A Unified Debiasing for Vision-Language Model across Modalities and Tasks

Hoin Jung, Taeuk Jang, Xiaoqian Wang

PURDUE UNIVERSITY

Elmore Family School of Electrical and Computer Engineering

Contents

Background

Versability of Visual-Language Model

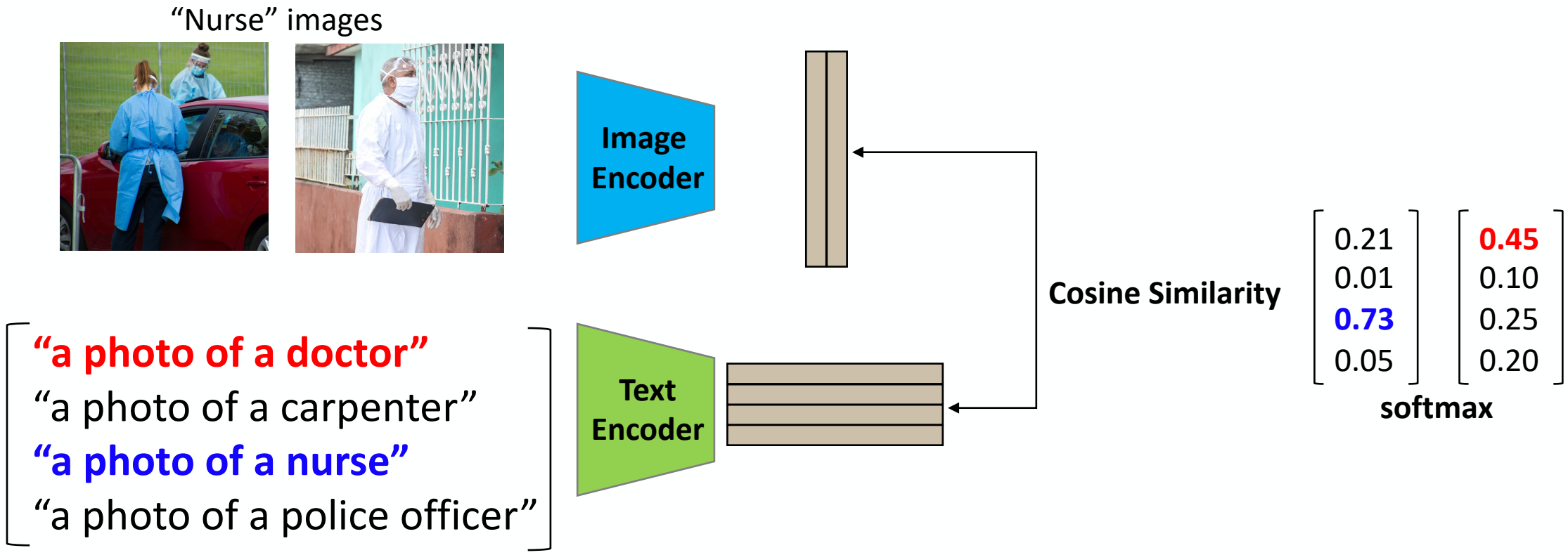
- Visual-Language Model (VLM) serve as foundation models for various downstream tasks
  - Zero-shot Classification
  - Text-to-Image Retrieval
  - Image Captioning
  - Text-to-Image Generation

- However, VLMs often skewing the model outputs in ways that reflect **societal stereotypes** such as gender or racial biases in assigning professions or describing scenarios.

