Vision-Language Joint Debiasing with Modality Alignment

FYP CA Presentation

Zhang Haoyu (A0220591M)

Backgrounds

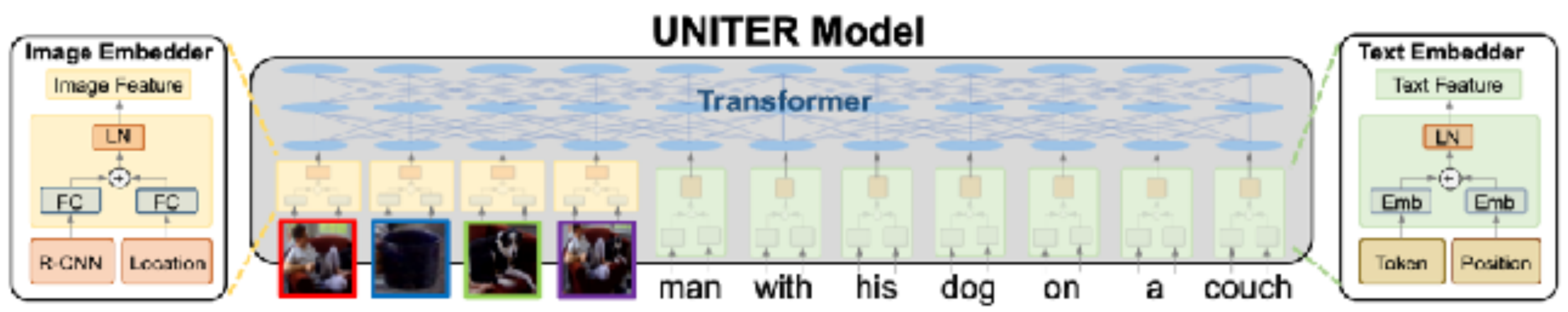
Vision-Language Pre-trained Model (VL-PTM)

- Single-stream:
  - UNITER, ViLT, etc.
- Dual-stream:
  - CLIP, ALIGN, ALBEF, BLIP, etc.

Backgrounds

Vision-Language Pre-trained Model (VL-PTM)

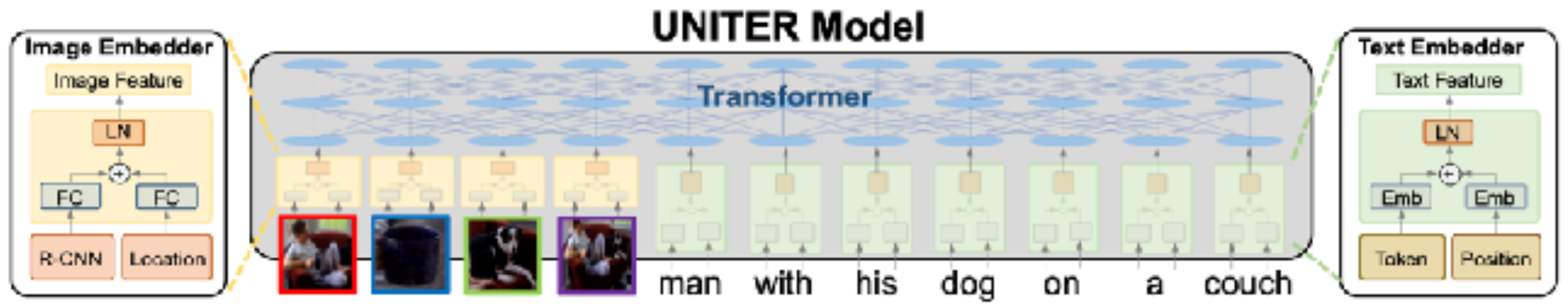
- Single-stream:
    - UNITER, ViLT, etc.
- Dual-stream:
    - CLIP, ALIGN, ALBEF, BLIP, etc.

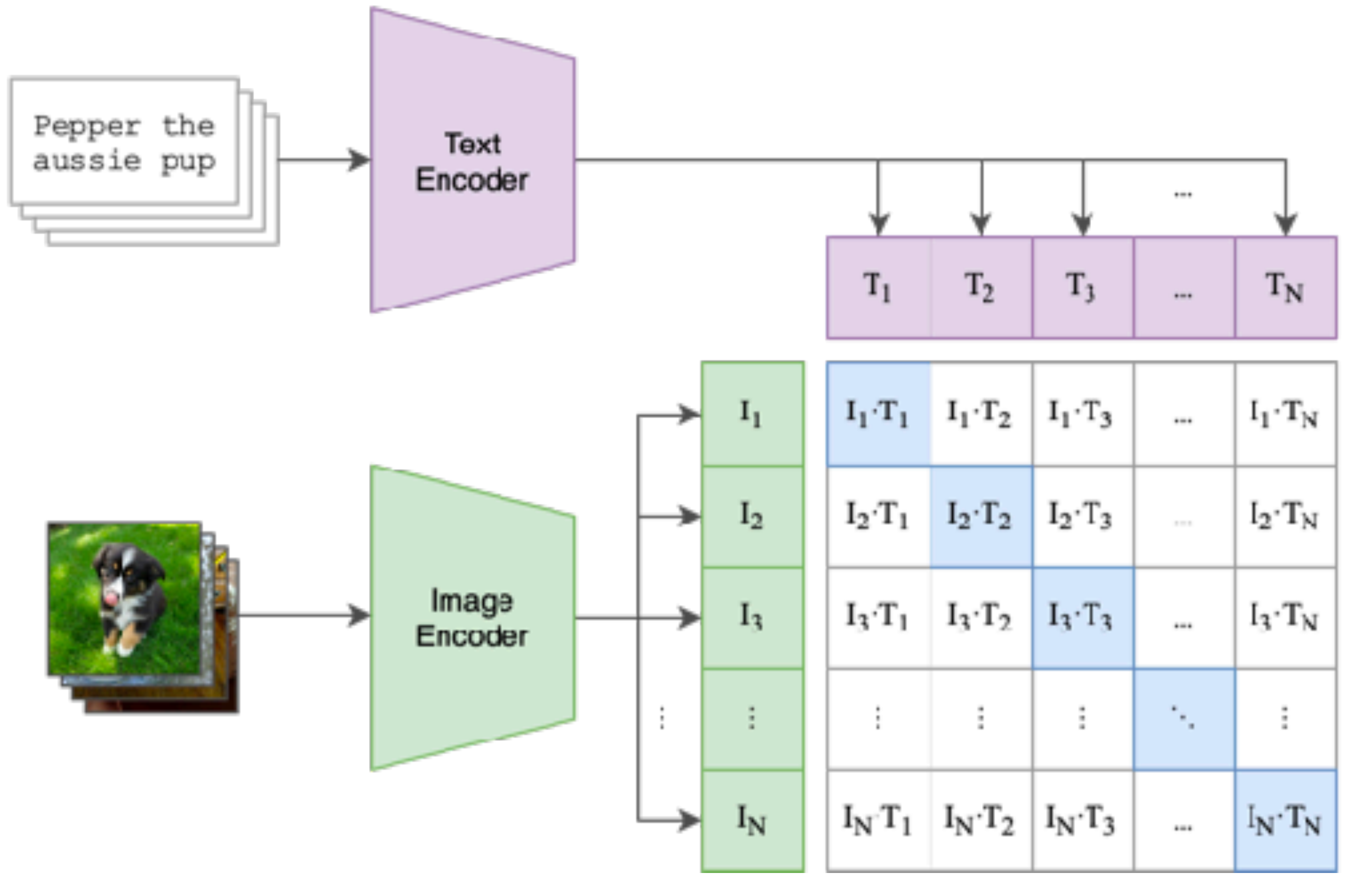


Backgrounds

Vision-Language Pre-trained Model (VL-PTM)

- Single-stream:
  - UNITER, ViLT, etc.
- Dual-stream:
  - CLIP, ALIGN, ALBEF, BLIP, etc.

