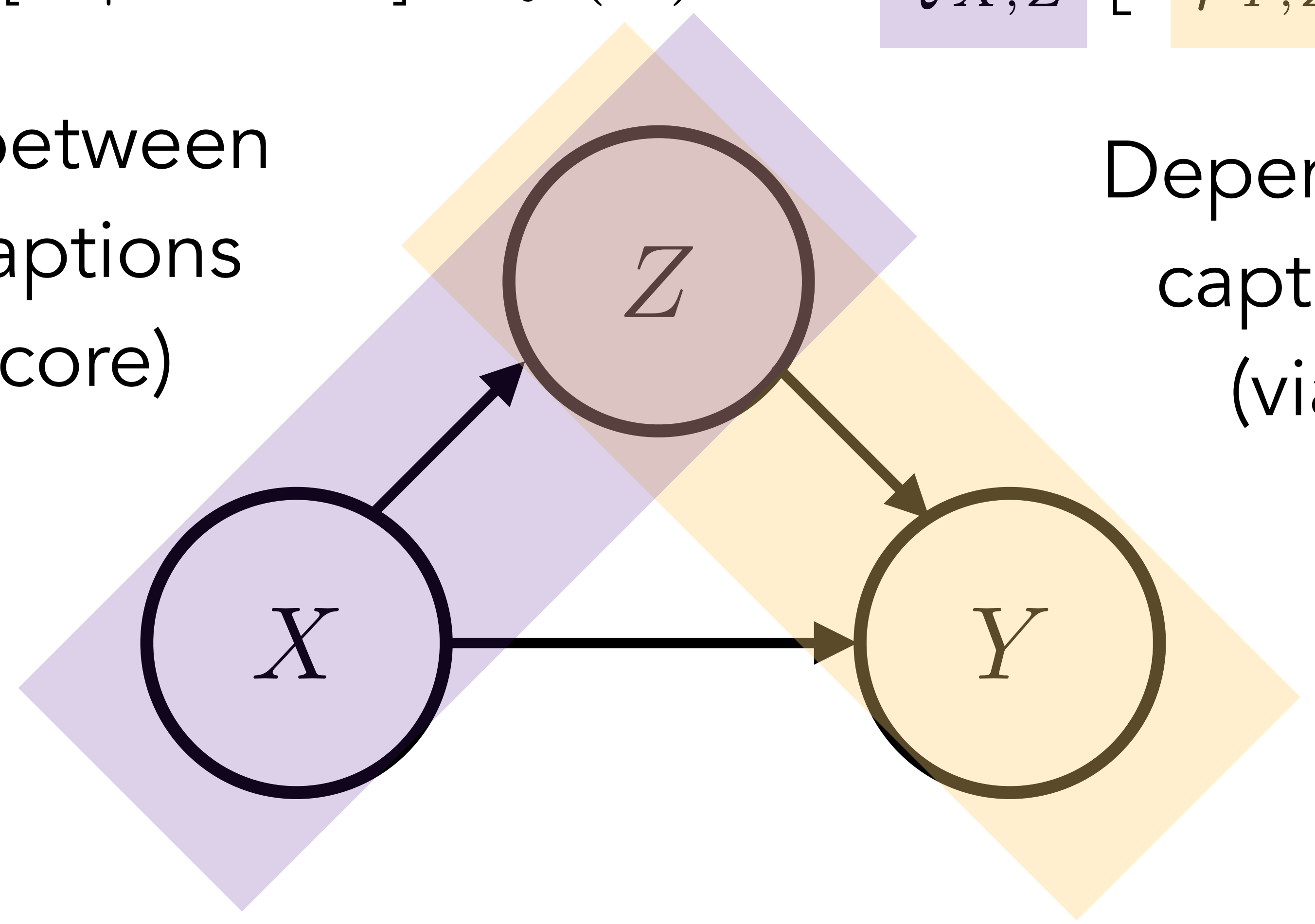
1. Define $\bar{f}$ in terms of pre-training, evaluation, and prompting distribution.
2. Upper bound information-theoretic error using dependence relationships between images, captions, and labels.
3. Define class of estimators $\hat{f}$, and bound learning error using tools from statistical learning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

X = image
Y = label
Z = caption

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

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

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



Dependence between captions and labels (via prompting)

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

direct predictor ZSP procedure

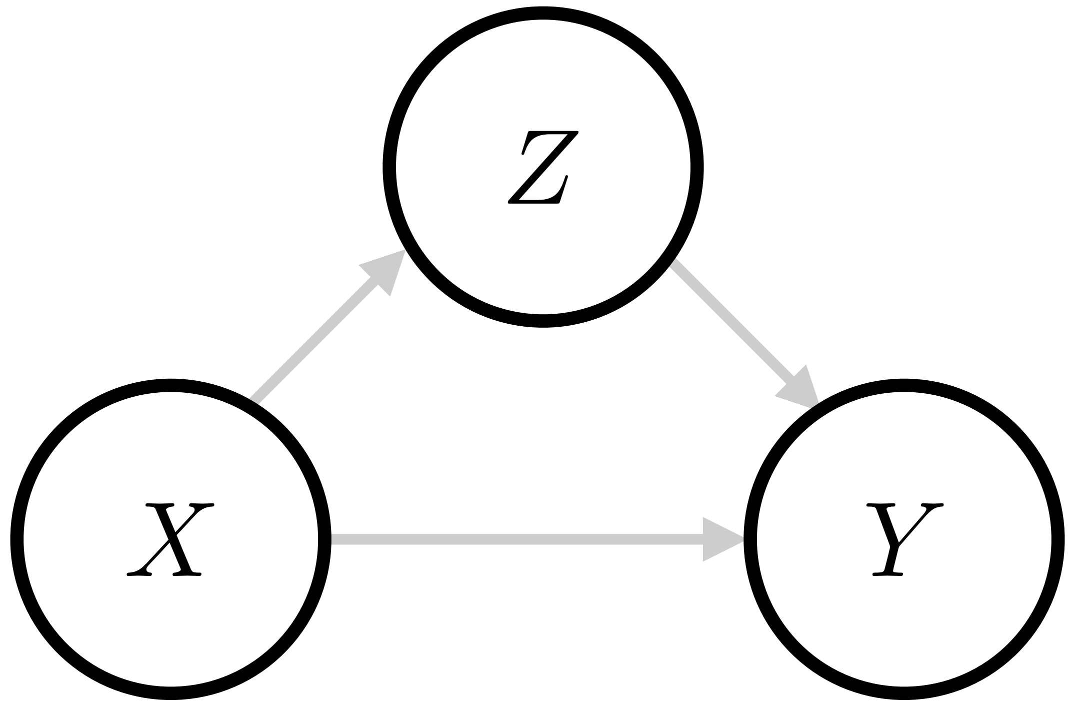
population version of ZSP
(based on distributions instead of samples)

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

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



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

direct predictor ZSP procedure

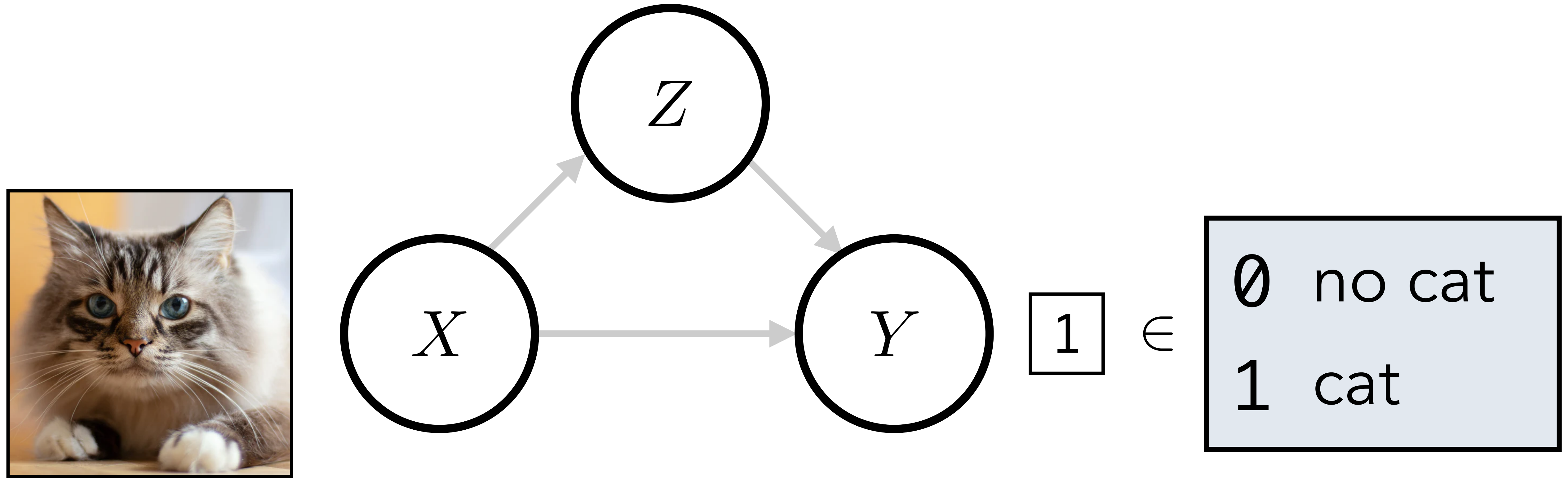
population version of ZSP
(based on distributions instead of samples)

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

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



$$\mathbb{E}_{X\sim P_X}\left[(f_\star(X)-\hat{f}(X))^2\right] \le \underbrace{2\mathbb{E}_{X\sim P_X}\left[(f_\star(X)-\bar{f}(X))^2\right]}_{\text{information-theoretic error}} + \underbrace{2\mathbb{E}_{X\sim P_X}\left[(\bar{f}(X)-\hat{f}(X))^2\right]}_{\text{learning error}}$$