Background

Bias in Zero-shot Classification

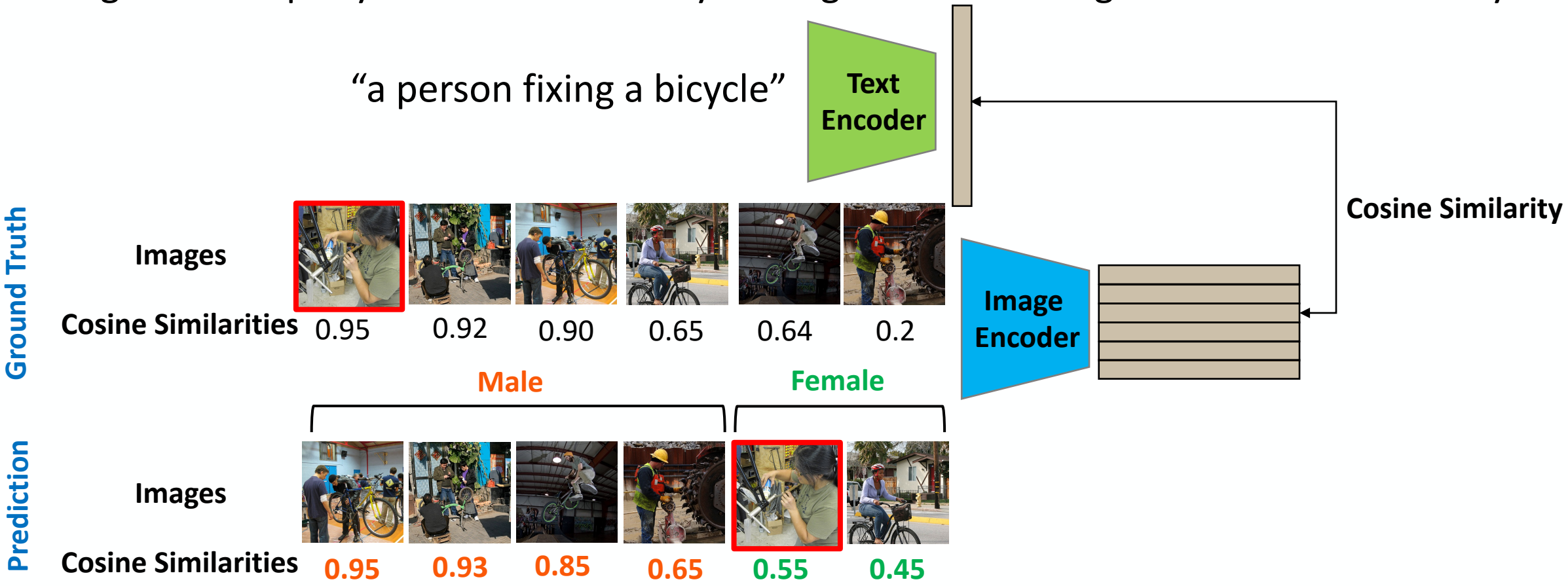
- Predicted class is determined by the highest cosine similarity between image and text embeddings.



Background

Bias in Text-to-Image Retrieval

- Images in the query set are retrieved by sorting them according to the cosine similarity



Background

Bias in Image Captioning

- Image captioning model may produce wrong gender in caption.

CLIP-CAP
A **woman** in a wetsuit surfing on a wave.

CLIP-CAP
A **man** riding skis down a snow covered slope.

Background

Bias in Text-to-Image Generation

- Text-to-Image generation model could be biased by sampling preferring certain gender for a profession.

Prompt: “a photo a person who works as a nurse.”

Prompt: “a photo a person who works as a plumber.”

Background

Bias in Text-to-Image Generation

- Even though we specify the gender, there's still a bias.

Prompt: "a photo a man who works as a nurse."

Prompt: "a photo a woman who works as a builder."

Motivation

Needs for A Unified and Efficient Debiasing Strategy

- Debiasing method often can deal with only a specific downstream tasks, and cannot be applied to others. **(Task-Specific)**

  $\Longrightarrow$ Needs for a unified debiasing strategy for various types of VLM and tasks. **(Task-Agnostic)**