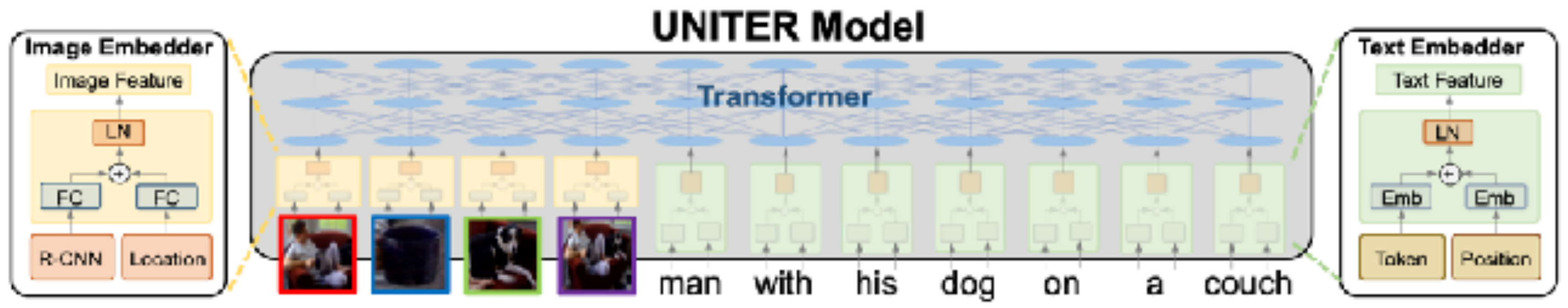


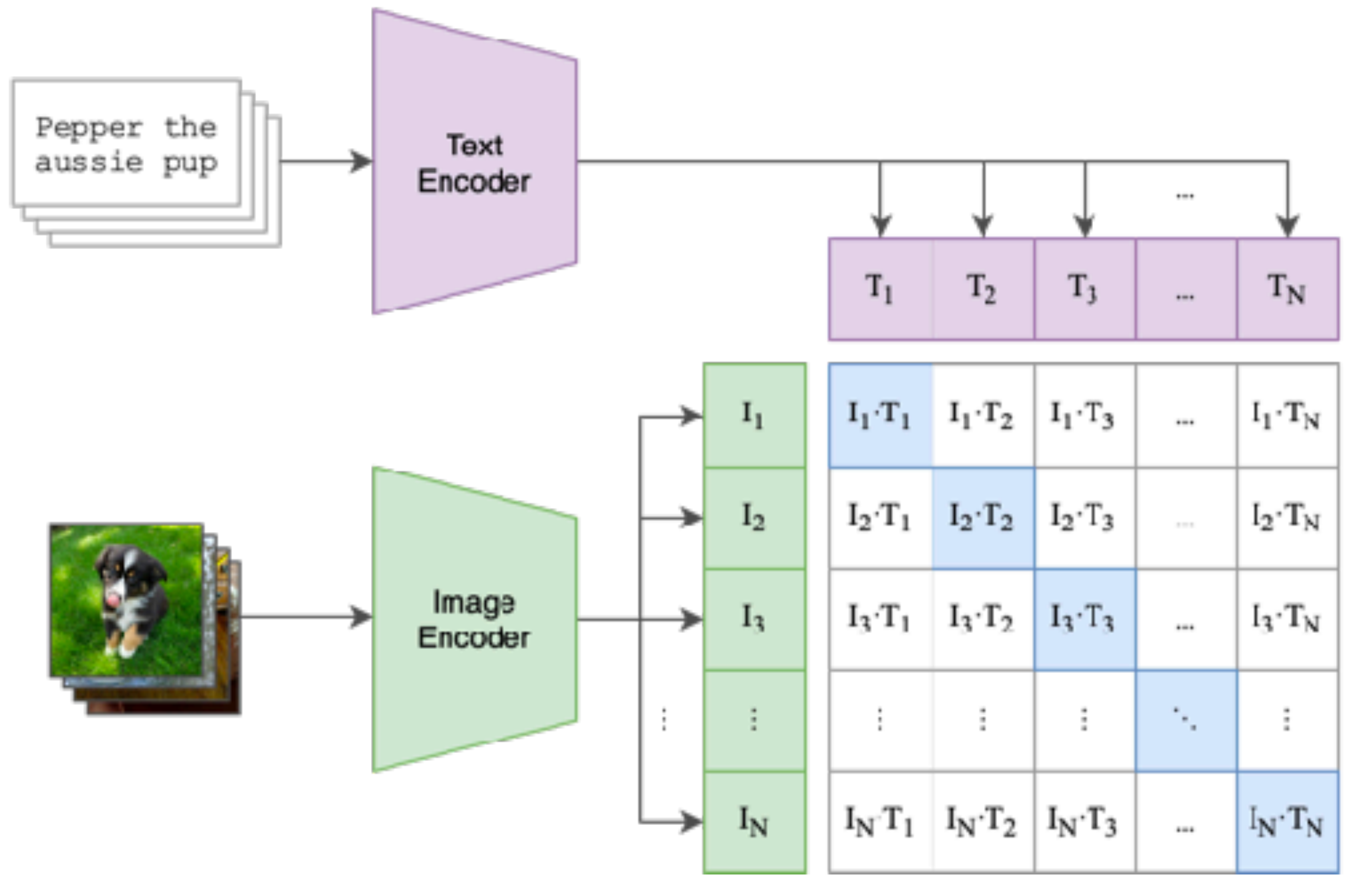


Backgrounds

Vision-Language Pre-trained Model (VL-PTM)

- Single-stream:
  - UNITER, ViLT, etc.
- Dual-stream:
  - CLIP, ALIGN, ALBEF, BLIP, etc.
- Vision-Language (V-L) tasks
  - Understanding: Image-text retrieval, Visual Question Answering, etc.
  - Generation: Image captioning, text-to-image (Stable Diffusion), etc.





Backgrounds

Social Bias in VL-PTMs

- Learning spurious correlations in images/texts during training
- Associates certain **concepts** with **groups with specific attributes** (race, gender, age, etc.)
- Manifest in V-L tasks
    - Understanding:
        - Image-text retrieval, etc.
    - Generation:
        - Text-to-image (Stable Diffusion, etc.)

Backgrounds

Social Bias in VL-PTMs

- Learning spurious correlations in images/texts during training
- Associates certain **concepts** with **groups with specific attributes** (race, gender, age, etc.)
- Manifest in V-L tasks
    - Understanding:
        - Image-text retrieval, etc.
    - Generation:
        - Text-to-image (Stable Diffusion, etc.)

Figure 3: **Effect of debiasing CLIP ViT-B/16 by ranked images with concept of "smart"** from the FairFace validation set, labeled with male and female.

Backgrounds

Social Bias in VL-PTMs

- Learning spurious correlations in images/texts during training
- Associates certain **concepts** with **groups with specific attributes** (race, gender, age, etc.)
- Manifest in V-L tasks
  - Understanding:
    - Image-text retrieval, etc.
  - Generation:
    - Text-to-image (Stable Diffusion, etc.)

Figure 3: **Effect of debiasing CLIP ViT-B/16 by ranked images with concept of "smart"** from the FairFace validation set, labeled with male and female.

Backgrounds

Social Bias in VL-PTMs

- **Profound social impact**
    - Biased decision making
        - Unfair allocation
    - Biased generated content
        - Reinforce existing social biases

Figure 3: **Effect of debiasing CLIP ViT-B/16 by ranked images with concept of "smart"** from the FairFace validation set, labeled with male and female.

Research Problem

- **VL-PTM debiasing**
  - Mitigate social biases in the pre-trained CLIP model, making it more fair
  - Maintain a balance between fairness and V-L performance.