direct predictor

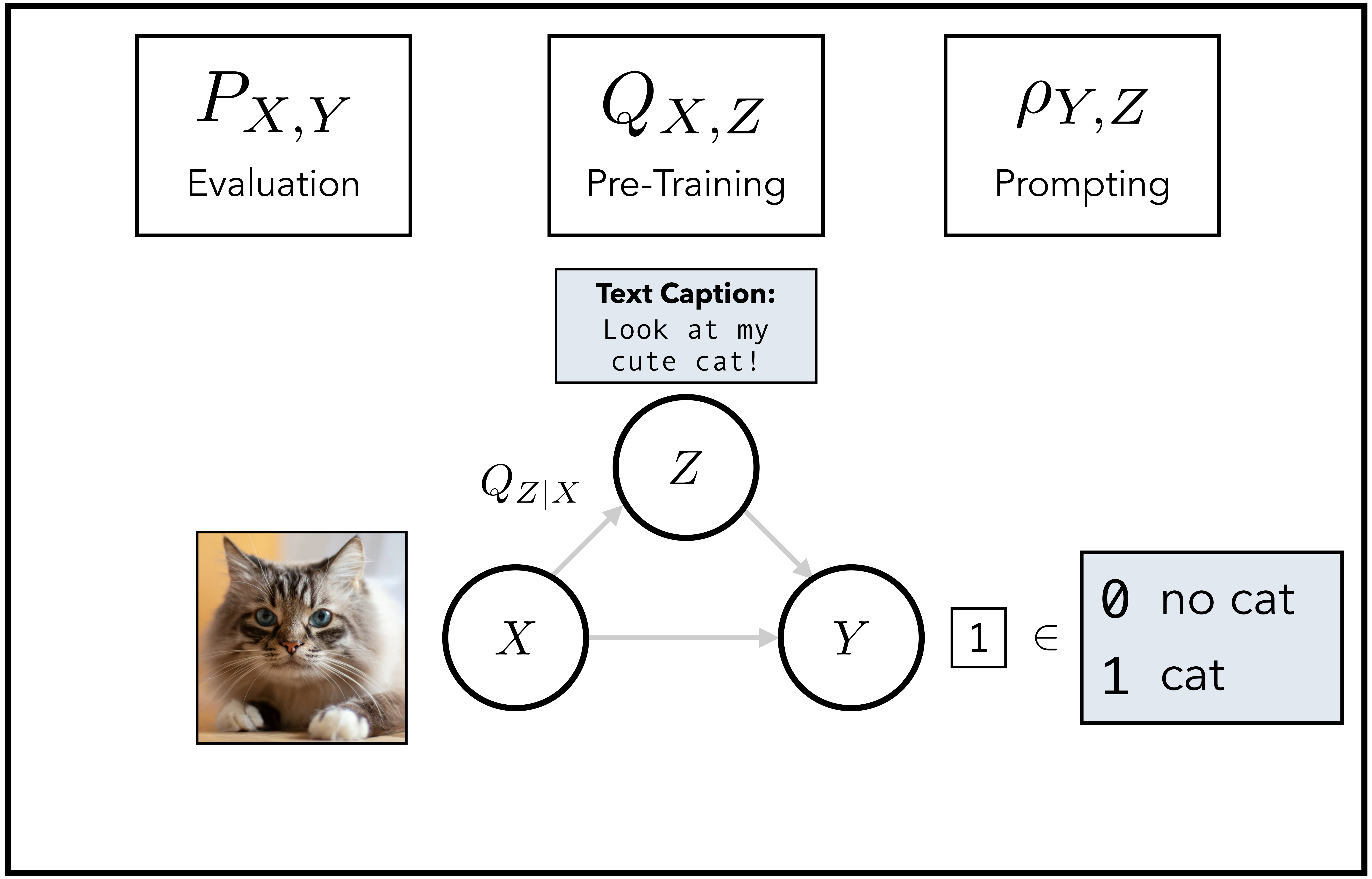
ZSP procedure

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



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

direct predictor ZSP procedure

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

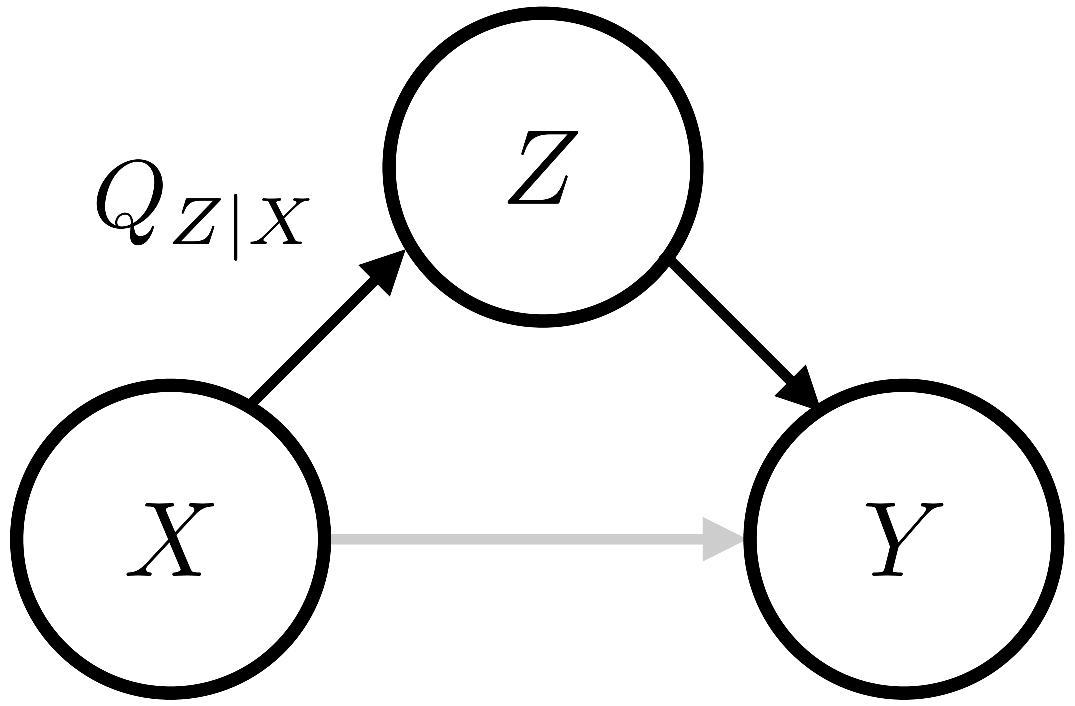
Roadmap of Theoretical Analysis

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

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

Text Caption:
`Look at my cute cat!`



$\boxed{1} \in$

0	no cat
1	cat

ZSP (Indirect) performs well 🙂

$$\mathbb{E}_{X\sim P_X}\left[(f_\star(X) - \hat{f}(X))^2\right] \le \underbrace{2\mathbb{E}_{X\sim P_X}\left[(f_\star(X) - \bar{f}(X))^2\right]}_{\text{information-theoretic error}} + \underbrace{2\mathbb{E}_{X\sim P_X}\left[(\bar{f}(X) - \hat{f}(X))^2\right]}_{\text{learning error}}$$

direct predictor

ZSP procedure

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

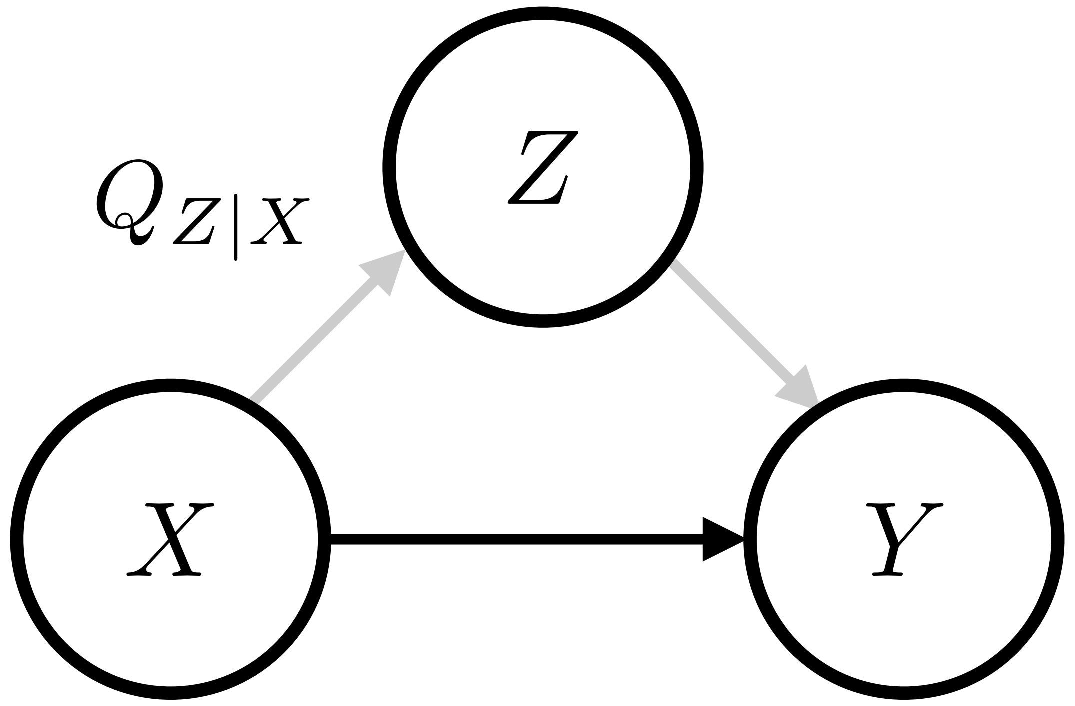
Roadmap of Theoretical Analysis

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

Text Caption:
`Look at my cute cat!`



$\boxed{1} \in$

0	orange cat
1	black cat

ZSP (Indirect) performs poorly ☹

$$\mathbb{E}_{X\sim P_X}\left[(f_\star(X)-\hat{f}(X))^2\right] \le 2\underbrace{\mathbb{E}_{X\sim P_X}\left[(f_\star(X)-\bar{f}(X))^2\right]}_{\text{information-theoretic error}} + 2\underbrace{\mathbb{E}_{X\sim P_X}\left[(\bar{f}(X)-\hat{f}(X))^2\right]}_{\text{learning error}}$$

direct predictor

ZSP procedure

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

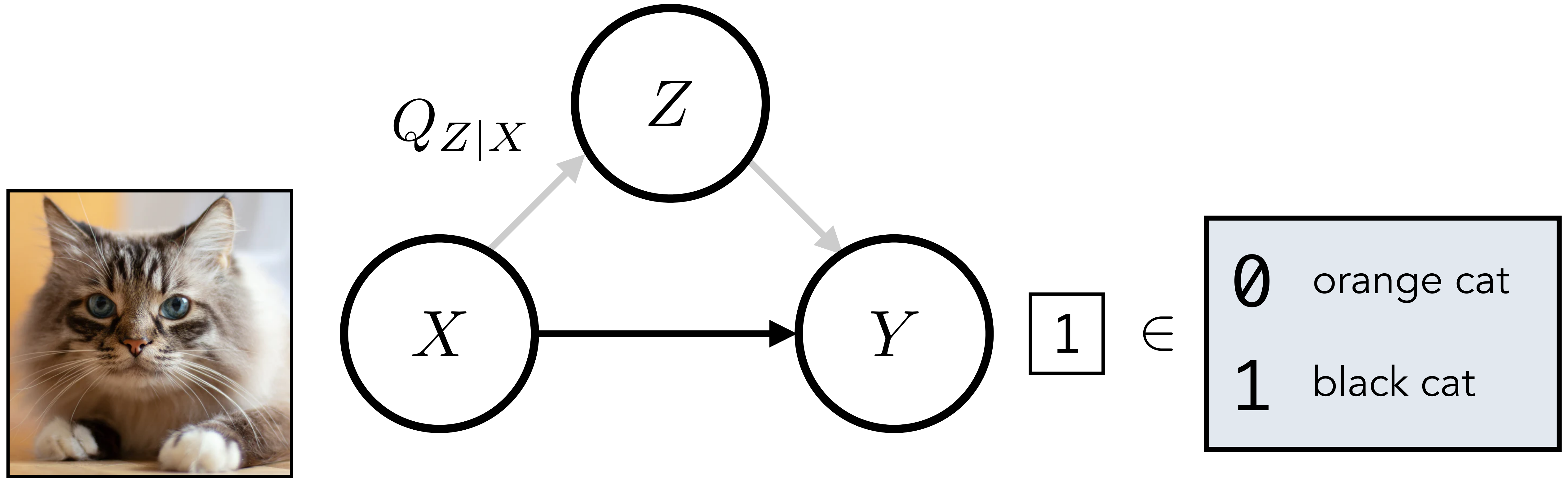
Roadmap of Theoretical Analysis

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

Text Caption: `#nofilter`



ZSP (Indirect) performs poorly ☹

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

direct predictor ZSP procedure

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

Roadmap of Theoretical Analysis

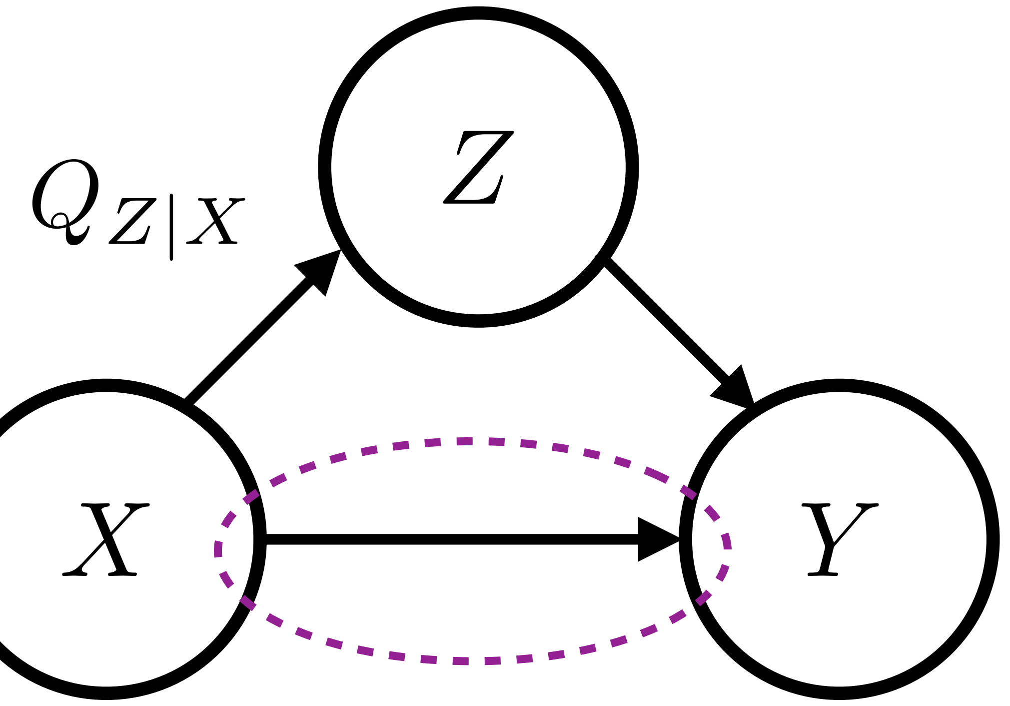
1. Define $\bar{f}$ in terms of pre-training, evaluation, and prompting distribution.
2. Upper bound information-theoretic error using dependence relationships between images, captions, and labels.
3. Define class of estimators $\hat{f}$, and bound learning error using tools from statistical learning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

$X =$ image
$Y =$ label
$Z =$ caption

Theorem. (Mehta & Harchaoui, ICML '25)

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

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



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

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.

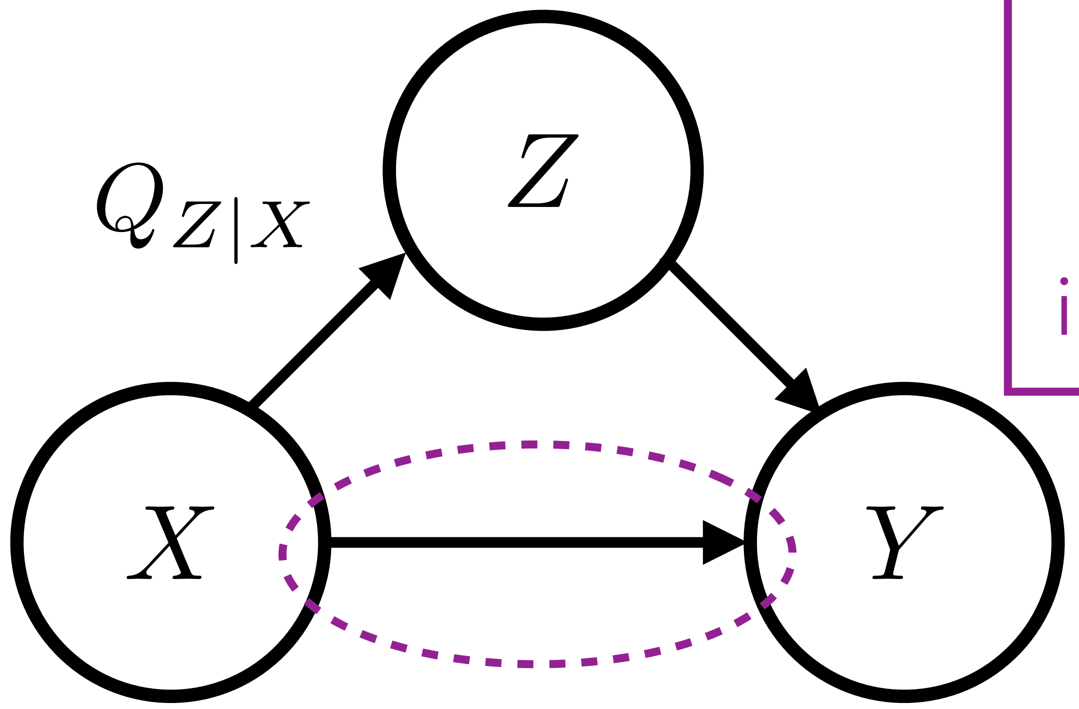
$X =$ image
$Y =$ label
$Z =$ caption

Theorem. (Mehta & Harchaoui, ICML '25)