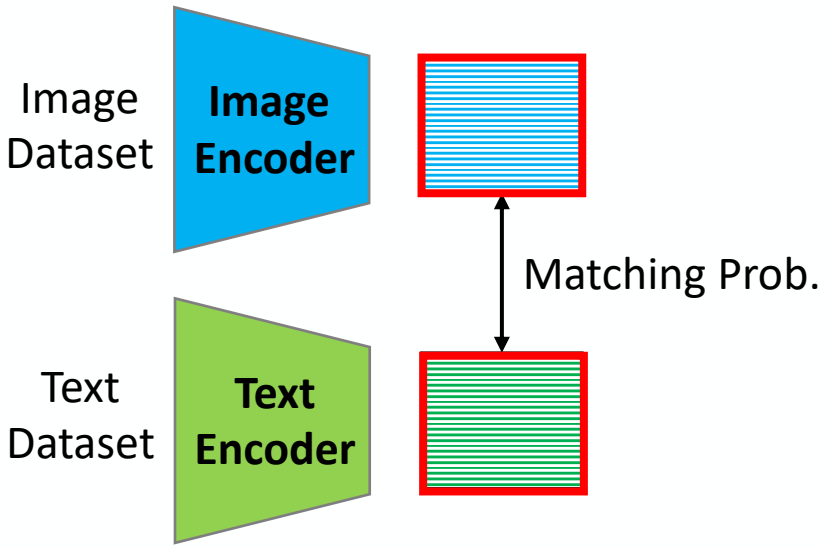
- Moreover, re-training the entire foundational model / VLMs is computationally expensive.

  $\Longrightarrow$ Needs for a **cost-efficient** debiasing approach.

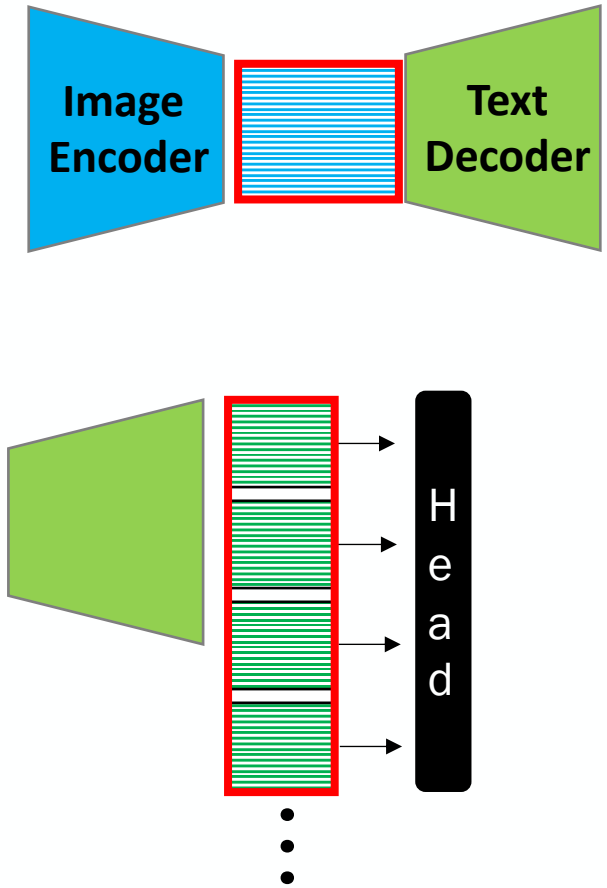
Motivation

A Unified Debiasing Strategy – Debiasing Embedding

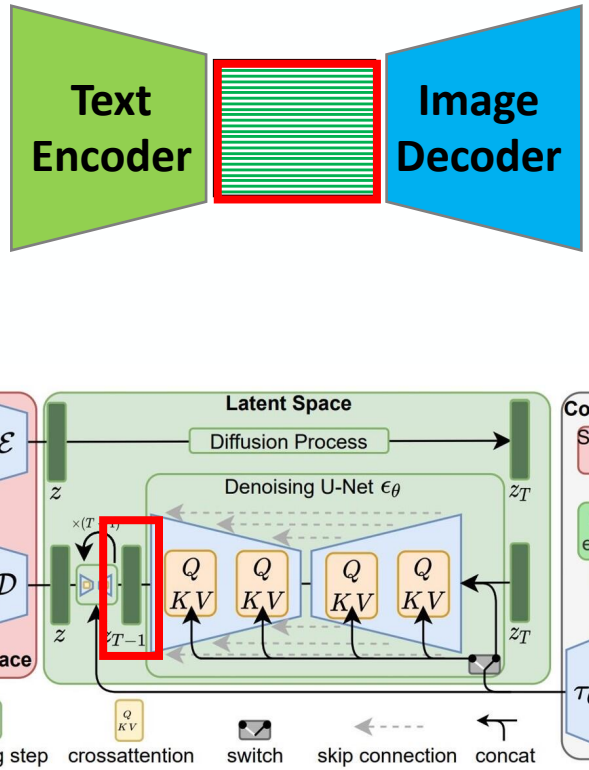
- Zero-shot Classification & Text-to-Image retrieval