Literature Review

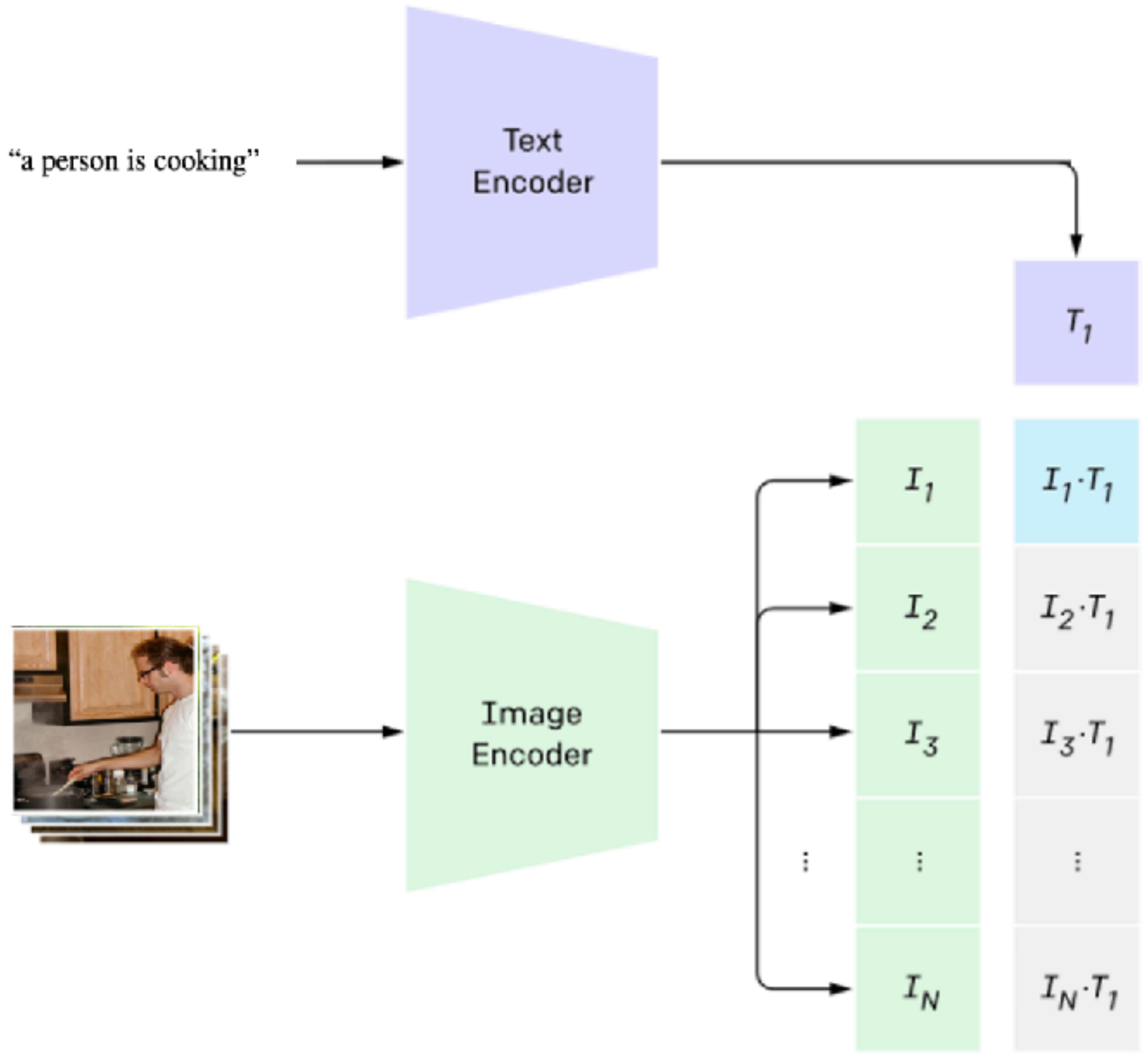
Measurement of Social Bias in CLIP

- **Bias measurement:**
  - Concepts (jobs, qualities, …)
  - Protected attributes (gender/race/age…)
  - Bias: model associate **concepts** with **protected attributes**

Literature Review

Measurement of Social Bias in CLIP

- **Retrieval-based bias metrics**
  - Concept: text prompt
  - Protected attributes: images
  - Text prompt as a query, the CLIP retrieves k images based on text-image similarities
  - The retrieved k images have different attributes (gender/age/race…)



Literature Review

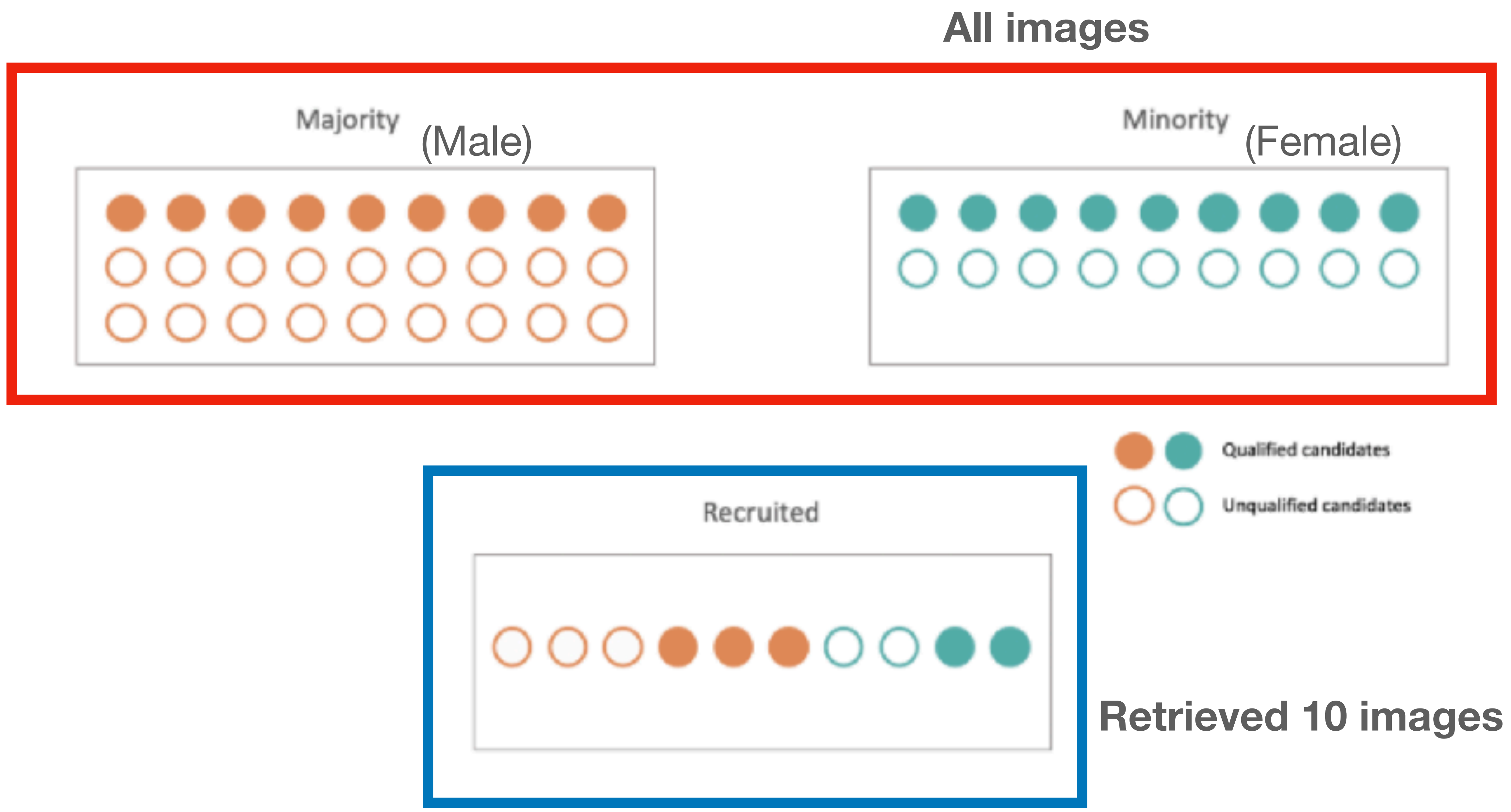
Measurement of Social Bias in CLIP

- Retrieval-based bias metrics
  - Distribution of the protected attributes in the k images retrieved