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

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

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



$$\mathbb{E}_{X\sim P_X}\left[(f_\star(X)-\hat{f}(X))^2\right] \le \underbrace{2\mathbb{E}_{X\sim P_X}\left[(f_\star(X)-\bar{f}(X))^2\right]}_{\text{information-theoretic error}} + \underbrace{2\mathbb{E}_{X\sim P_X}\left[(\bar{f}(X)-\hat{f}(X))^2\right]}_{\text{learning error}}$$

direct predictor ZSP procedure

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

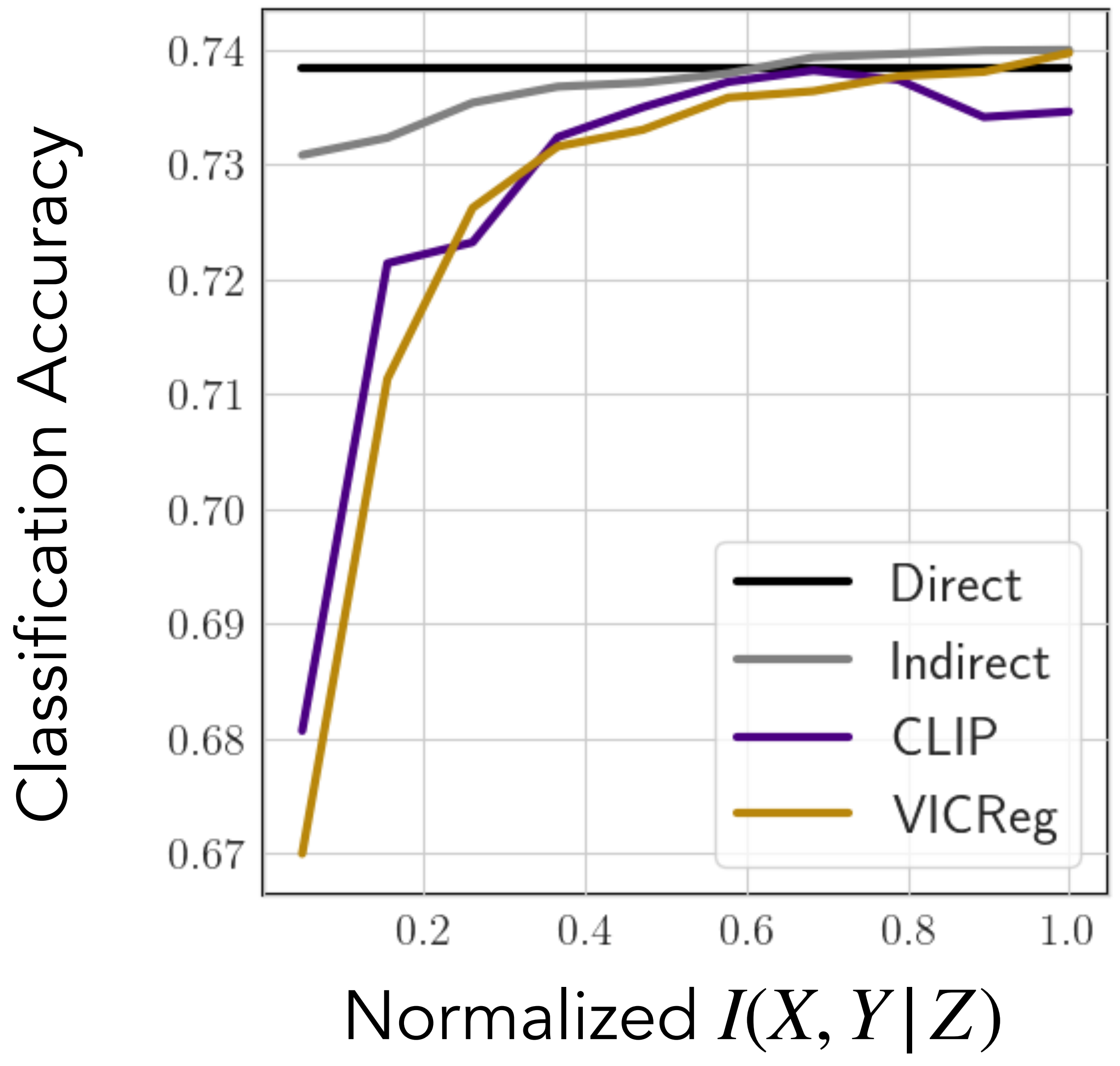
Roadmap of Theoretical Analysis

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

Theorem. (Mehta & Harchaoui, ICML '25)

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



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

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

information-theoretic error

learning error

direct predictor

ZSP procedure

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

Roadmap of Theoretical Analysis

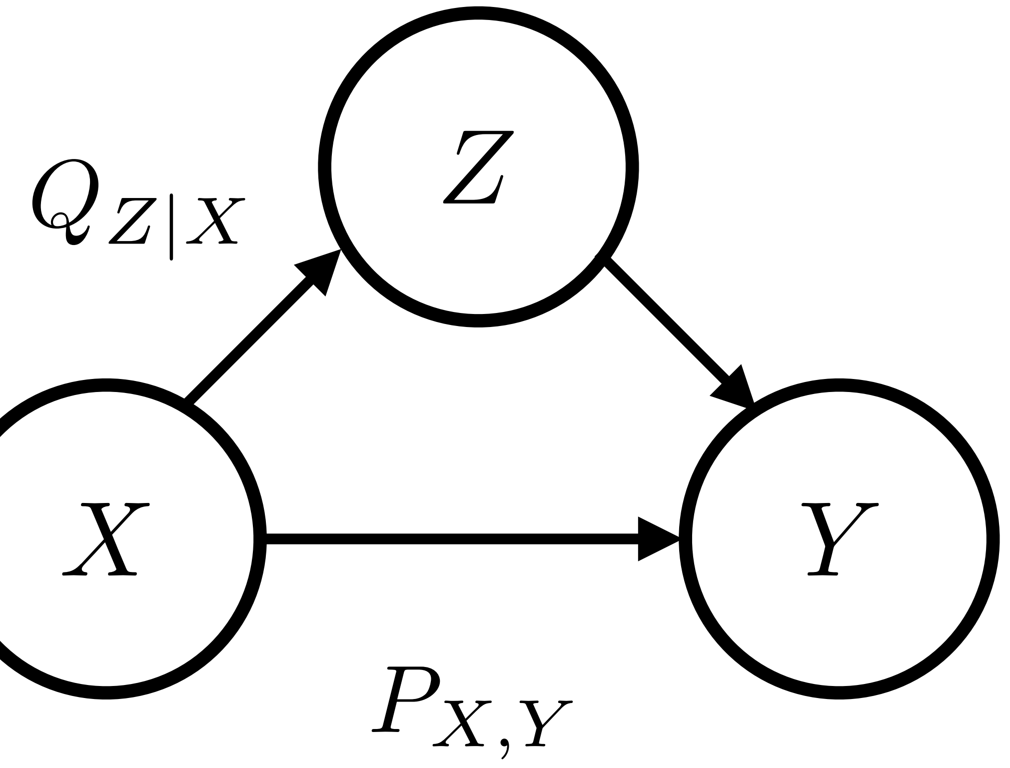
1. Define $\bar{f}$ in terms of pre-training, evaluation, and prompting distribution.
2. Upper bound information-theoretic error using dependence relationships between images, captions, and labels.
3. Define class of estimators $\hat{f}$, and bound learning error using tools from statistical learning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

$X =$ image
$Y =$ label
$Z =$ caption

Theorem. (Mehta & Harchaoui, ICML '25)

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

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



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

information-theoretic error

learning error

direct predictor

ZSP procedure

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

Roadmap of Theoretical Analysis

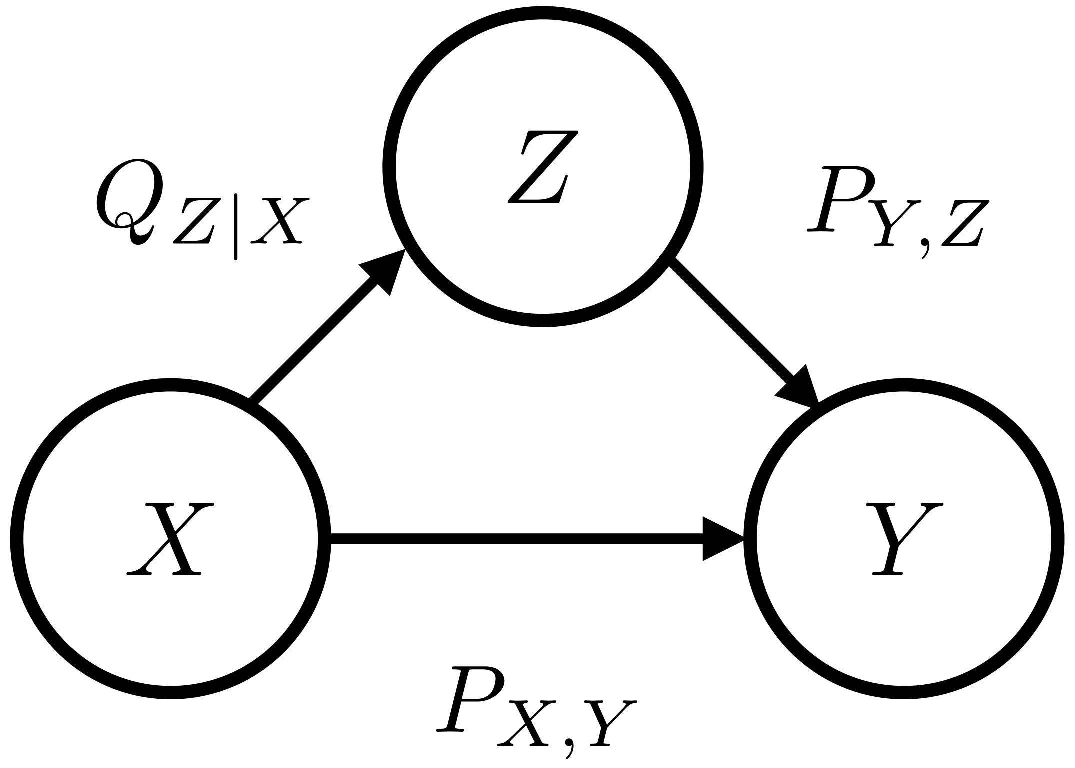
1. Define $\bar{f}$ in terms of pre-training, evaluation, and prompting distribution.
2. Upper bound information-theoretic error using dependence relationships between images, captions, and labels.
3. Define class of estimators $\hat{f}$, and bound learning error using tools from statistical learning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

X = image
Y = label
Z = caption

Theorem. (Mehta & Harchaoui, ICML '25)

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

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



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

direct predictor ($f_\star$)

ZSP procedure ($\hat{f}$)

population version of ZSP ($\bar{f}$)
(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.

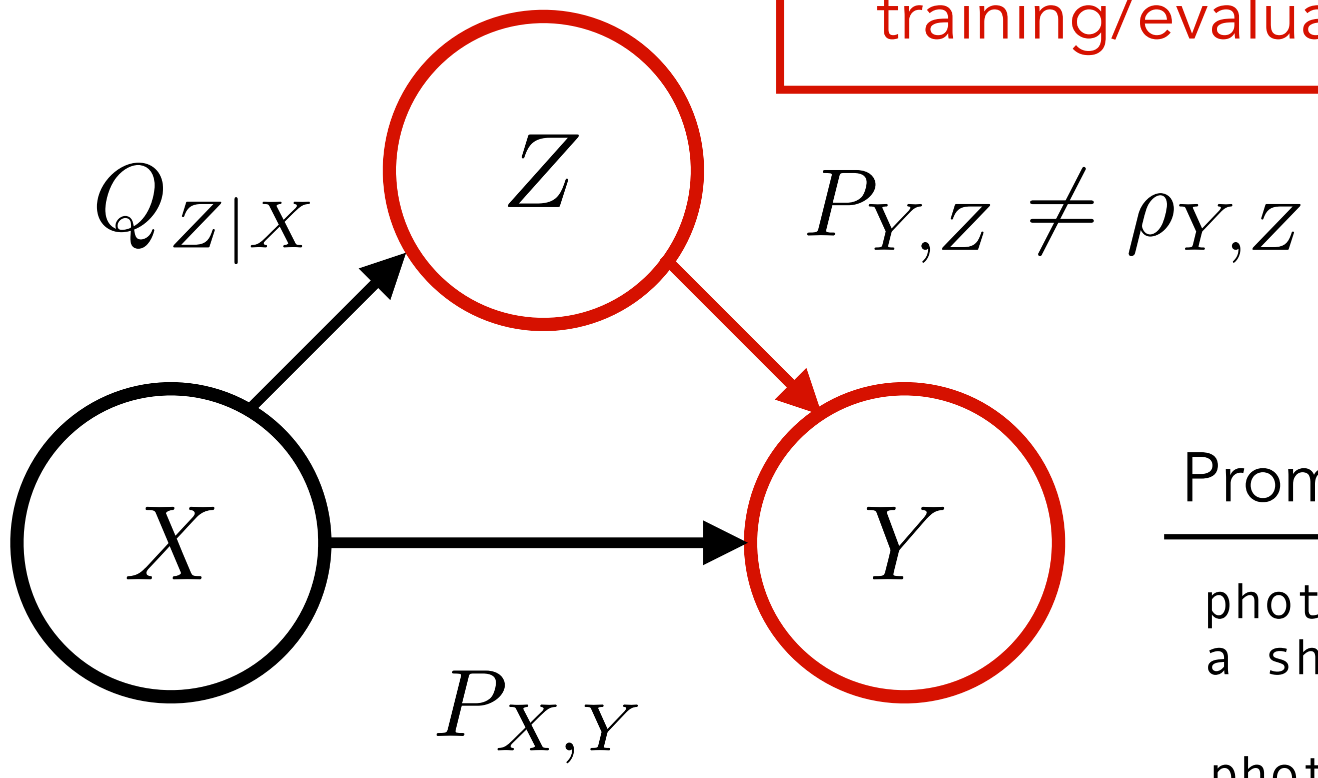
$X =$ image
$Y =$ label
$Z =$ caption

Theorem. (Mehta & Harchaoui, ICML '25)