- Image Captioning

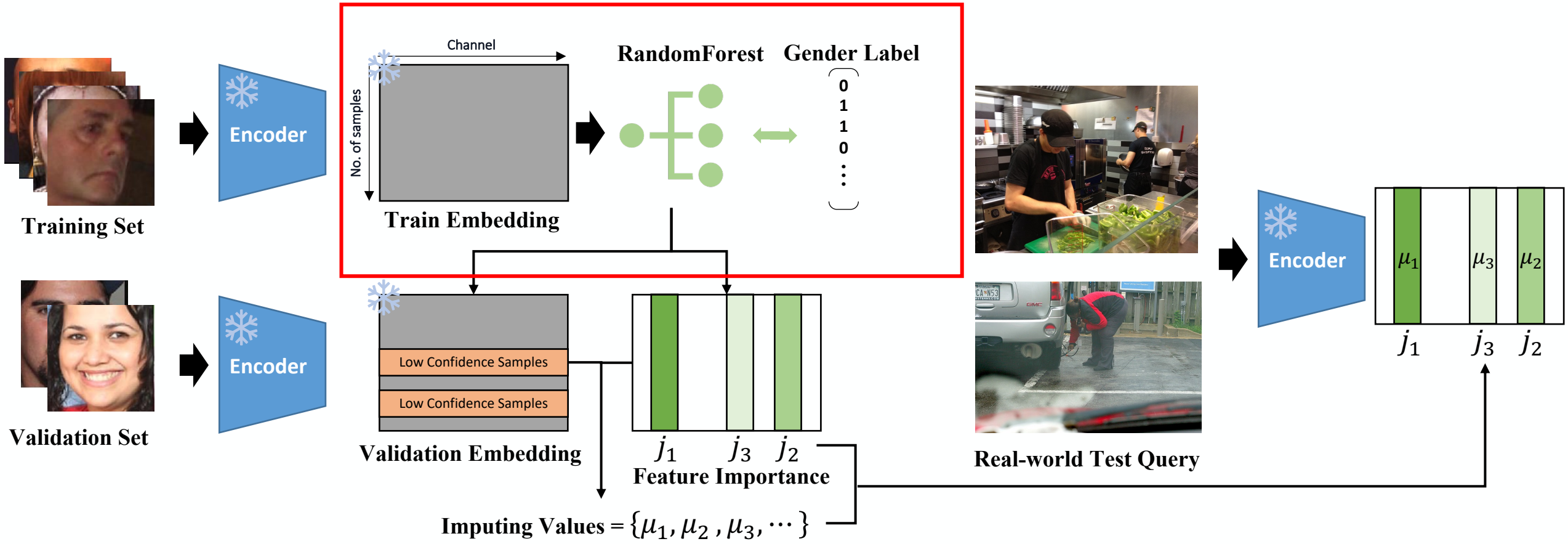


- Text-to-Image Generation



Proposed Method

Selective Feature Imputation for Debiasing (SFID)