Literature Review

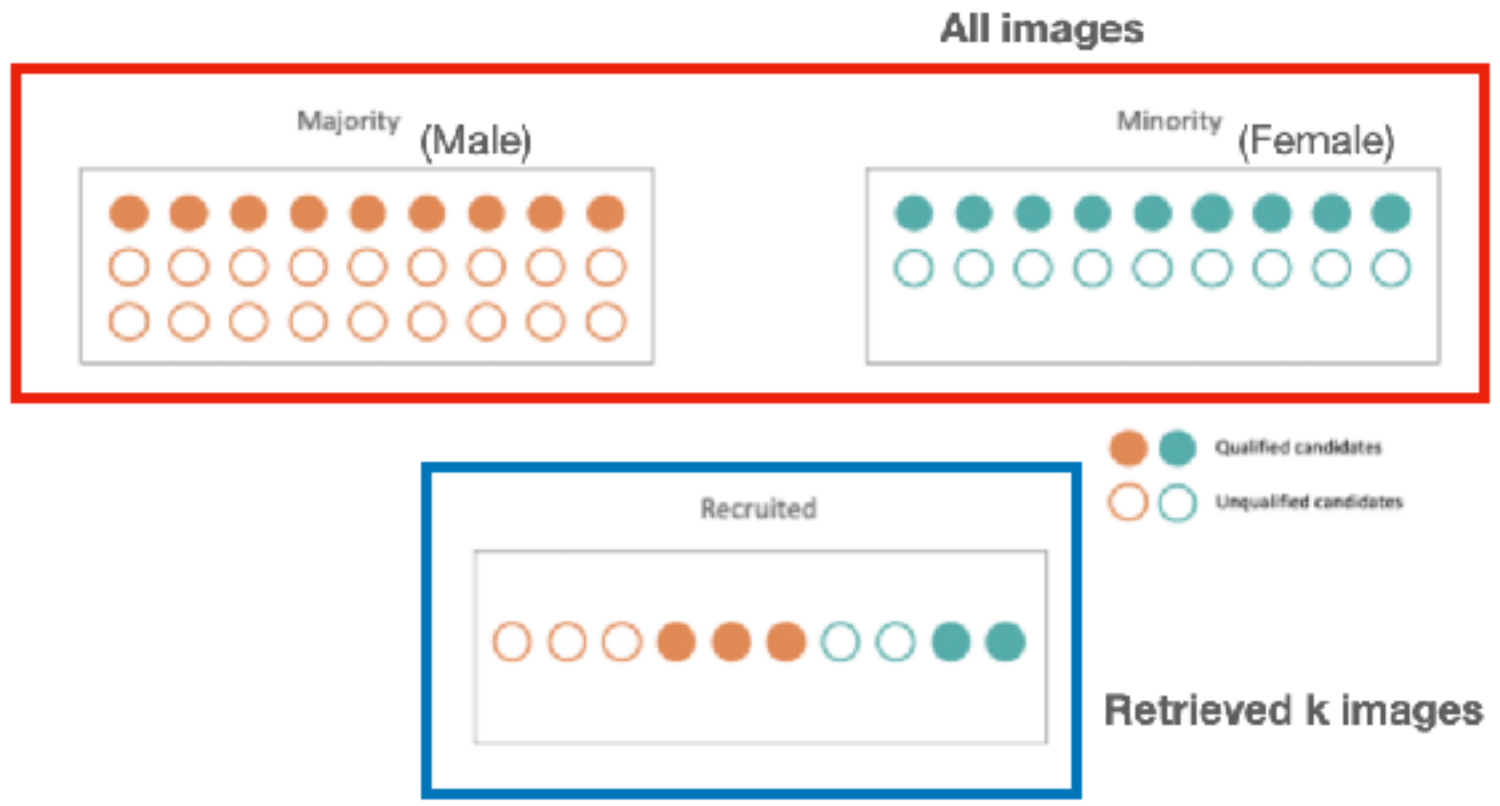
Measurement of Social Bias in CLIP

- **Retrieval-based bias metrics**
  - **Distribution of the protected attributes in the k images retrieved**
  - Ideally, **proportion of an attribute in retrieved images** = **proportion of an attribute in the original pool of images**
  - **Demographic parity:** the **retrieval based on matching** is **independent** of the **protected attributes** of the image



Literature Review

Measurement of Social Bias in CLIP



- **Retrieval-based bias metrics**
  - Bias measured by comparing **the new distribution of the protected attributes among the k retrieved images** with **the original distribution of the protected attributes among all images**
  - **Metric 1: Max Skew**

$$MaxSkew_{@}k(\tau_T) = \max_{A_i \in A} Skew_{A_i}@k(\tau_T) \quad \left(Skew_A@k(\tau_T) = \ln \frac{p_{\tau_T,T,A}}{p_{d,T,A}} \right)$$

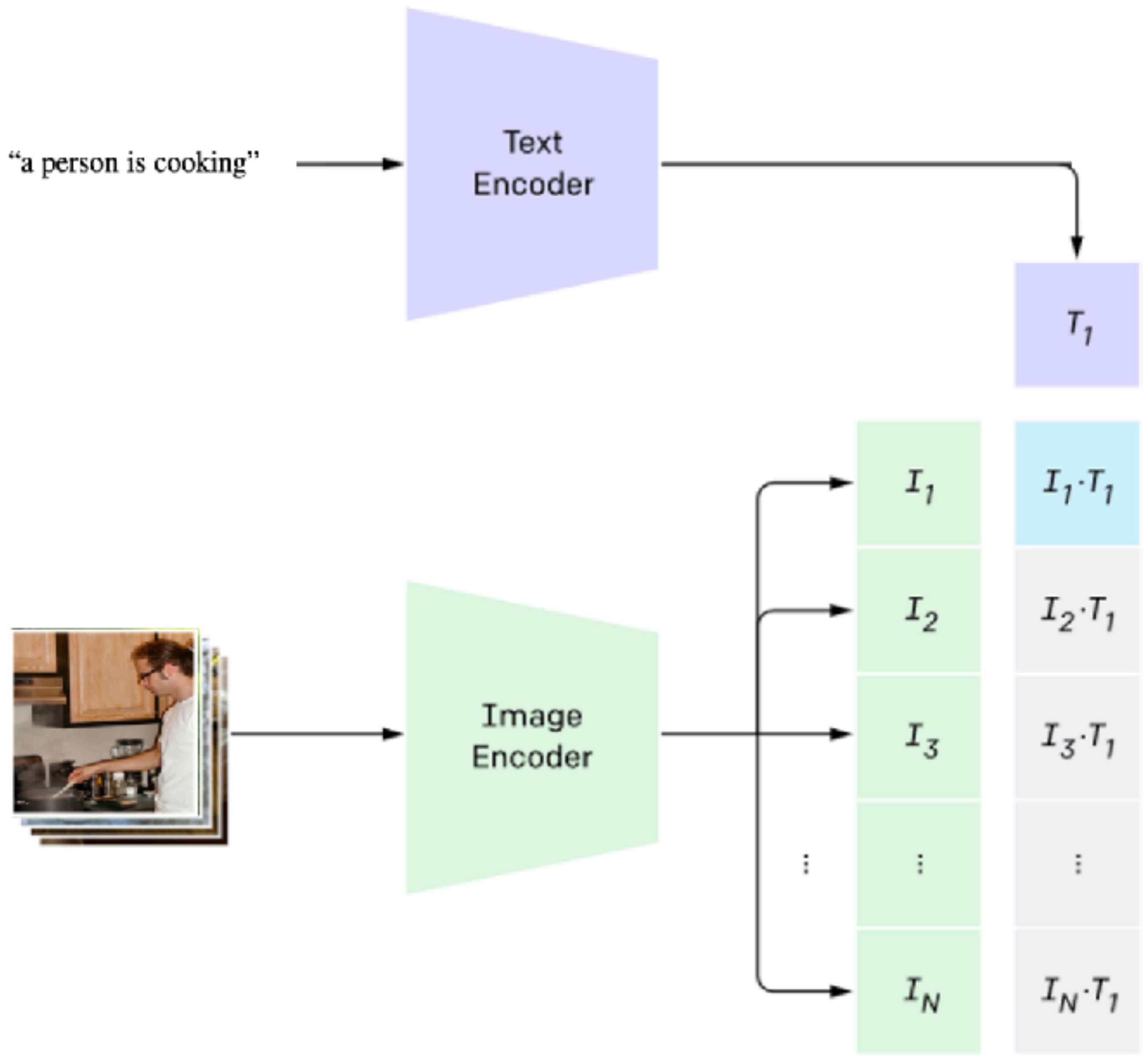
  - **Metric 2: Normalized Discounted Cumulative KL-Divergence (NDKL)**

$$NDKL(\tau_T) = \frac{1}{Z} \sum_{i=1}^{|\tau_y|} \frac{1}{\log_2(i+1)} d_{KL}(D_{\tau_T^i} || D_T)$$