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

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

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



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

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

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

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

direct predictor

ZSP procedure

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

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

information-theoretic error

learning error

direct predictor

ZSP procedure

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

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

based on **population** distributions

learned from **data**

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

direct predictor

ZSP procedure

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

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

based on **population** distributions	$\bar{f}(\boldsymbol{x}) = \mathbb{E}_{Q_{X,Z}}\left[\mathbb{E}_{\rho_{Y,Z}}\left[Y\|Z\right]\|X=\boldsymbol{x}\right]$
learned from **data**	

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

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	$$\begin{aligned}\bar{f}(\boldsymbol{x}) &= \mathbb{E}_{Q_{X,Z}}\left[\mathbb{E}_{\rho_{Y,Z}}\left[Y\vert Z\right]\vert X=\boldsymbol{x}\right] \\ &= \mathbb{E}_{\rho_{Y,Z}}[Y\cdot \underbrace{\mathsf{R}(\boldsymbol{x},Z)]}_{\text{similarity score}} + \underbrace{\mathrm{err}(Q_Z,\rho_Z)}_{\text{small if prompts "cover" the pre-training distribution}}\end{aligned}$$
learned from **data**	

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

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

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

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

$$= \mathbb{E}_{\rho_{Y,Z}}[Y \cdot \underbrace{\mathsf{R}(\boldsymbol{x}, Z)]}_{\text{similarity score}} + \underbrace{\mathrm{err}(Q_Z, \rho_Z)}_{\text{small if prompts "cover" the pre-training distribution}}$$

learned from **data**

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

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

$$\mathbb{E}_{X\sim P_X}\left[(f_\star(X)-\hat{f}(X))^2\right] \le \underbrace{2\mathbb{E}_{X\sim P_X}\left[(f_\star(X)-\bar{f}(X))^2\right]}_{\text{information-theoretic error}} + \underbrace{2\mathbb{E}_{X\sim P_X}\left[(\bar{f}(X)-\hat{f}(X))^2\right]}_{\text{learning error}}$$

direct predictor ZSP procedure

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

Approach 1: Similarity Score Learning

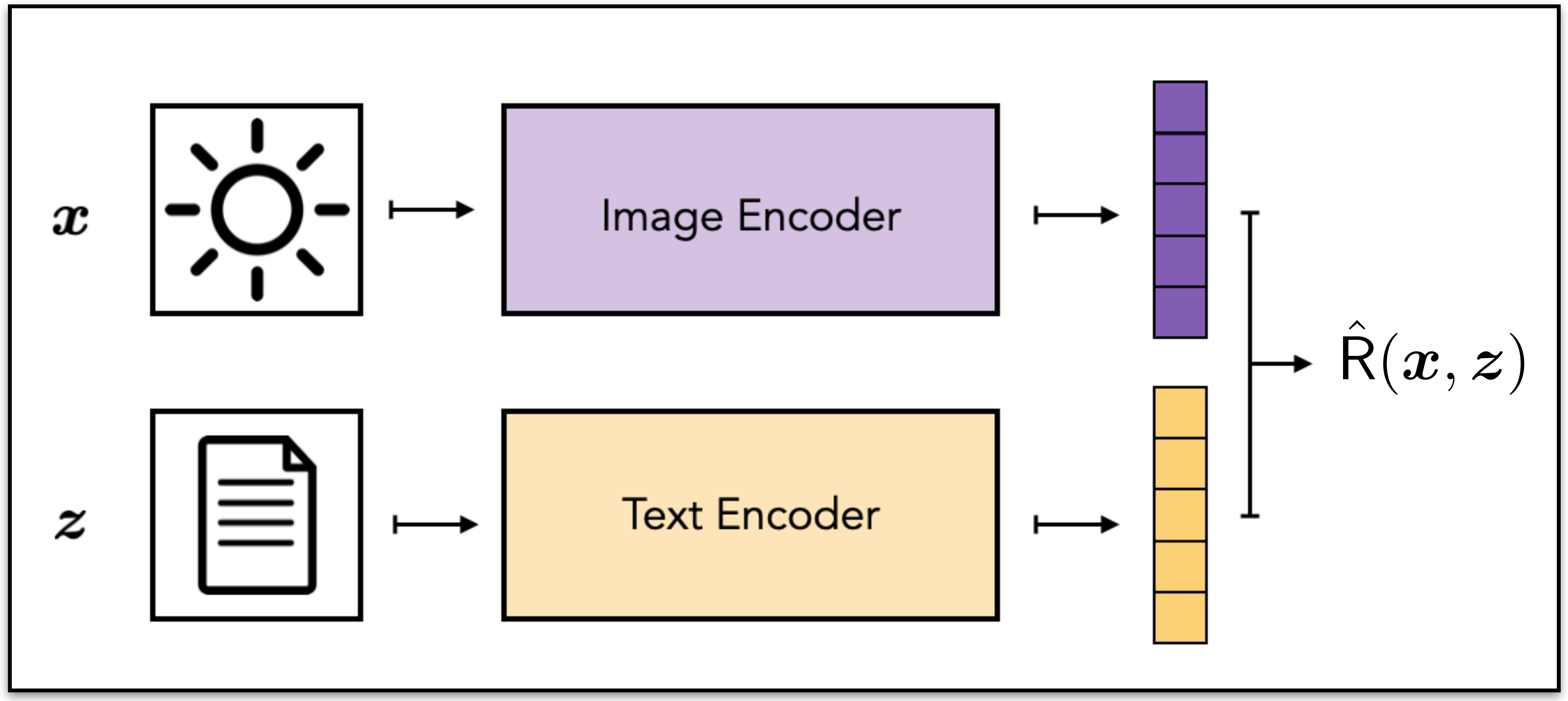
$P_{X,Y}$	$Q_{X,Z}$	$\rho_{Y,Z}$
Evaluation	Pre-Training	Prompting

based on **population** distributions

$$\bar{f}(\boldsymbol{x}) = \mathbb{E}_{Q_{X,Z}}\left[\mathbb{E}_{\rho_{Y,Z}}\left[Y|Z\right]|X=\boldsymbol{x}\right]$$
$$= \mathbb{E}_{\rho_{Y,Z}}\left[Y\cdot \underbrace{\mathsf{R}(\boldsymbol{x},Z)}_{\text{similarity score}}\right] + \underbrace{\mathrm{err}(Q_Z,\rho_Z)}_{\text{small if prompts "cover" the pre-training distribution}}$$

learned from **data**

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



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

$$\mathbb{E}_{X\sim P_X}\left[(f_\star(X)-\hat{f}(X))^2\right] \le \underbrace{2\mathbb{E}_{X\sim P_X}\left[(f_\star(X)-\bar{f}(X))^2\right]}_{\text{information-theoretic error}} + \underbrace{2\mathbb{E}_{X\sim P_X}\left[(\bar{f}(X)-\hat{f}(X))^2\right]}_{\text{learning error}}$$

direct predictor ZSP procedure

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

Approach 1: Similarity Score Learning

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

based on **population** distributions

$$\bar{f}(\boldsymbol{x}) = \mathbb{E}_{Q_{X,Z}}\left[\mathbb{E}_{\rho_{Y,Z}}[Y|Z]\,|X=\boldsymbol{x}\right]$$
$$= \mathbb{E}_{\rho_{Y,Z}}[Y\cdot \underbrace{\mathsf{R}(\boldsymbol{x},Z)]}_{\text{similarity score}} + \underbrace{\mathrm{err}(Q_Z,\rho_Z)}_{\text{small if prompts “cover” the pre-training distribution}}$$

learned from **data**

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

$$M\left[\begin{array}{l}(\boldsymbol{z}_1^{\boldsymbol{y}})_{\boldsymbol{y}\in\mathcal{Y}}\\ (\boldsymbol{z}_2^{\boldsymbol{y}})_{\boldsymbol{y}\in\mathcal{Y}}\\ (\boldsymbol{z}_3^{\boldsymbol{y}})_{\boldsymbol{y}\in\mathcal{Y}}\end{array}\right.$$

“photo of a **sun**”
“photo of a **moon**”
“photo of a **comet**”
“photo of a **planet**”

Idea: Convert labels into **prompts** (pseudo-captions)

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

$$\mathbb{E}_{X\sim P_X}\left[(\underbrace{f_\star}_{\text{direct predictor}}(X)-\underbrace{\hat{f}}_{\text{ZSP procedure}}(X))^2\right] \leq \underbrace{2\mathbb{E}_{X\sim P_X}\left[(f_\star(X)-\bar{f}(X))^2\right]}_{\text{information-theoretic error}} + \underbrace{2\mathbb{E}_{X\sim P_X}\left[(\bar{f}(X)-\hat{f}(X))^2\right]}_{\text{learning error}}$$

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

Approach 1: Similarity Score Learning

Theorem. (Mehta & Harchaoui, ICML '25)

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

based on **population** distributions	$\bar{f}(\boldsymbol{x}) = \mathbb{E}_{Q_{X,Z}}\left[\mathbb{E}_{\rho_{Y,Z}}[Y \mid Z] \mid X=\boldsymbol{x}\right]$ $= \mathbb{E}_{\rho_{Y,Z}}[Y \cdot \underbrace{\mathsf{R}(\boldsymbol{x}, Z)}_{\text{similarity score}}] + \underbrace{\mathrm{err}(Q_Z, \rho_Z)}_{\text{small if prompts ``cover'' the pre-training distribution}}$
learned from **data**	$\hat{f}(\boldsymbol{x}) = \mathbb{E}_{\hat{\rho}_{Y,Z}}[Y \cdot \hat{\mathsf{R}}(\boldsymbol{x}, Z)]$

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

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

direct predictor ZSP procedure

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

Approach 1: Similarity Score Learning

Theorem. (Mehta & Harchaoui, ICML '25)

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

based on **population** distributions

$$\begin{aligned}\bar{f}(\boldsymbol{x}) &= \mathbb{E}_{Q_{X,Z}}\left[\mathbb{E}_{\rho_{Y,Z}}[Y|Z]\,|X = \boldsymbol{x}\right] \\ &= \mathbb{E}_{\rho_{Y,Z}}\underbrace{[Y \cdot \mathsf{R}(\boldsymbol{x}, Z)]}_{\text{similarity score}} + \underbrace{\mathrm{err}(Q_Z, \rho_Z)}_{\text{small if prompts “cover” the pre-training distribution}}\end{aligned}$$

learned from **data**

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

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

$$\mathbb{E}_{X\sim P_X}\left[(\underbrace{f_\star}_{\text{direct predictor}}(X)-\underbrace{\hat f}_{\text{ZSP procedure}}(X))^2\right] \le \underbrace{2\mathbb{E}_{X\sim P_X}\left[(f_\star(X)-\bar f(X))^2\right]}_{\text{information-theoretic error}} + \underbrace{2\mathbb{E}_{X\sim P_X}\left[(\bar f(X)-\hat f(X))^2\right]}_{\text{learning error}}$$

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

Approach 1: Similarity Score Learning

Theorem. (Mehta & Harchaoui, ICML '25)

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

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

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

based on **population** distributions

$$\bar f(\boldsymbol{x}) = \mathbb{E}_{Q_{X,Z}}\left[\mathbb{E}_{\rho_{Y,Z}}[Y|Z]\,|X=\boldsymbol{x}\right]$$
$$= \mathbb{E}_{\rho_{Y,Z}}[Y\cdot \underbrace{\mathsf{R}(\boldsymbol{x},Z)]}_{\text{similarity score}} + \underbrace{\mathrm{err}(Q_Z,\rho_Z)}_{\text{small if prompts "cover" the pre-training distribution}}$$

learned from **data**

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

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

$$\mathbb{E}_{X\sim P_X}\left[(\underset{\text{direct predictor}}{f_\star}(X)-\underset{\text{ZSP procedure}}{\hat{f}}(X))^2\right] \le \underbrace{2\mathbb{E}_{X\sim P_X}\left[(f_\star(X)-\bar{f}(X))^2\right]}_{\text{information-theoretic error}} + \underbrace{2\mathbb{E}_{X\sim P_X}\left[(\bar{f}(X)-\hat{f}(X))^2\right]}_{\text{learning error}}$$

direct predictor

ZSP procedure

population version of ZSP ($\bar{f}$)
(based on distributions instead of samples)

Approach 1: Similarity Score Learning

Theorem. (Mehta & Harchaoui, ICML '25)