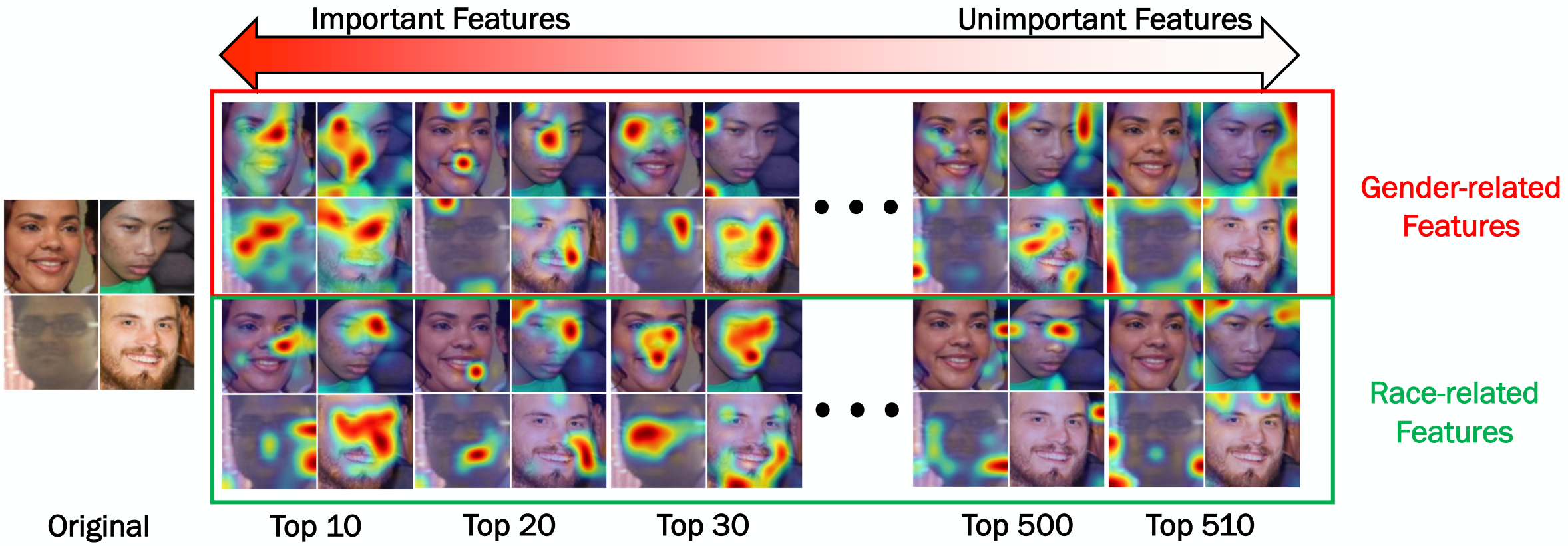


(a) Feature Selection & Imputing Value Extraction

(b) Debiasing Downstream Tasks

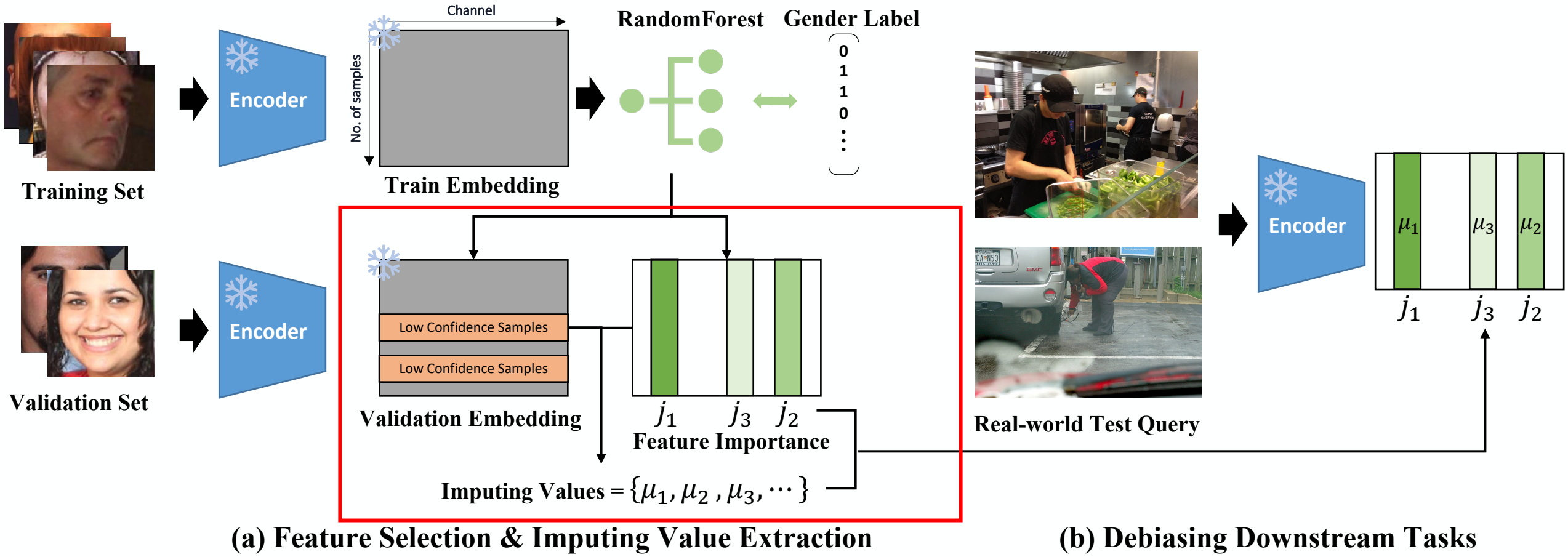
Proposed Method

Selective Feature Imputation for Debiasing (SFID)