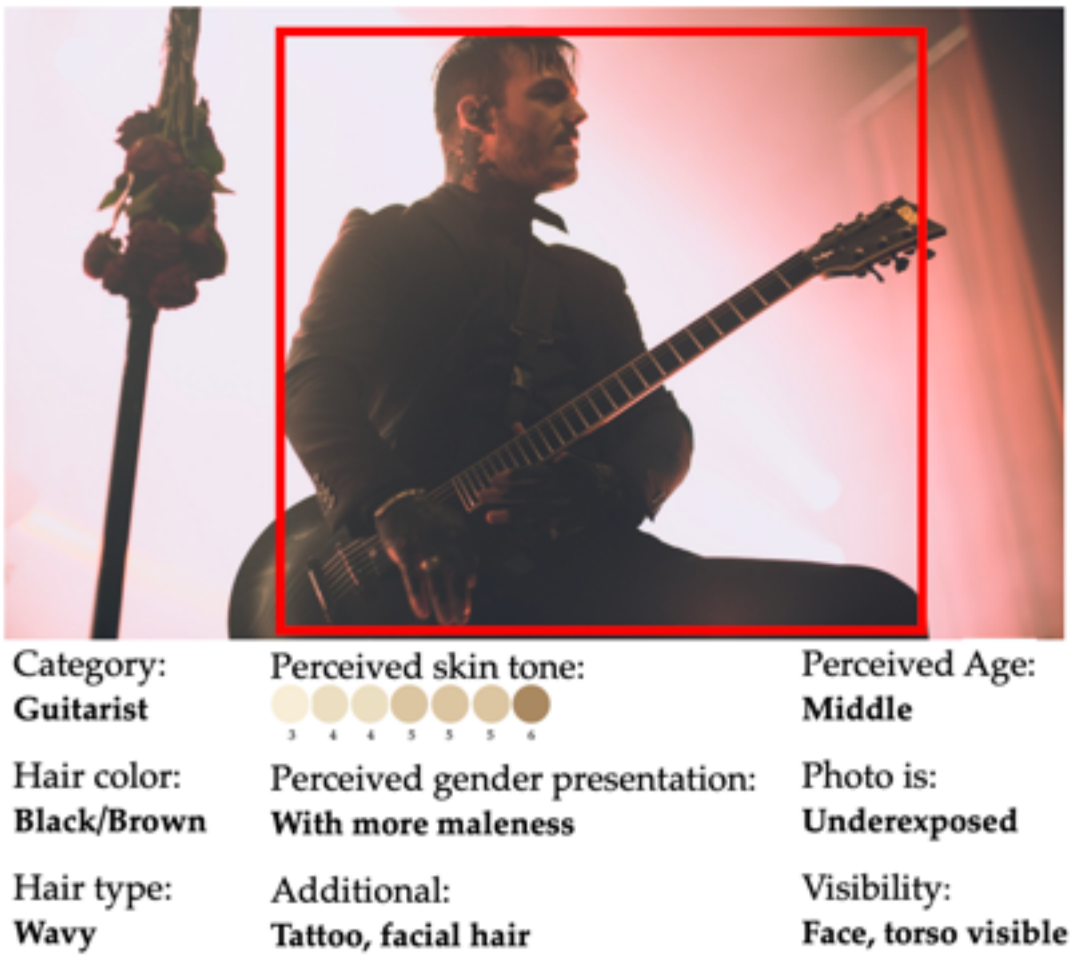
Literature Review

Measurement of Social Bias in CLIP

- **Fairness datasets**
  - FairFace, UTKFace, FACET
  - Labels: Race, gender, age, etc.



(a) FairFace (b) UTKFace



Literature Review

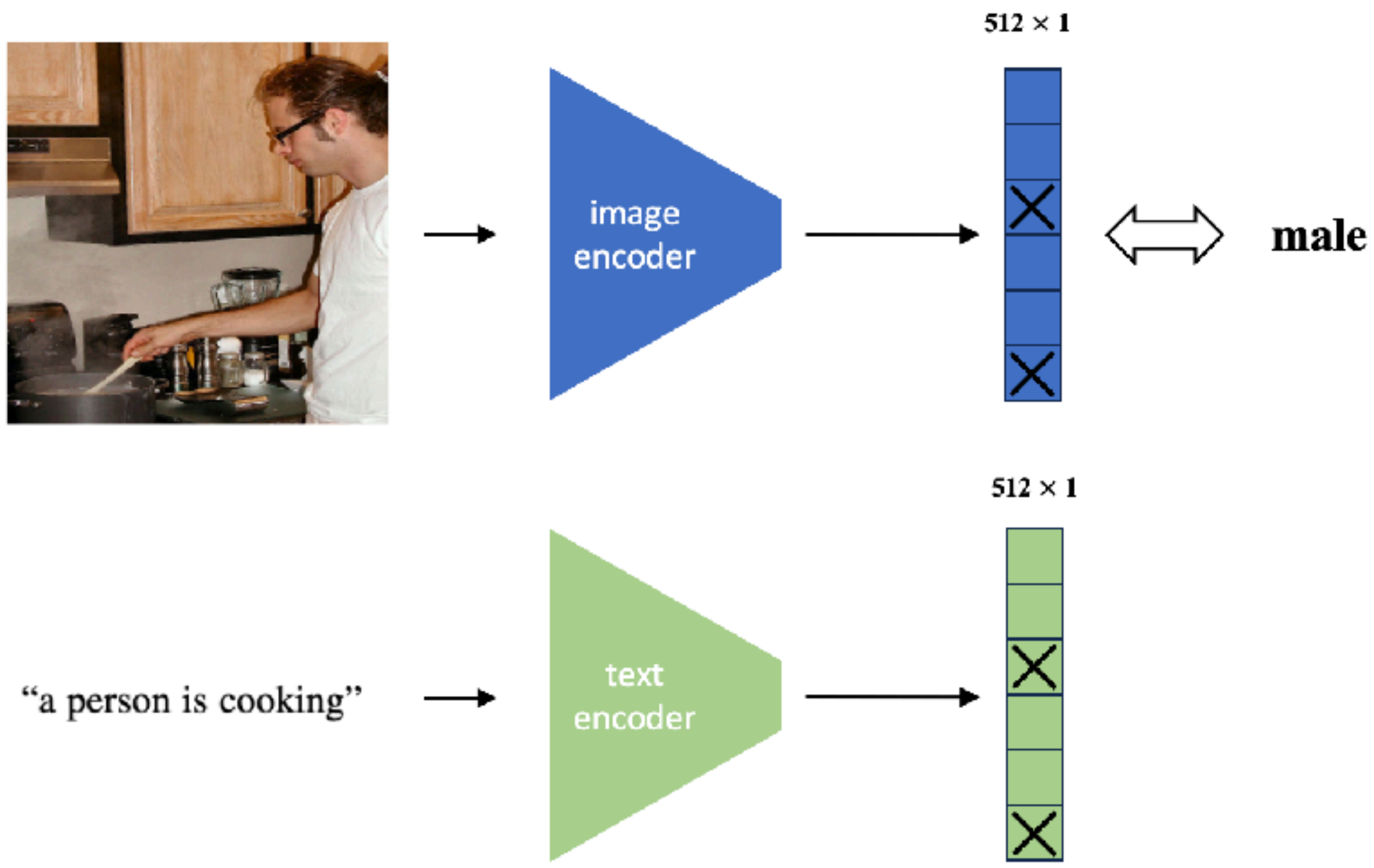
CLIP Debiasing Approaches

- Training-free: **Embedding vector manipulation**
- Training-required: **Fair module fine-tuning**