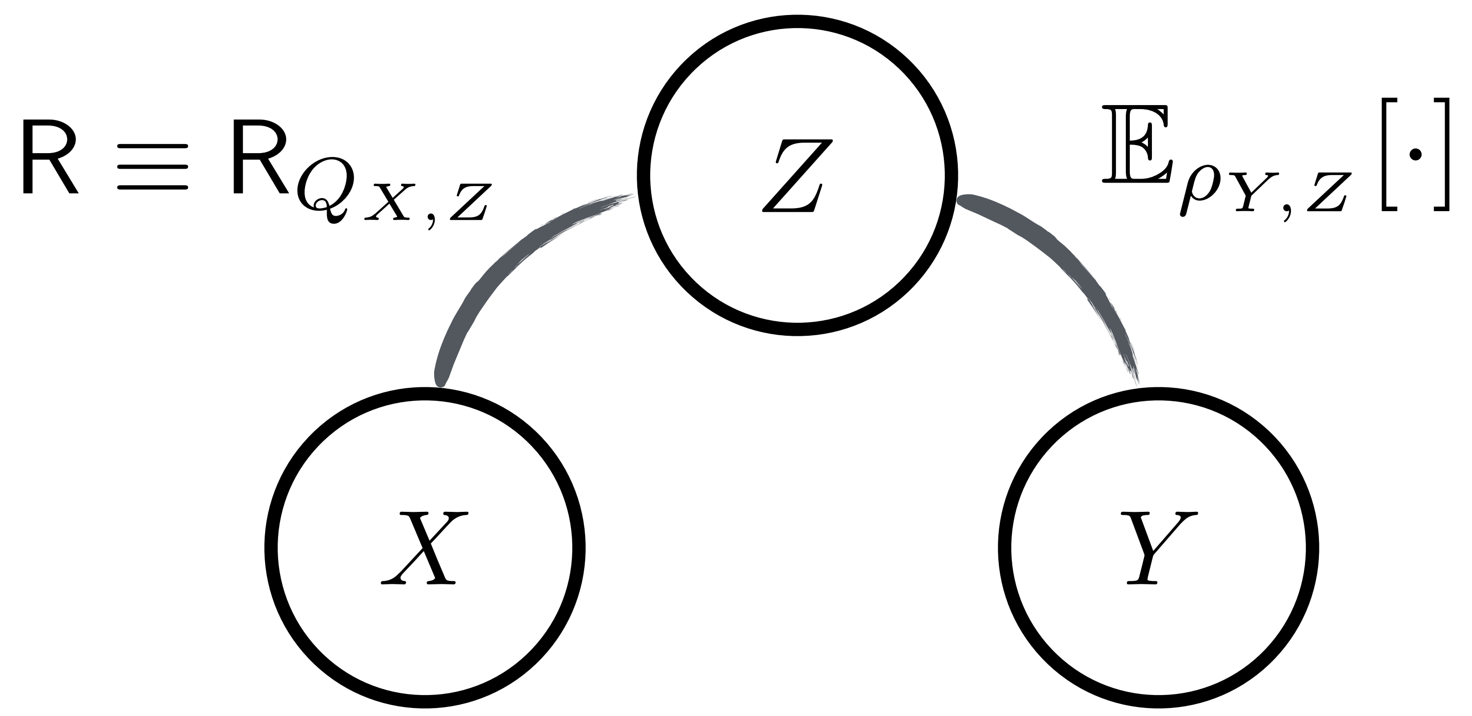
$$\mathbb{E}_{X\sim P_X}\left[(\bar{f}(X)-\hat{f}(X))^2\right] \lesssim d(\hat{\mathsf{R}},\mathsf{R}) + d(\hat{\rho}_{Y,Z},\rho_{Y,Z})$$

based on **population** distributions

$$\bar{f}(\boldsymbol{x}) = \mathbb{E}_{Q_{X,Z}}\left[\mathbb{E}_{\rho_{Y,Z}}[Y|Z]\,|X=\boldsymbol{x}\right]$$
$$= \mathbb{E}_{\rho_{Y,Z}}\underbrace{[Y\cdot\mathsf{R}(\boldsymbol{x},Z)]}_{\text{similarity score}} + \underbrace{\mathrm{err}(Q_Z,\rho_Z)}_{\text{small if prompts "cover" the pre-training distribution}}$$

learned from **data**

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



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

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

direct predictor ZSP procedure

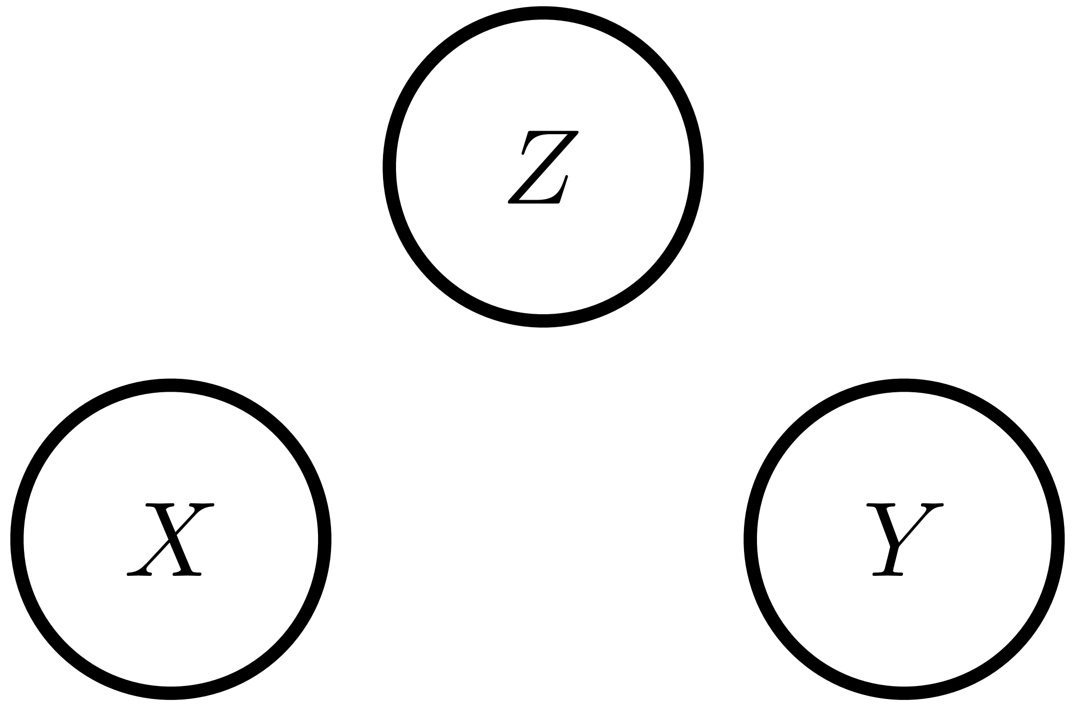
population version of ZSP
(based on distributions instead of samples)

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	$\bar{f}(\boldsymbol{x}) = \mathbb{E}_{Q_{X,Z}}\left[\mathbb{E}_{\rho_{Y,Z}}\left[Y\|Z\right]\|X = \boldsymbol{x}\right]$
learned from **data**	$\hat{f}(\boldsymbol{x}) = \hat{g}_M(\hat{h}_N(\boldsymbol{x}))$



$$\mathbb{E}_{X\sim P_X}\left[(f_\star(X)-\hat{f}(X))^2\right] \le \underbrace{2\mathbb{E}_{X\sim P_X}\left[(f_\star(X)-\bar{f}(X))^2\right]}_{\text{information-theoretic error}} + \underbrace{2\mathbb{E}_{X\sim P_X}\left[(\bar{f}(X)-\hat{f}(X))^2\right]}_{\text{learning error}}$$

direct predictor ZSP procedure

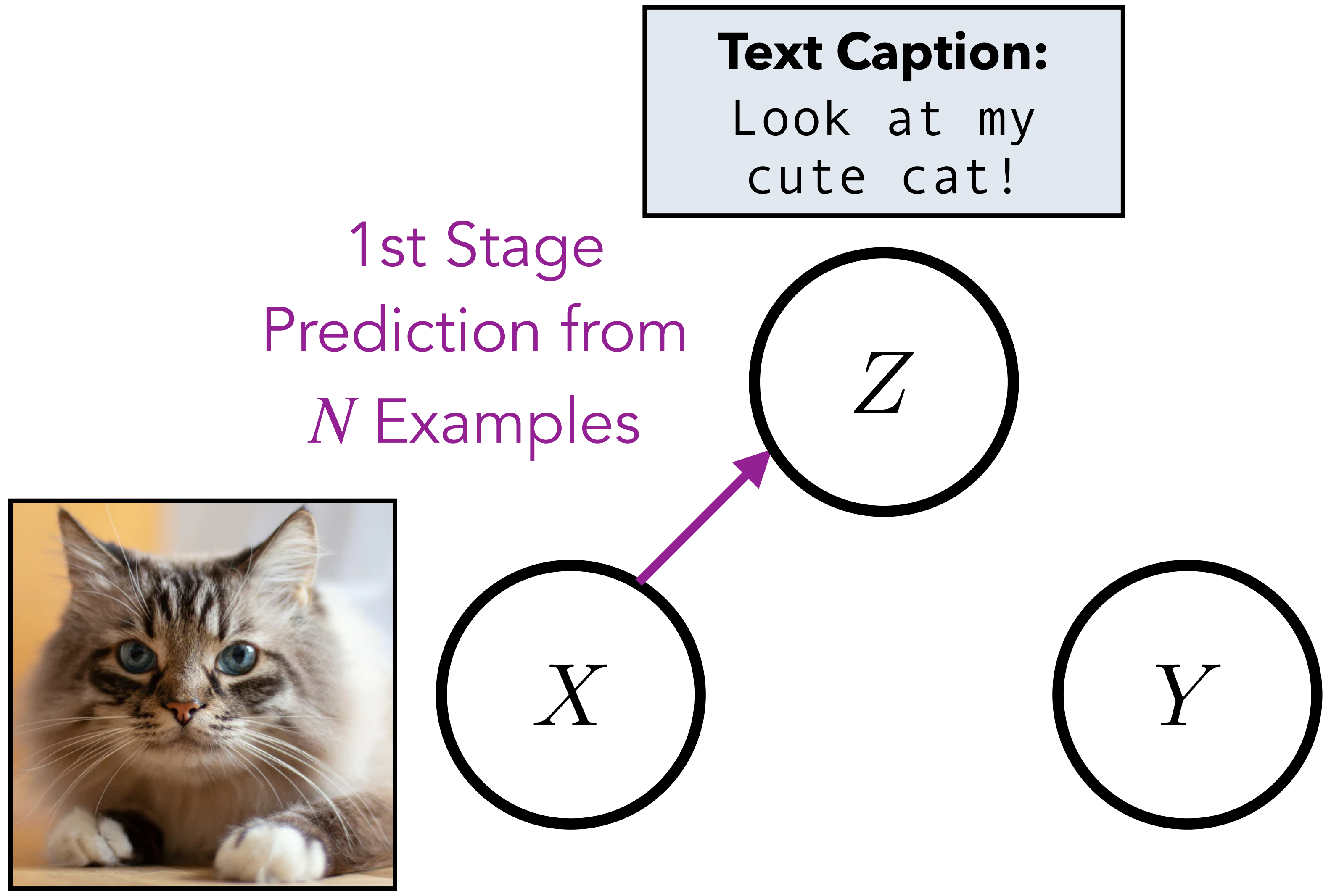
population version of ZSP
(based on distributions instead of samples)

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	$\bar{f}(\boldsymbol{x}) = \mathbb{E}_{Q_{X,Z}}\left[\underbrace{\mathbb{E}_{\rho_{Y,Z}}[Y\vert Z]}_{\text{1st Stage}}\,\vert X=\boldsymbol{x}\right]$
learned from **data**	$\hat{f}(\boldsymbol{x}) = \hat{g}_M(\underline{\hat{h}_N}(\boldsymbol{x}))$



$$\mathbb{E}_{X\sim P_X}\left[(f_\star(X) - \hat{f}(X))^2\right] \le 2\underbrace{\mathbb{E}_{X\sim P_X}\left[(f_\star(X) - \bar{f}(X))^2\right]}_{\text{information-theoretic error}} + 2\underbrace{\mathbb{E}_{X\sim P_X}\left[(\bar{f}(X) - \hat{f}(X))^2\right]}_{\text{learning error}}$$

direct predictor

ZSP procedure

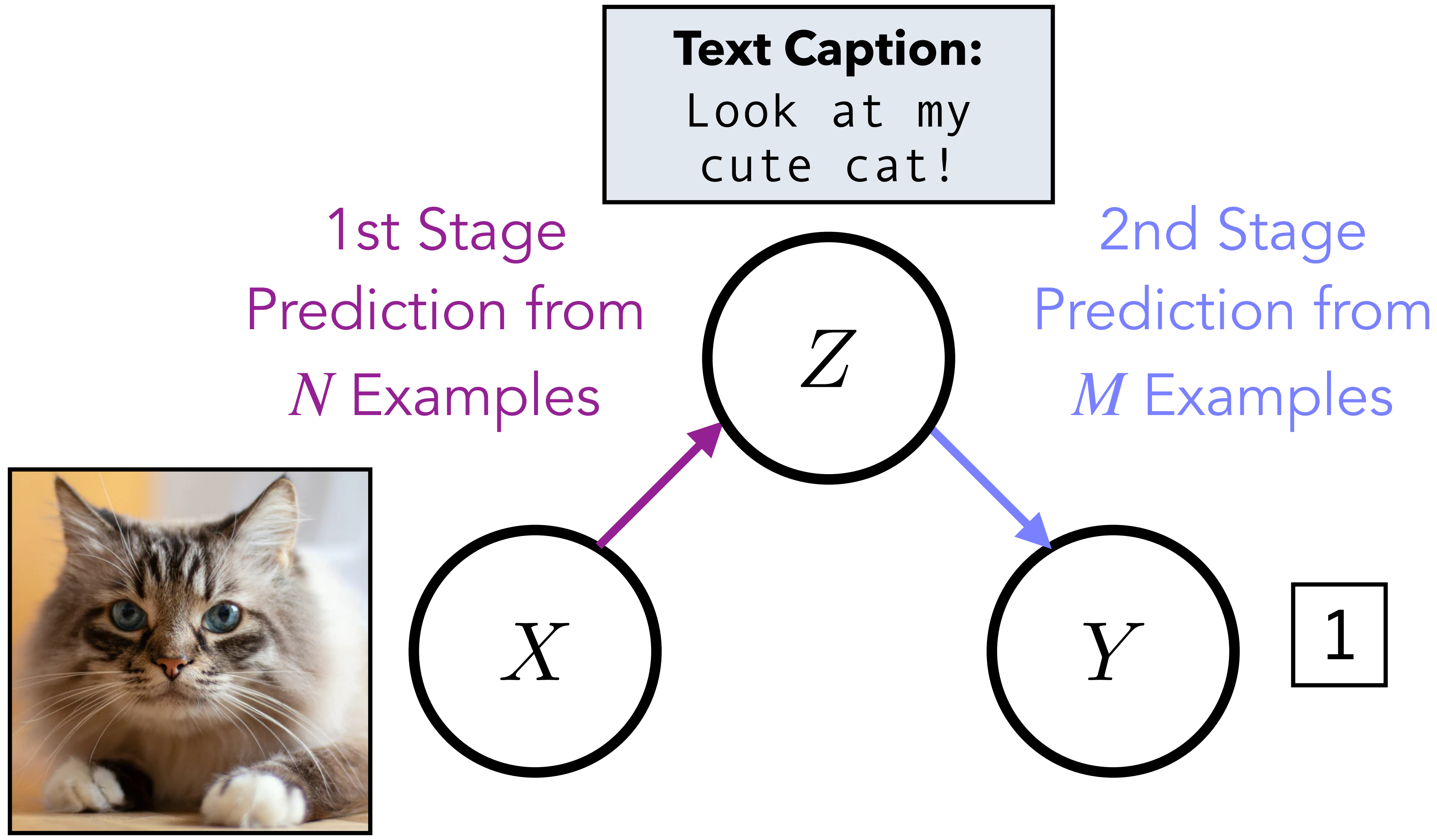
population version of ZSP
(based on distributions instead of samples)