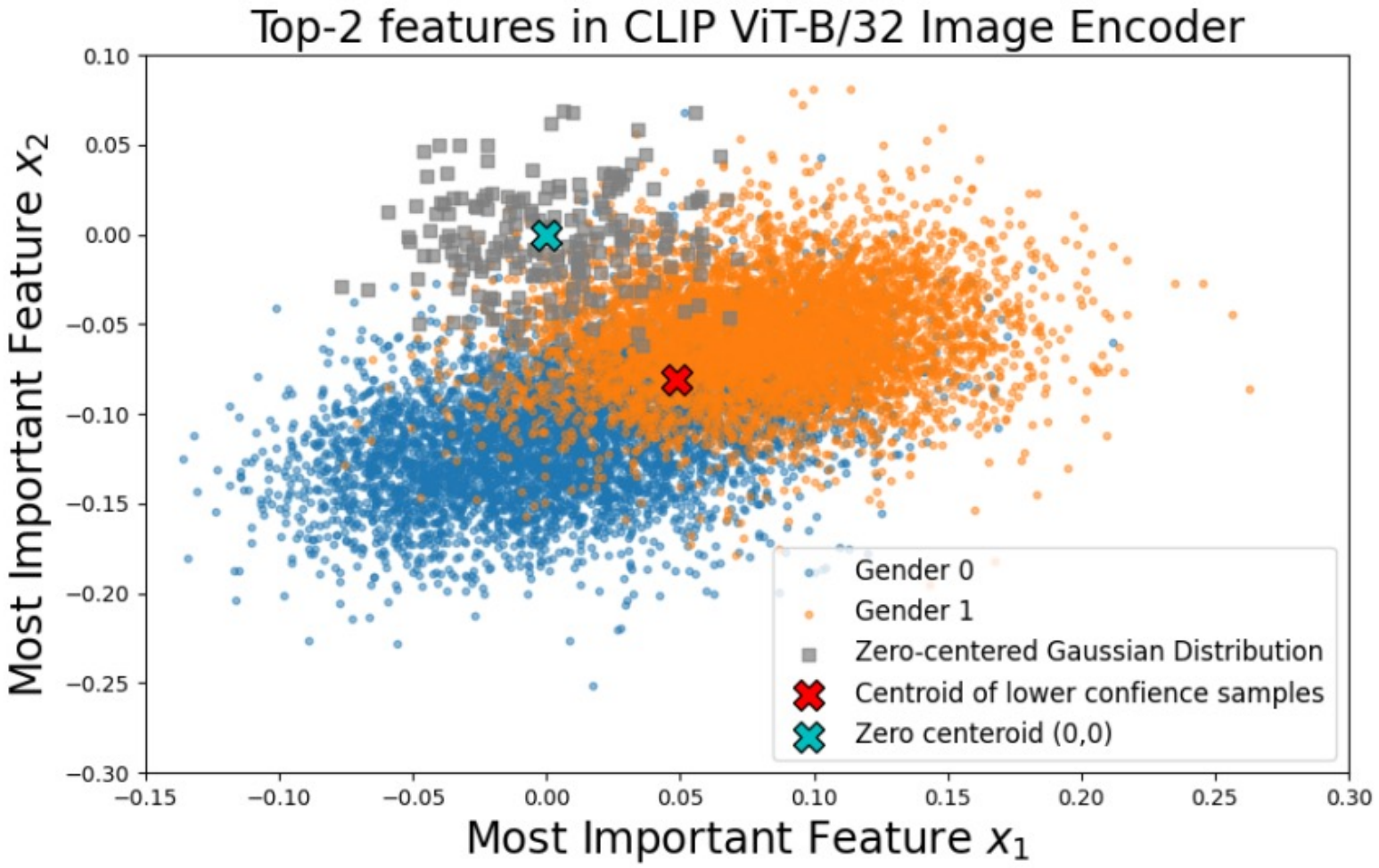
Proposed Method

Selective Feature Imputation for Debiasing (SFID)