Literature Review

Embedding Vector Manipulation - Feature Clipping

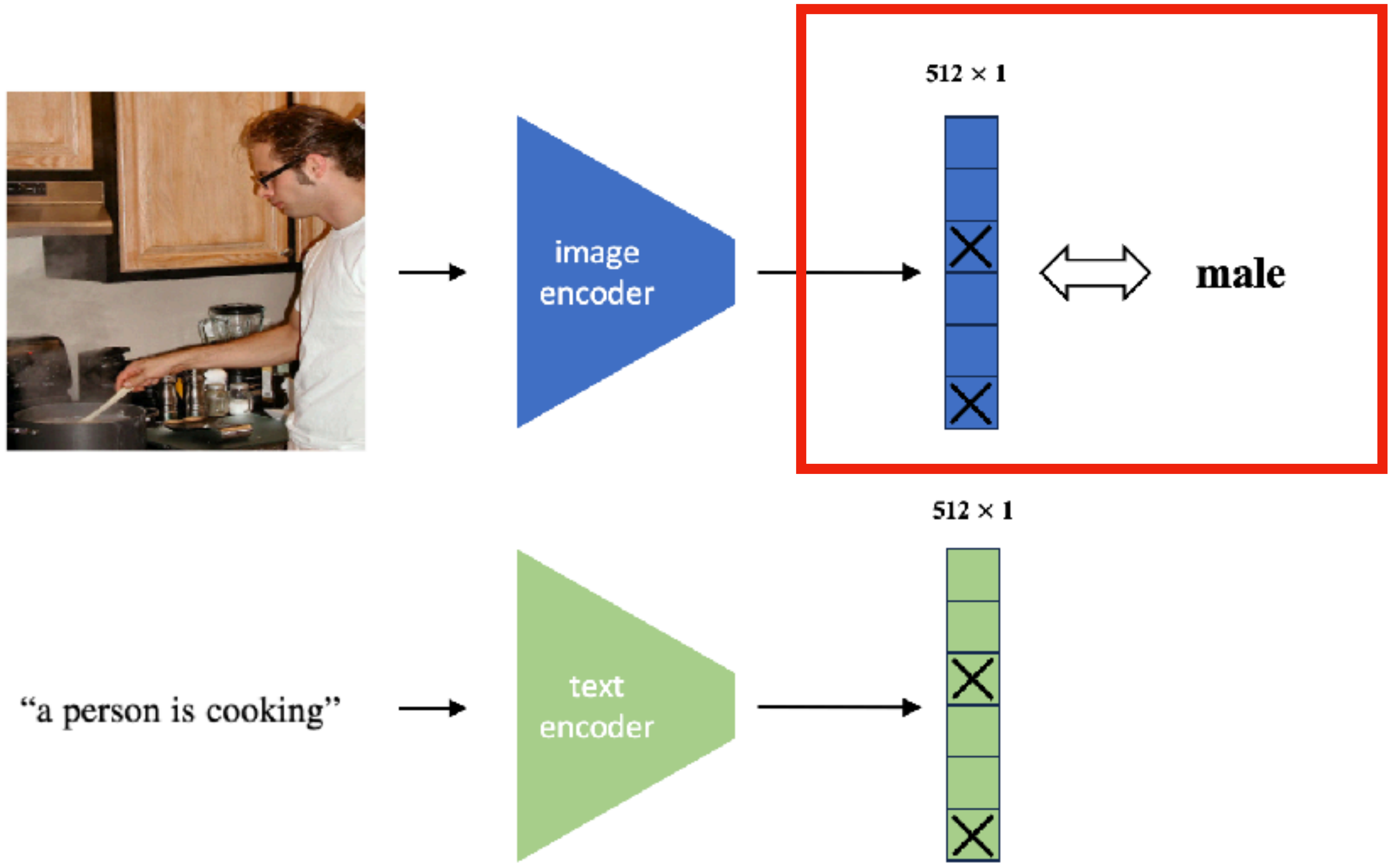
- Debias both image and text embeddings
- Determine the features in the image embedding vectors that contain the most bias information
- Remove those image features from the image embedding vector
- Remove the corresponding text features from the text embedding vector



Literature Review

Embedding Vector Manipulation - Feature Clipping

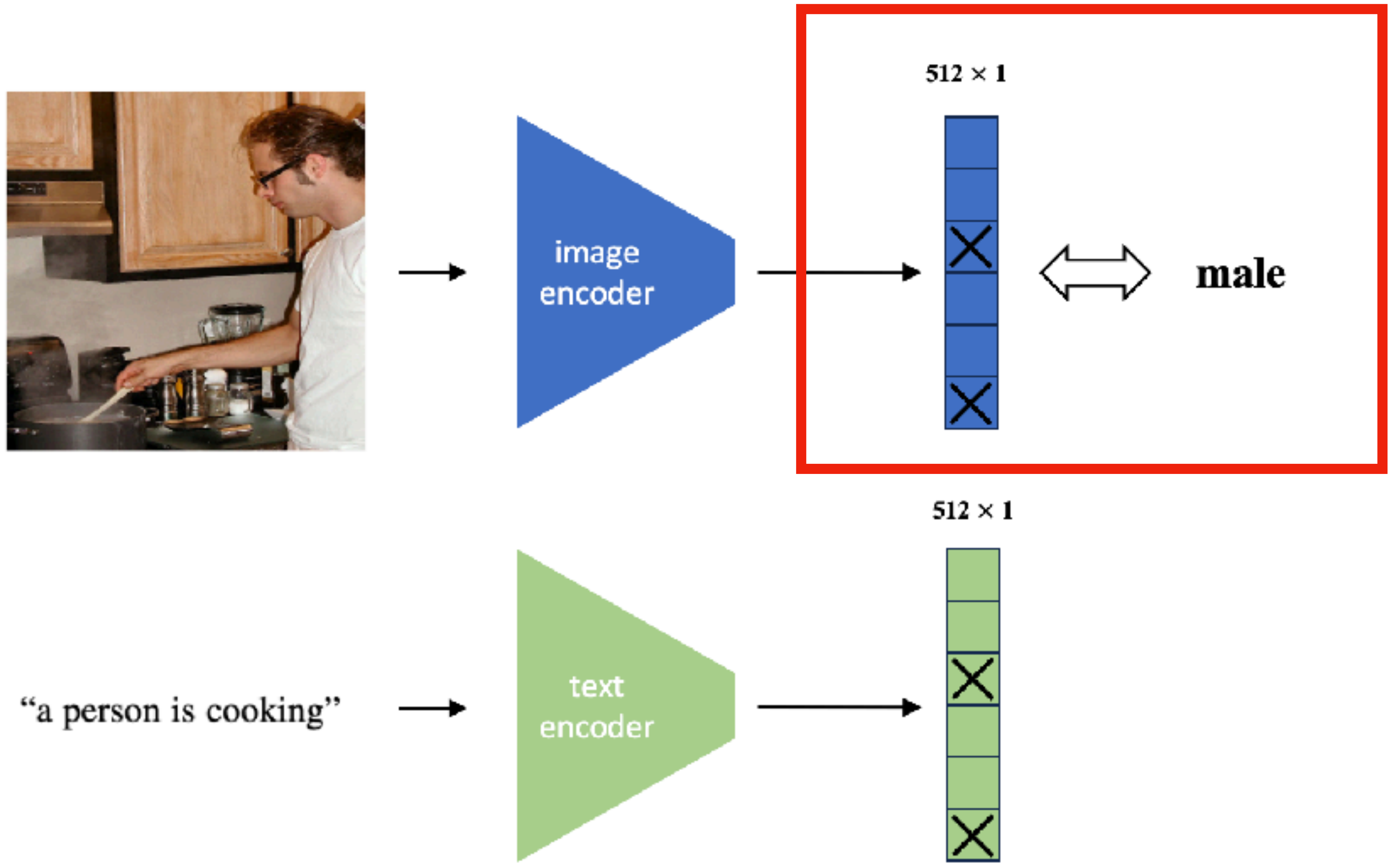
- Debias both image and text embeddings
- Determine the features in the image embedding vectors that contain the most bias information
- Remove those image features from the image embedding vector
- Remove the corresponding text features from the text embedding vector