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	$\bar{f}(\boldsymbol{x}) = \underbrace{\mathbb{E}_{Q_{X,Z}}\left[\underbrace{\mathbb{E}_{\rho_{Y,Z}}[Y\vert Z]}_{\text{1st Stage}}\,\vert X = \boldsymbol{x}\right]}_{\text{2nd Stage}}$
learned from **data**	$\hat{f}(\boldsymbol{x}) = \underline{\hat{g}_M}(\underline{\hat{h}_N}(\boldsymbol{x}))$



$$\mathbb{E}_{X\sim P_X}\left[(\underset{\text{direct predictor}}{f_\star(X)} - \underset{\text{ZSP procedure}}{\hat{f}(X)})^2\right] \le \underbrace{2\mathbb{E}_{X\sim P_X}\left[(f_\star(X) - \bar{f}(X))^2\right]}_{\text{information-theoretic error}} + \underbrace{2\mathbb{E}_{X\sim P_X}\left[(\bar{f}(X) - \hat{f}(X))^2\right]}_{\text{learning error}}$$

direct predictor ZSP procedure

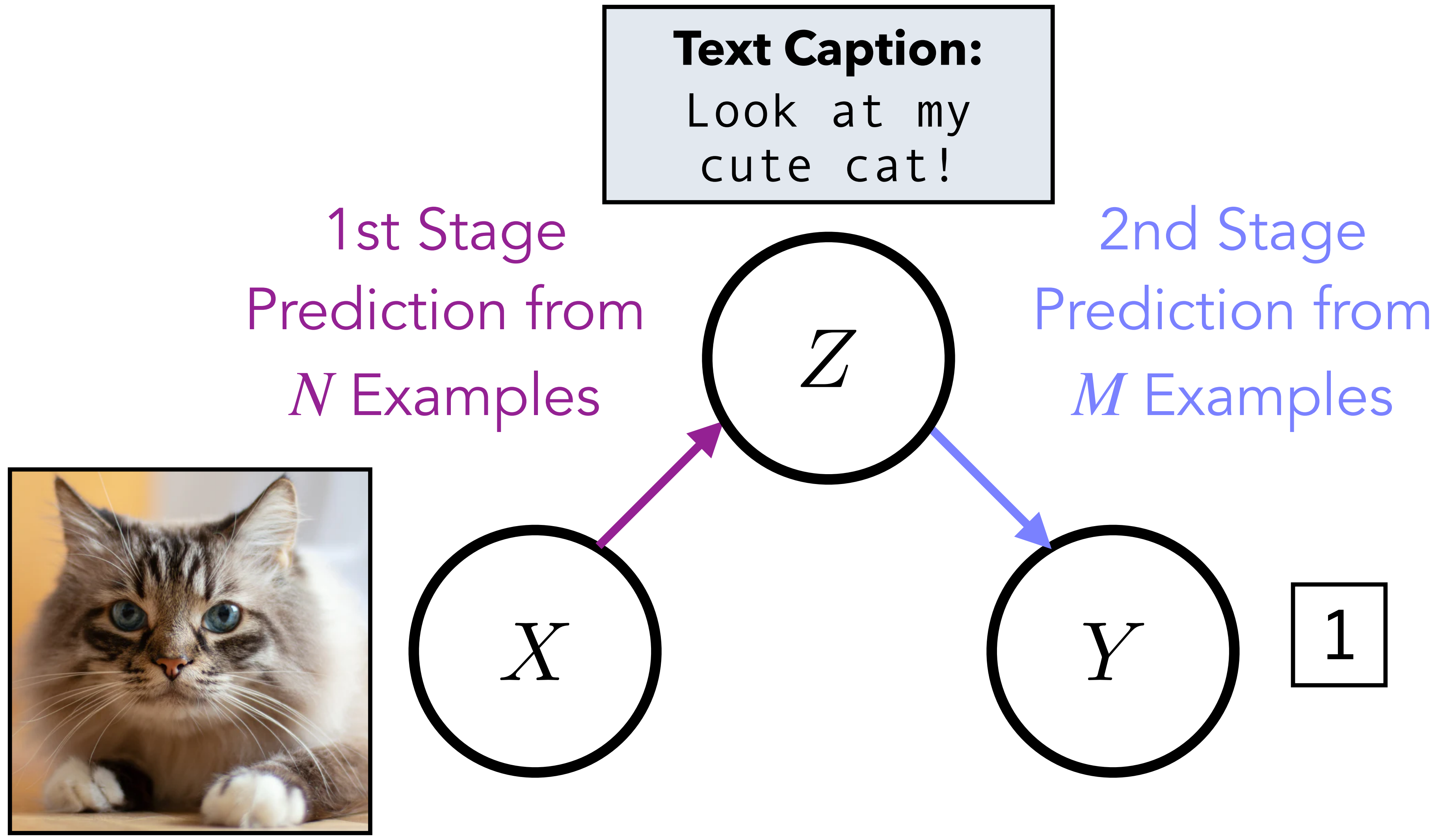
population version of ZSP
(based on distributions instead of samples)

Approach 2: Two-Stage Prediction

Theorem. (Mehta & Harchaoui, ICML '25)

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

based on **population** distributions	$\bar{f}(\boldsymbol{x}) = \underbrace{\mathbb{E}_{Q_{X,Z}}\left[\underbrace{\mathbb{E}_{\rho_{Y,Z}}[Y \mid Z]}_{\text{1st Stage}} \mid X = \boldsymbol{x}\right]}_{\text{2nd Stage}}$
learned from **data**	$\hat{f}(\boldsymbol{x}) = \underline{\hat{g}_M}(\underline{\hat{h}_N}(\boldsymbol{x}))$



$\boxed{1}$

$$\mathbb{E}_{X\sim P_X}\left[(f_\star(X)-\hat{f}(X))^2\right] \le \underbrace{2\mathbb{E}_{X\sim P_X}\left[(f_\star(X)-\bar{f}(X))^2\right]}_{\text{information-theoretic error}} + \underbrace{2\mathbb{E}_{X\sim P_X}\left[(\bar{f}(X)-\hat{f}(X))^2\right]}_{\text{learning error}}$$

direct predictor

ZSP procedure

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

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

direct predictor

ZSP procedure

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.

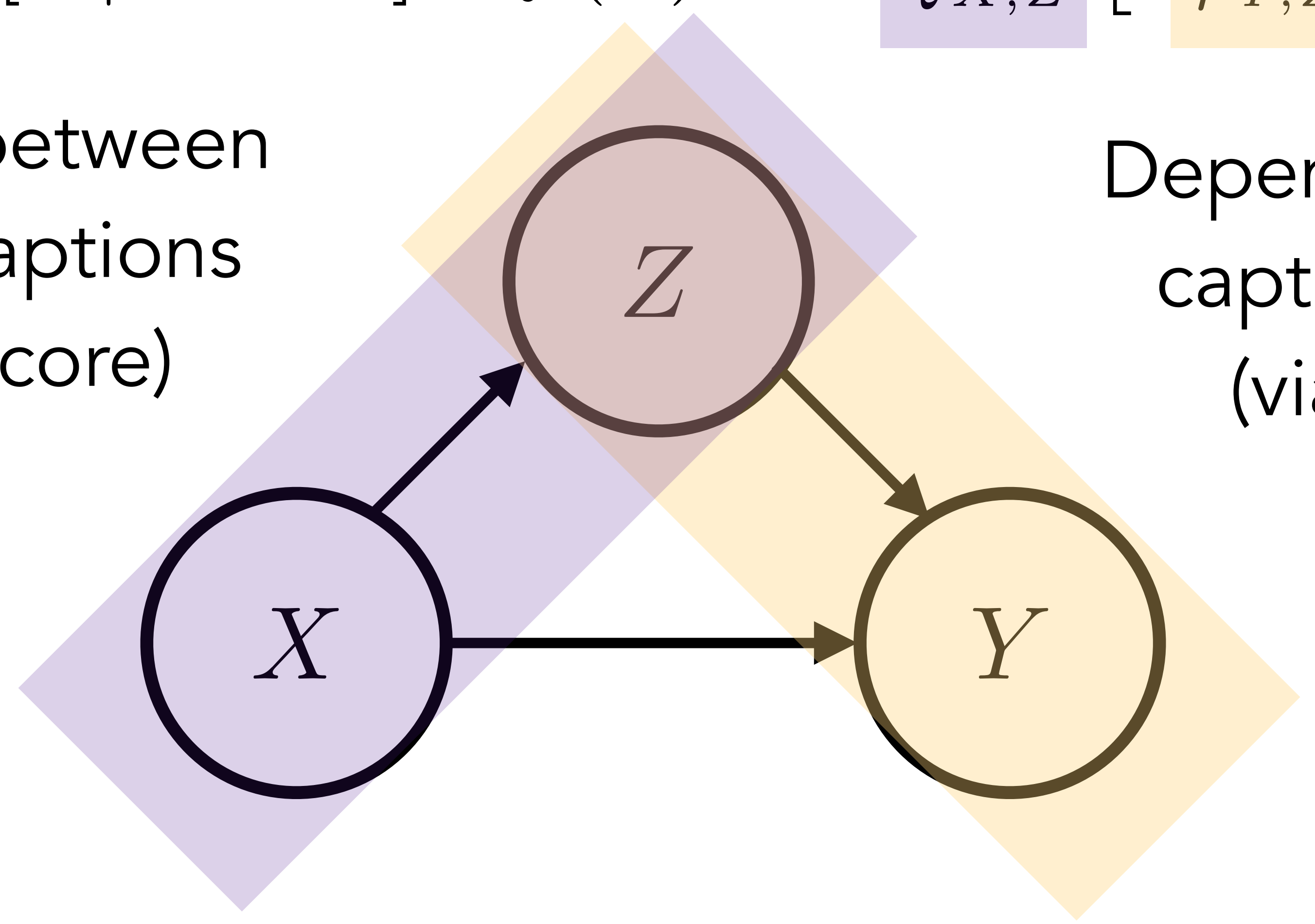
2. Two proof strategies which apply to different classes of methods.

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

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

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



Dependence between captions and labels (via prompting)

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

direct predictor — ZSP procedure

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

residual dependence — prompt bias — sample complexity — prompt complexity

Contributions

1. Theoretical framework to formalize zero-shot prediction (ZSP) and obtain its generalization analysis.

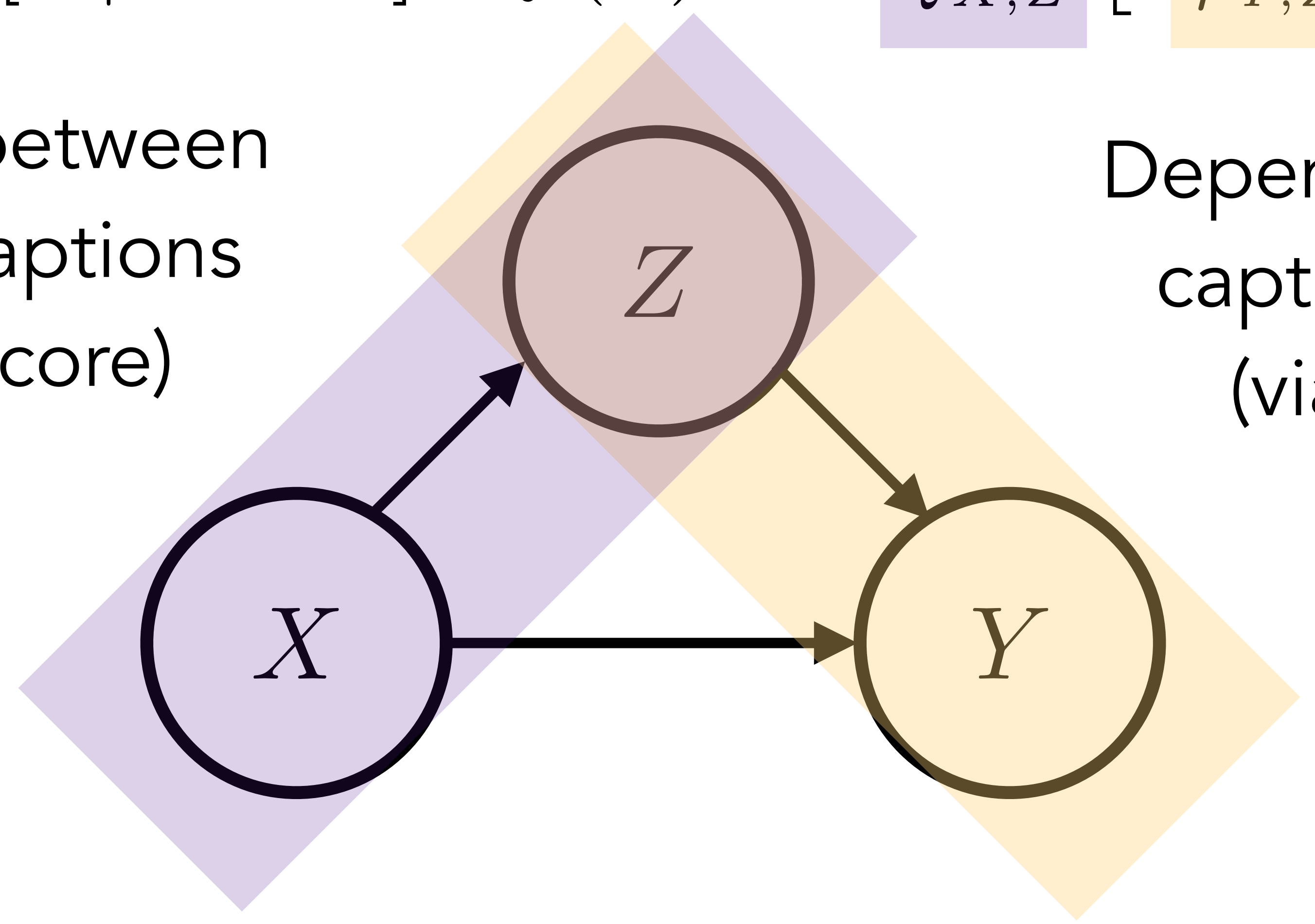
2. Two proof strategies which apply to different classes of methods.