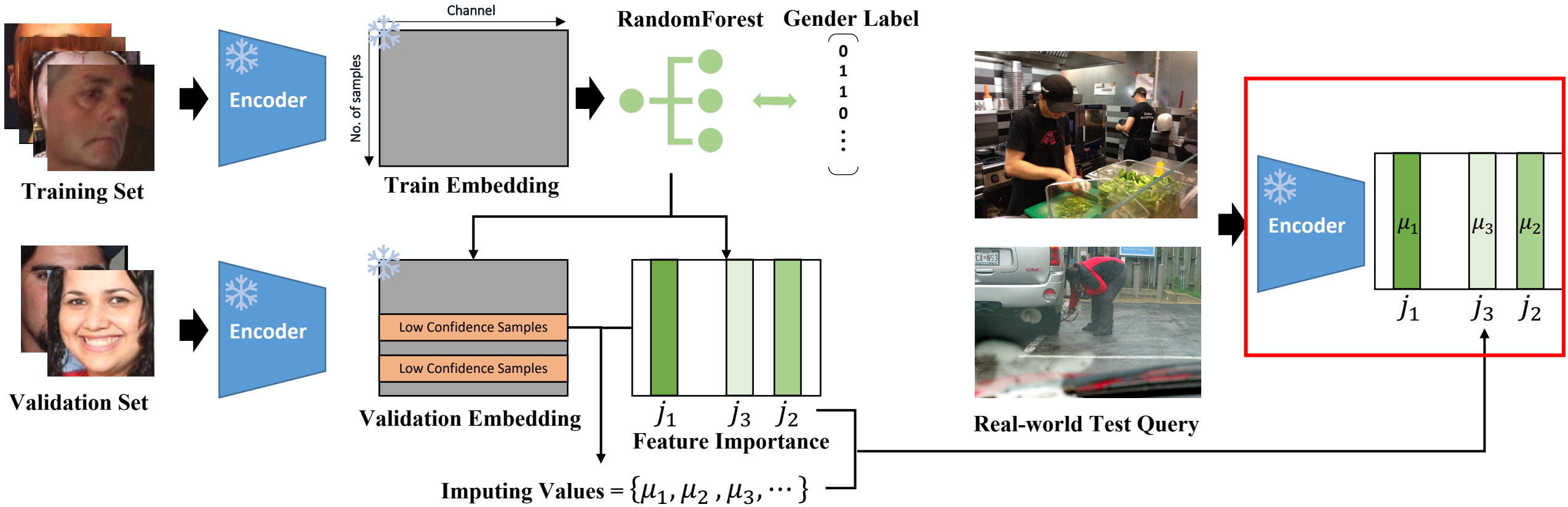
Proposed Method

Selective Feature Imputation for Debiasing (SFID)