Literature Review

Embedding Vector Manipulation - Feature Clipping

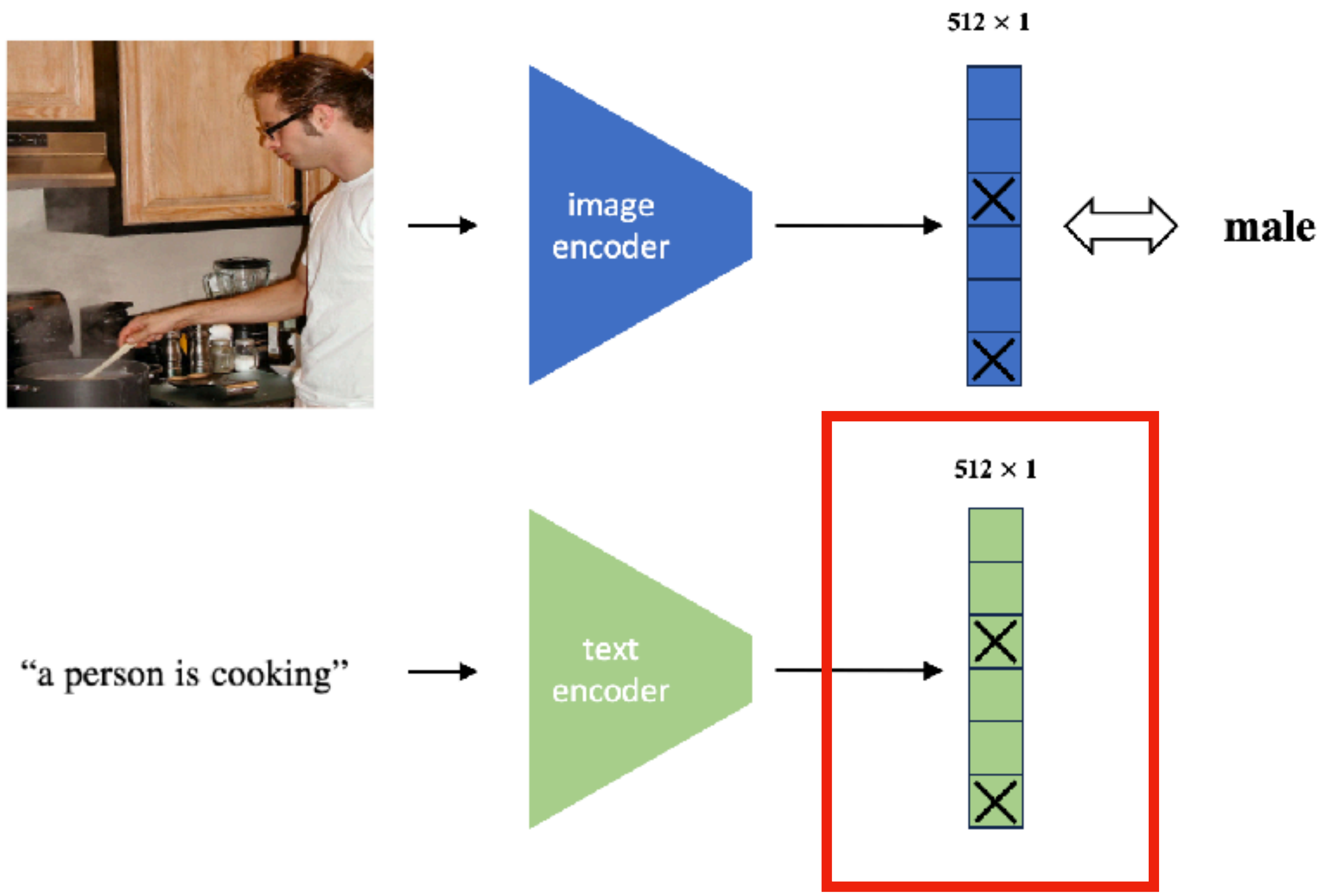
- Debias both image and text embeddings
- Determine the features in the image embedding vectors that contain the most bias information
- Remove those image features from the image embedding vector
- Remove the corresponding text features from the text embedding vector



Literature Review

Embedding Vector Manipulation - Feature Clipping

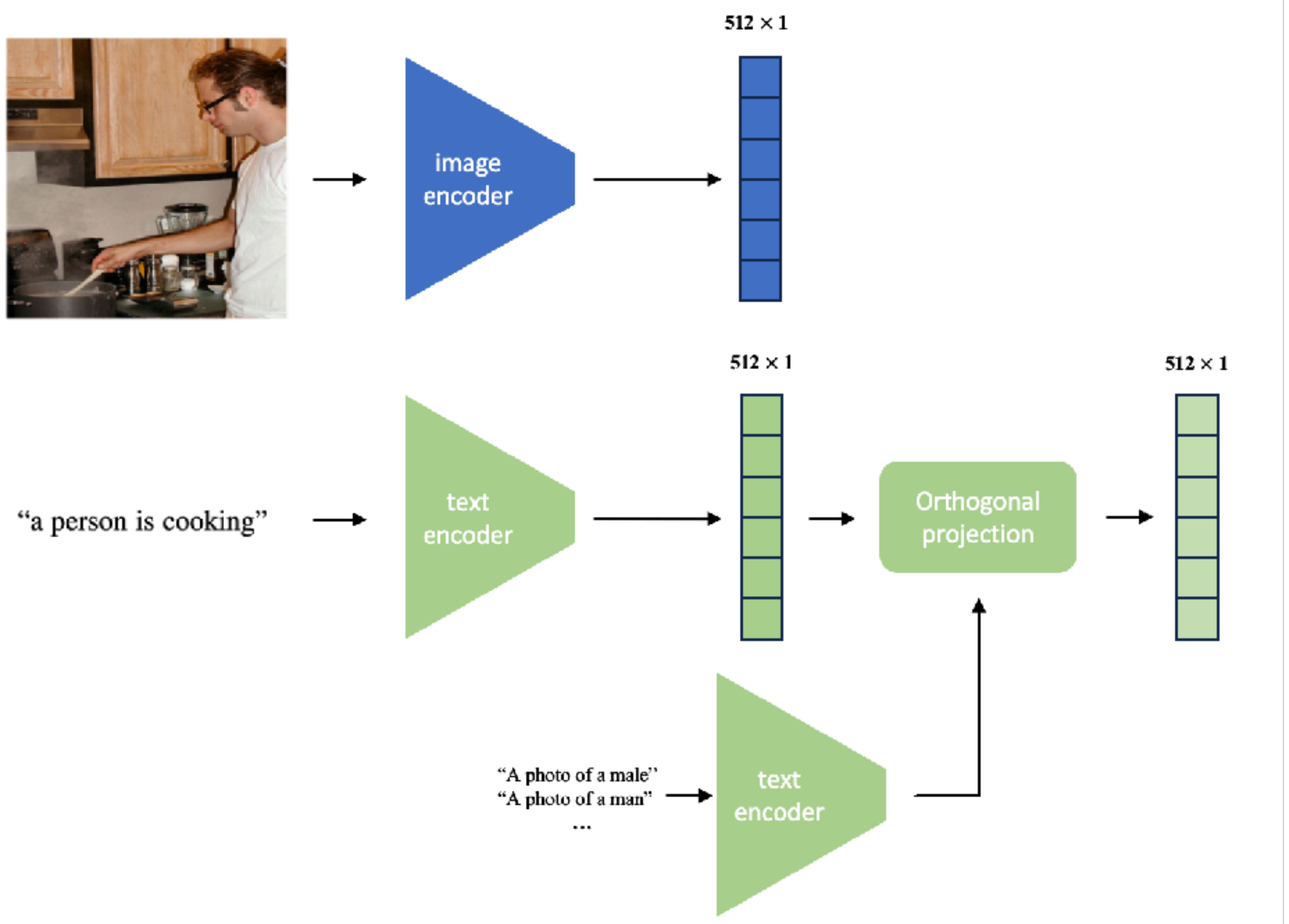
- Debias both image and text embeddings
- Determine the features in the image embedding vectors that contain the most bias information
- Remove those image features from the image embedding vector
- Remove the corresponding text features from the text embedding vector