3. Key quantities for success of ZSP: residual dependence, prompt bias, sample complexity, and prompt complexity.

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

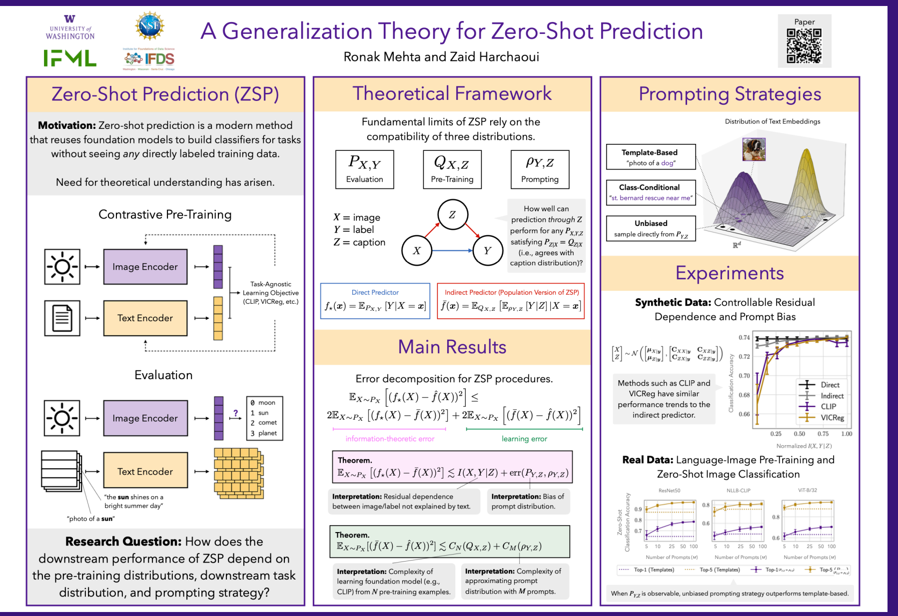
$$f_\star(\boldsymbol{x}) = \mathbb{E}_{P_{X,Y}}\left[Y|X=\boldsymbol{x}\right] \qquad \bar{f}(\boldsymbol{x}) = \mathbb{E}_{Q_{X,Z}}\left[\mathbb{E}_{\rho_{Y,Z}}\left[Y|Z\right]|X=\boldsymbol{x}\right]$$

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



Dependence between captions and labels (via prompting)

Thank you!

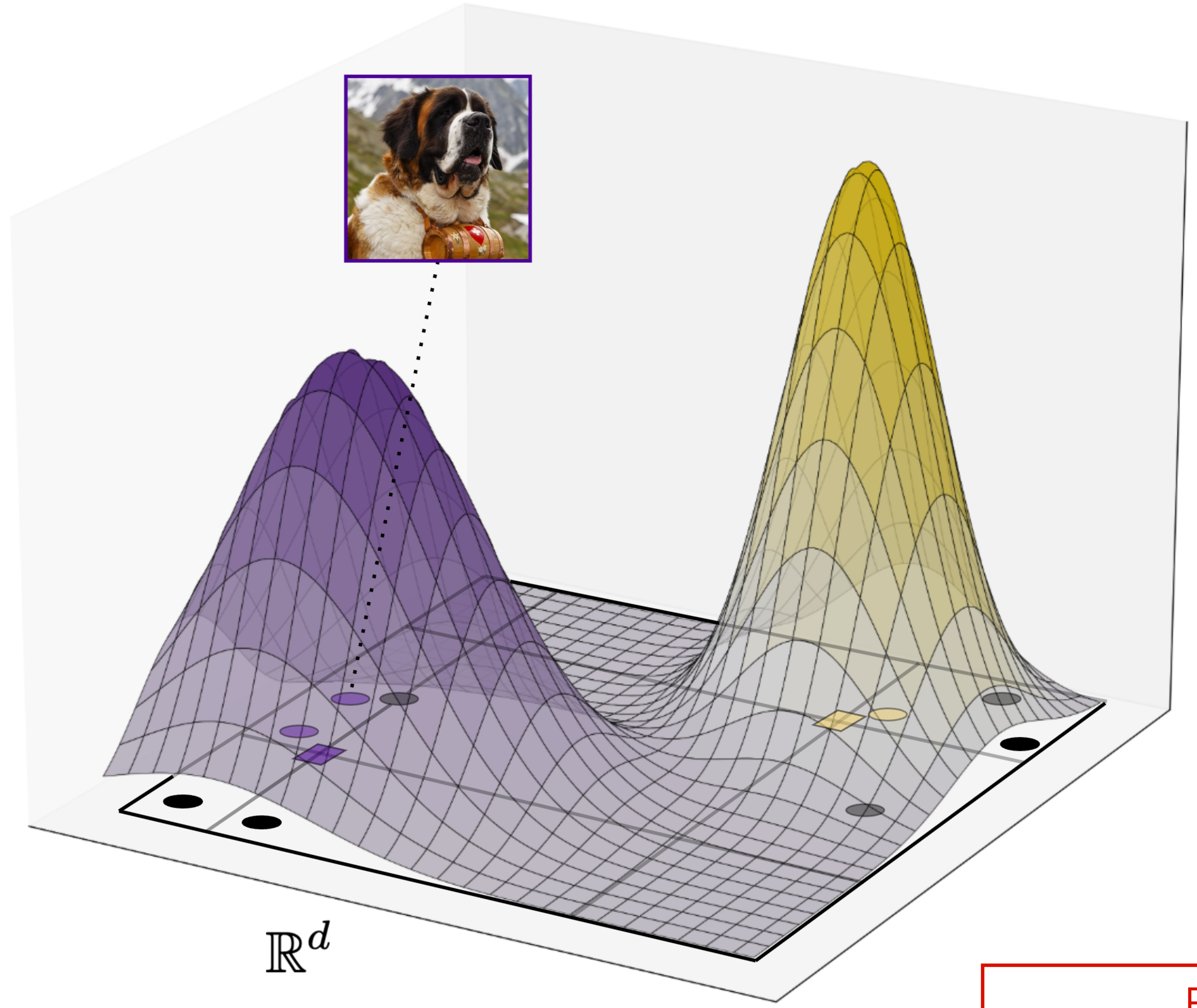


Spotlight Poster Info

Time: Wednesday, July 16, 11:00am PDT - 1:30pm PDT
Place: West Exhibition Hall B2-B3 #W-905

IFDS Institute for Foundations of Data Science
NSF
IFML
UNIVERSITY OF WASHINGTON · LVX · SIT · 1861