Proposed Method

Selective Feature Imputation for Debiasing (SFID)



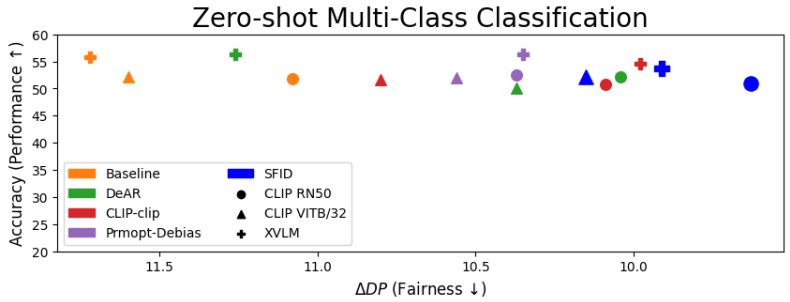
(a) Feature Selection & Imputing Value Extraction

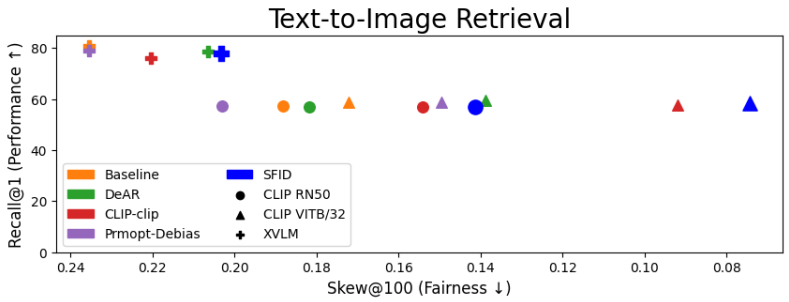
(b) Debiasing Downstream Tasks

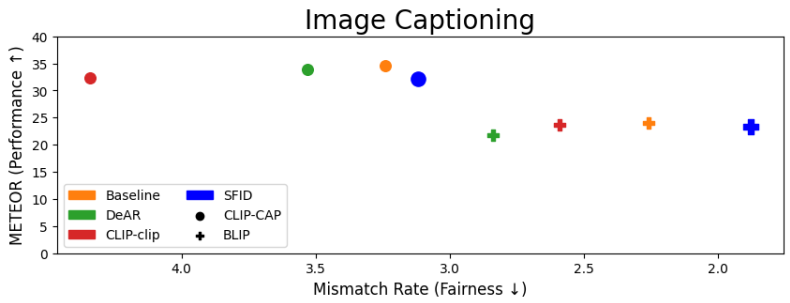
Result

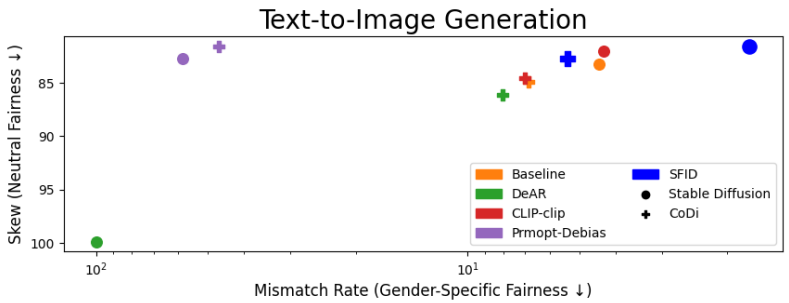
Selective Feature Imputation for Debiasing (SFID)

- Effective in debiasing.
- Can be used any types of tasks.
- Not requiring training a model.









Thank You

Hoin Jung
jung414@purdue.edu