Appendix

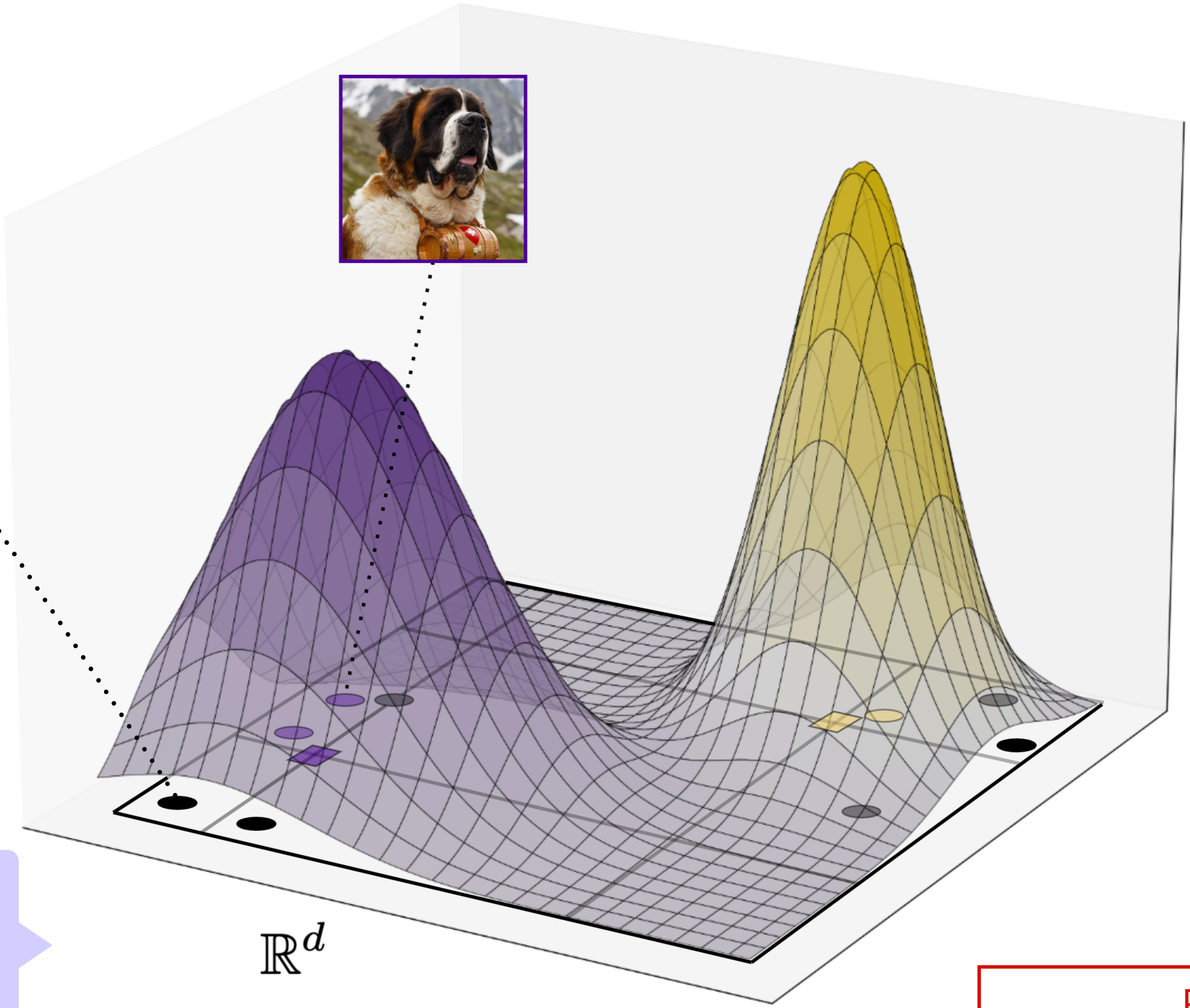


Distribution of Text Embeddings

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

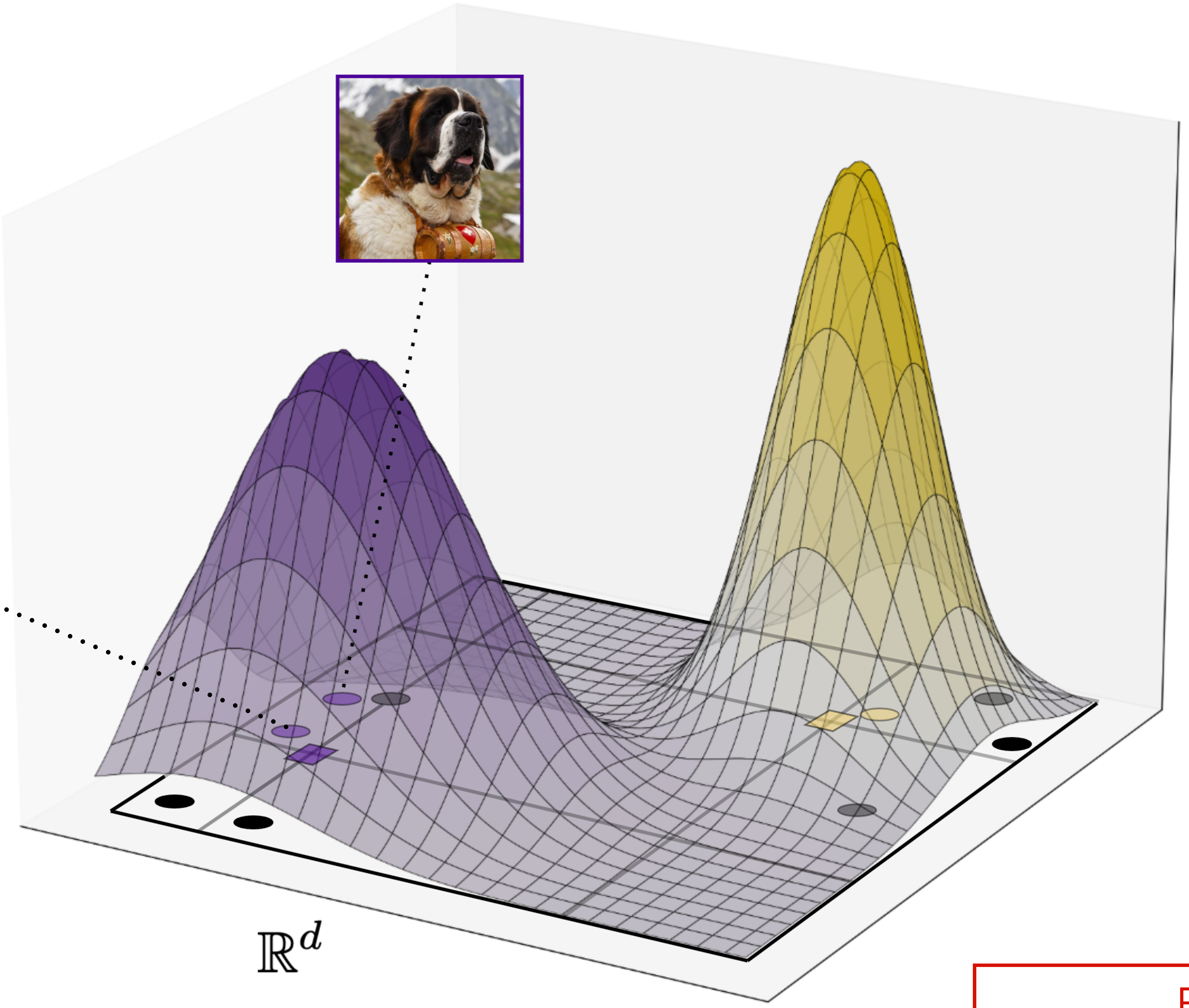
Distribution of Text Embeddings

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

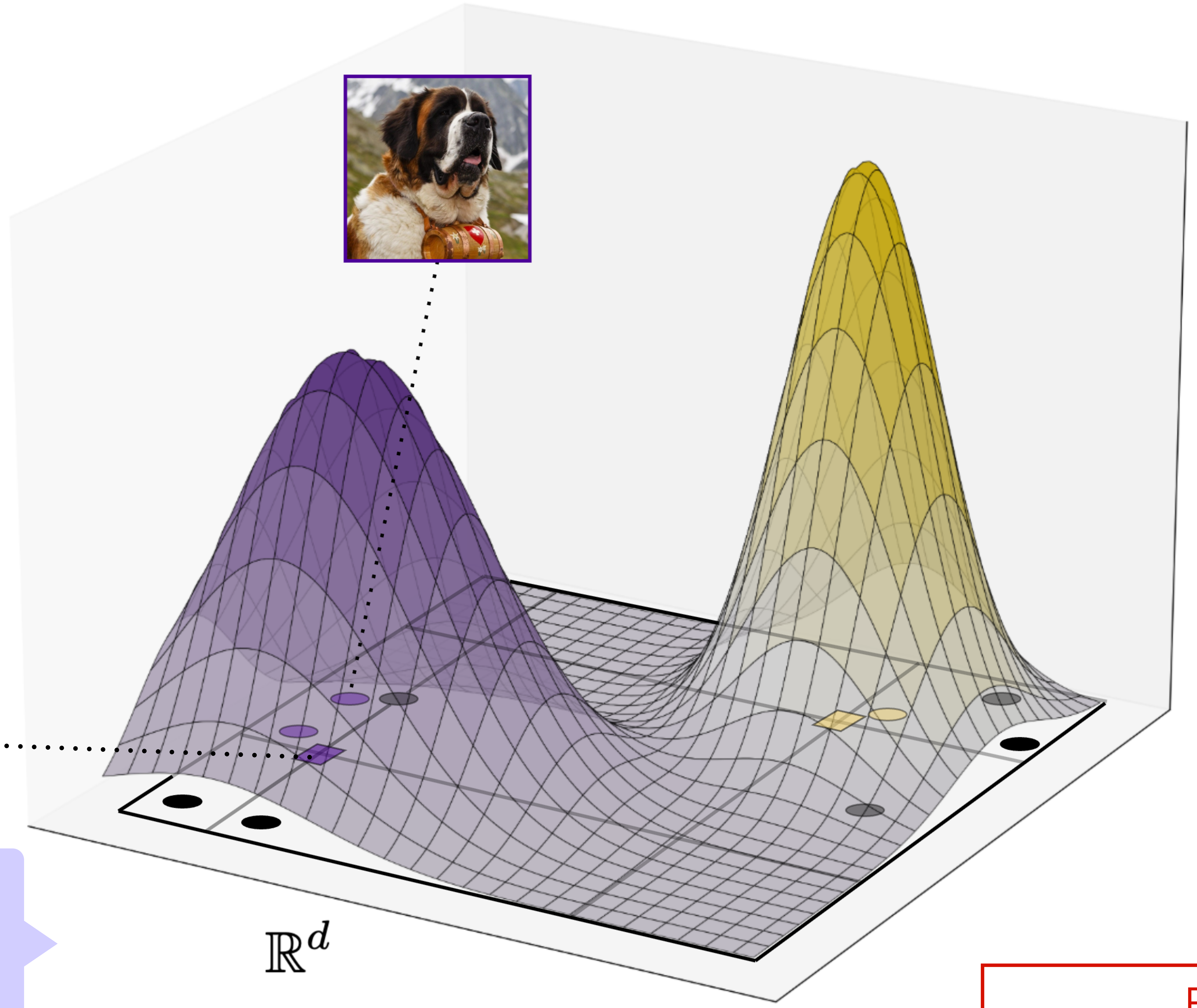
"st. bernard rescue near me"



Distribution of Text Embeddings

Prompt Bias

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



Unbiased
sample directly from $P_{Y,Z}$

true conditional mean w.r.t. distribution $P_{Y,Z}$

Distribution of Text Embeddings

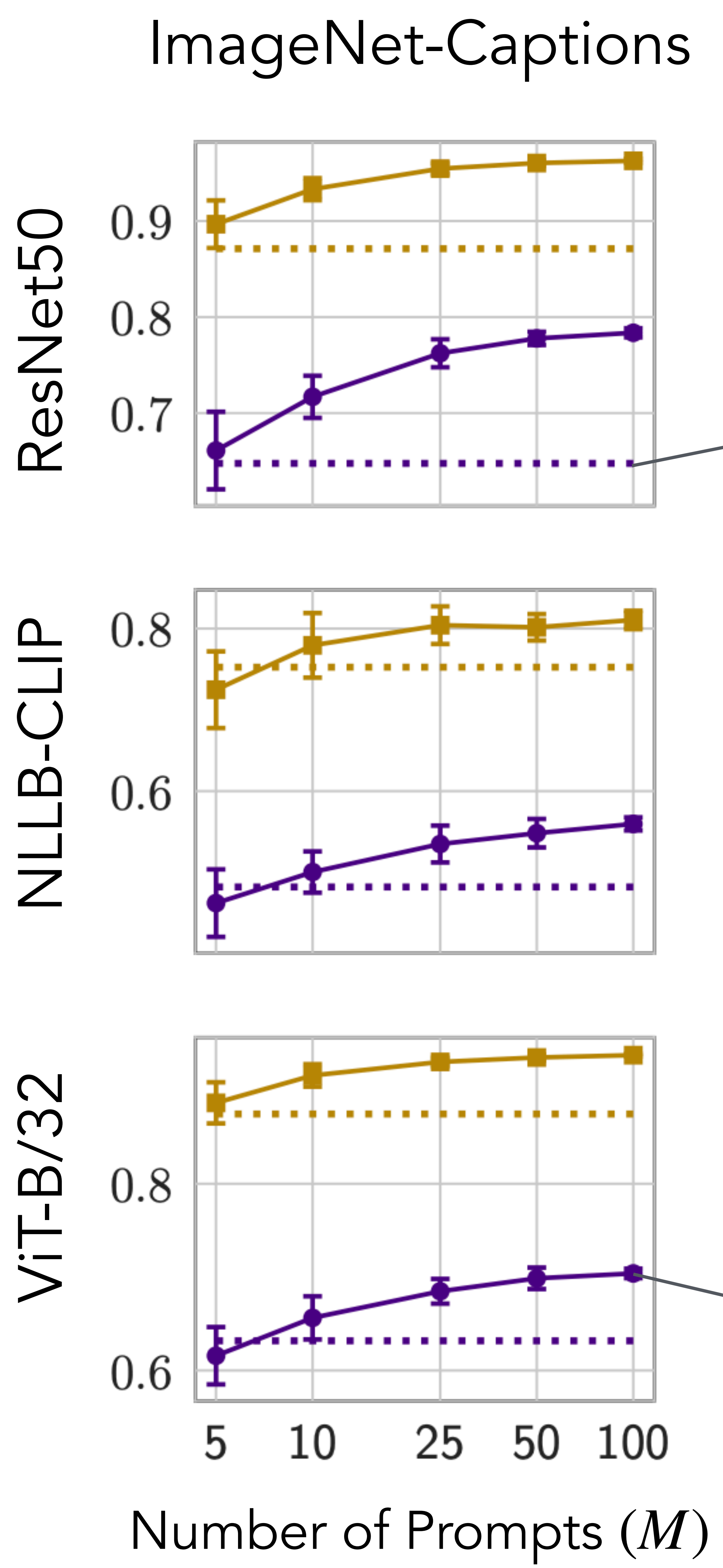
Prompt Bias

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

Zero-Shot Classification Accuracy

Top-1

Top-5




```
"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}.",
```

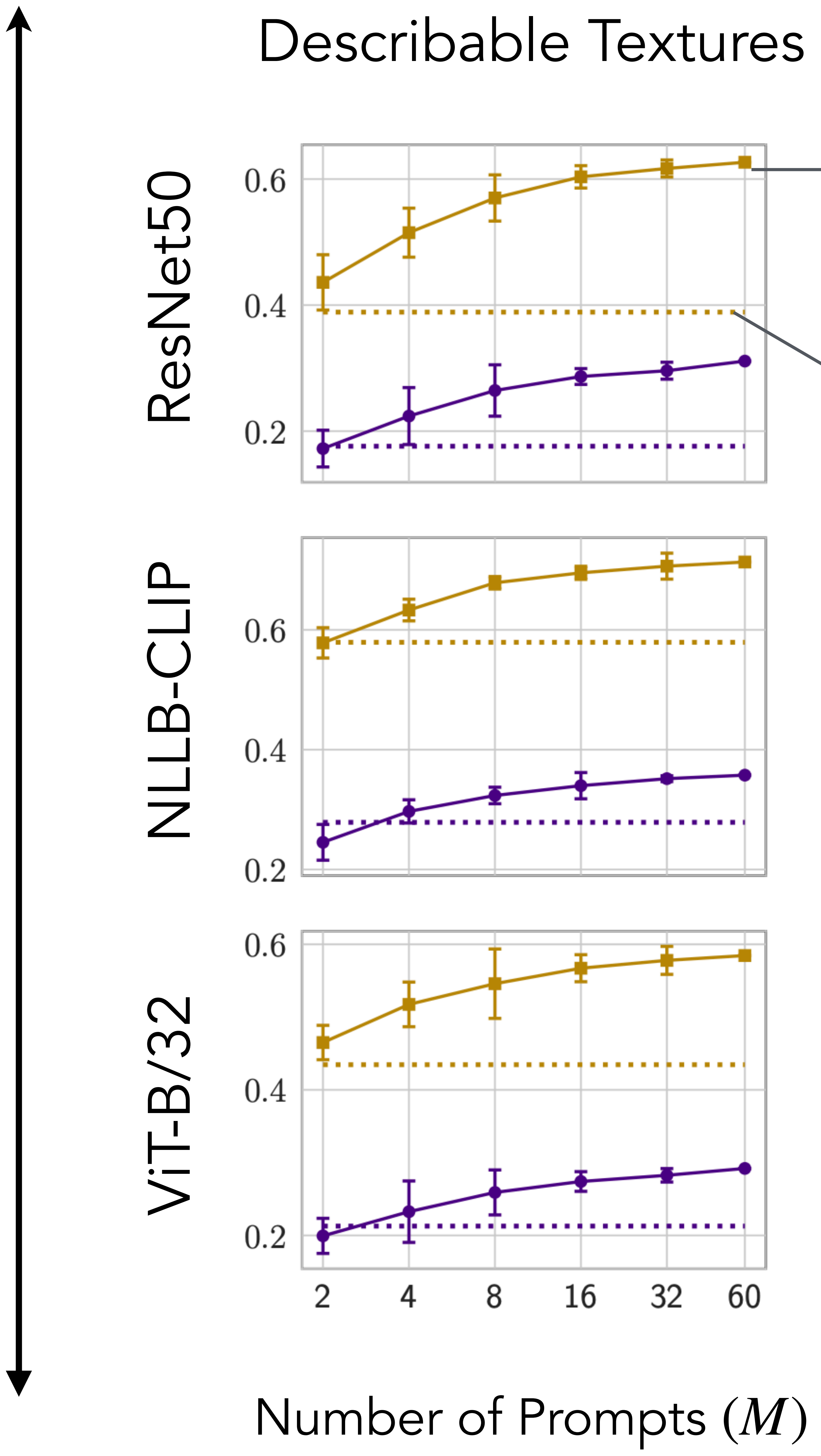
$$\frac{1}{M_{\boldsymbol{y}}} \sum_{i:\boldsymbol{y}_i=\boldsymbol{y}}$$

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

Zero-Shot Classification Accuracy

Top-1

Top-5



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)

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

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

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

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

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