Literature Review

Embedding Vector Manipulation - Bias Projection

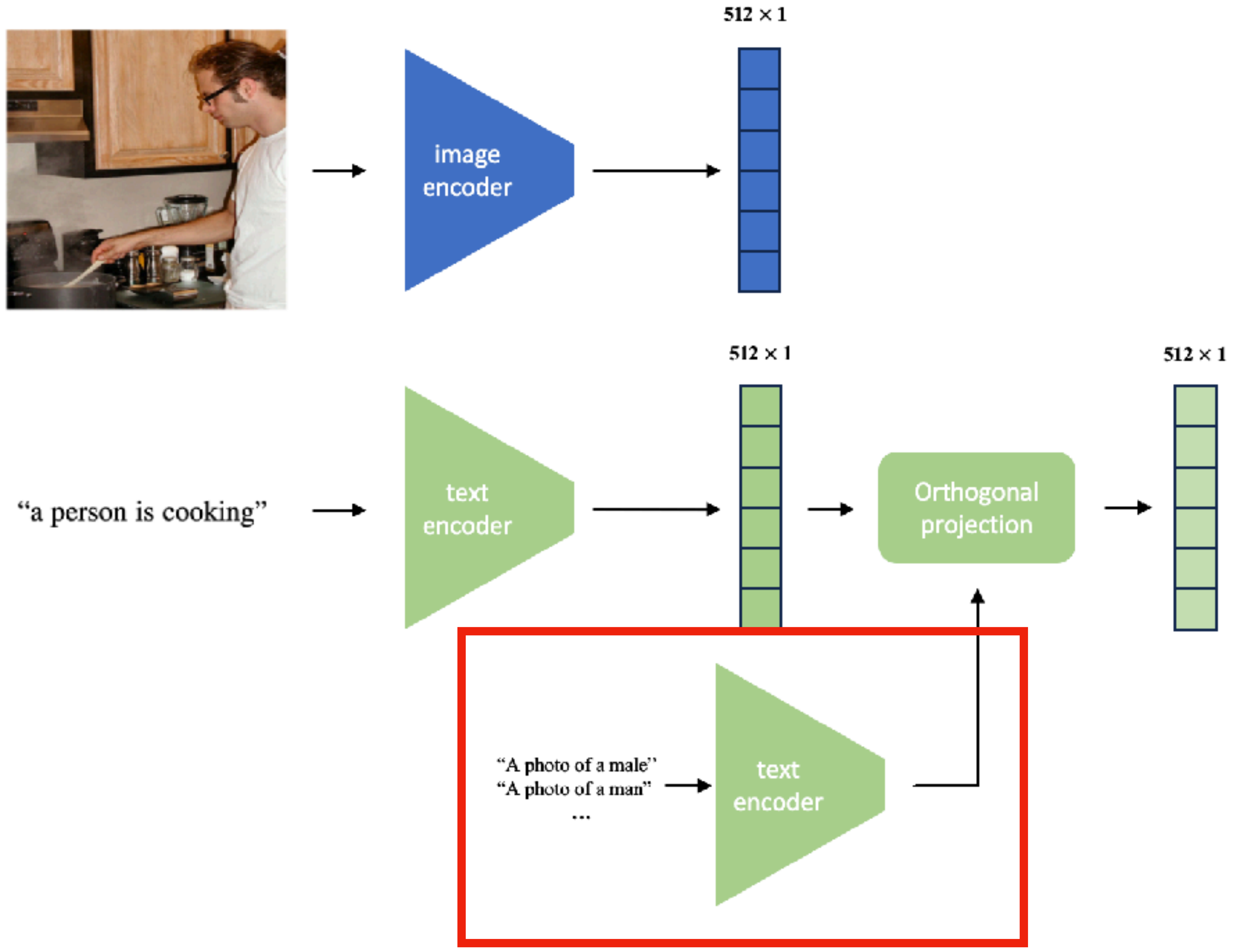
- Debias text embedding only
- Generate biased text prompt embeddings that contain bias information
- Use orthogonal projection to produce debiased text embedding invariant to biased prompts



Literature Review

Embedding Vector Manipulation - Bias Projection

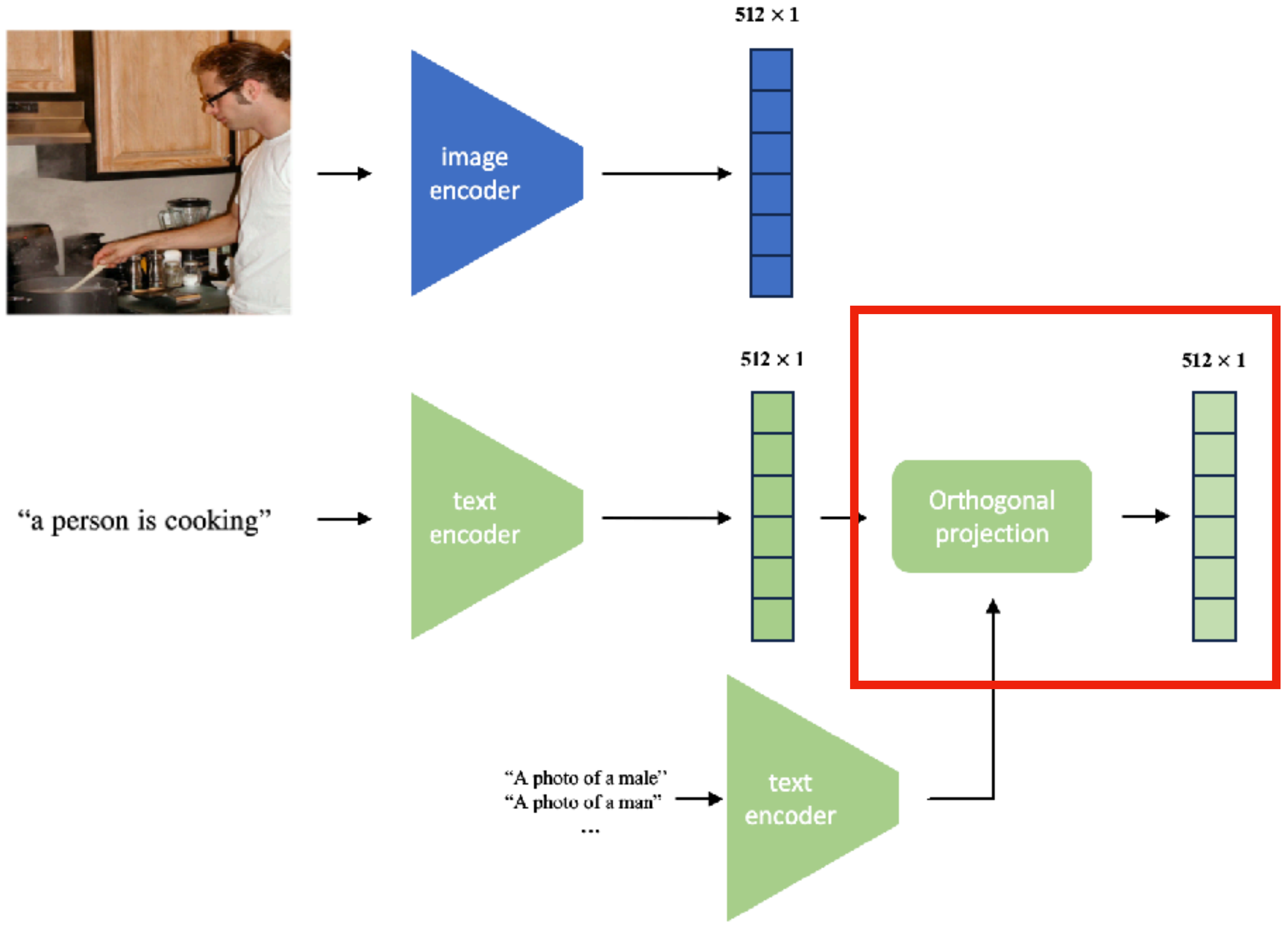
- Debias text embedding only
- Generate biased text prompt embeddings that contain bias information
- Use orthogonal projection to produce debiased text embedding invariant to biased prompts



Literature Review

Embedding Vector Manipulation - Bias Projection

- Debias text embedding only
- Generate biased text prompt embeddings that contain bias information
- Use orthogonal projection to produce debiased text embedding invariant to biased prompts



Literature Review

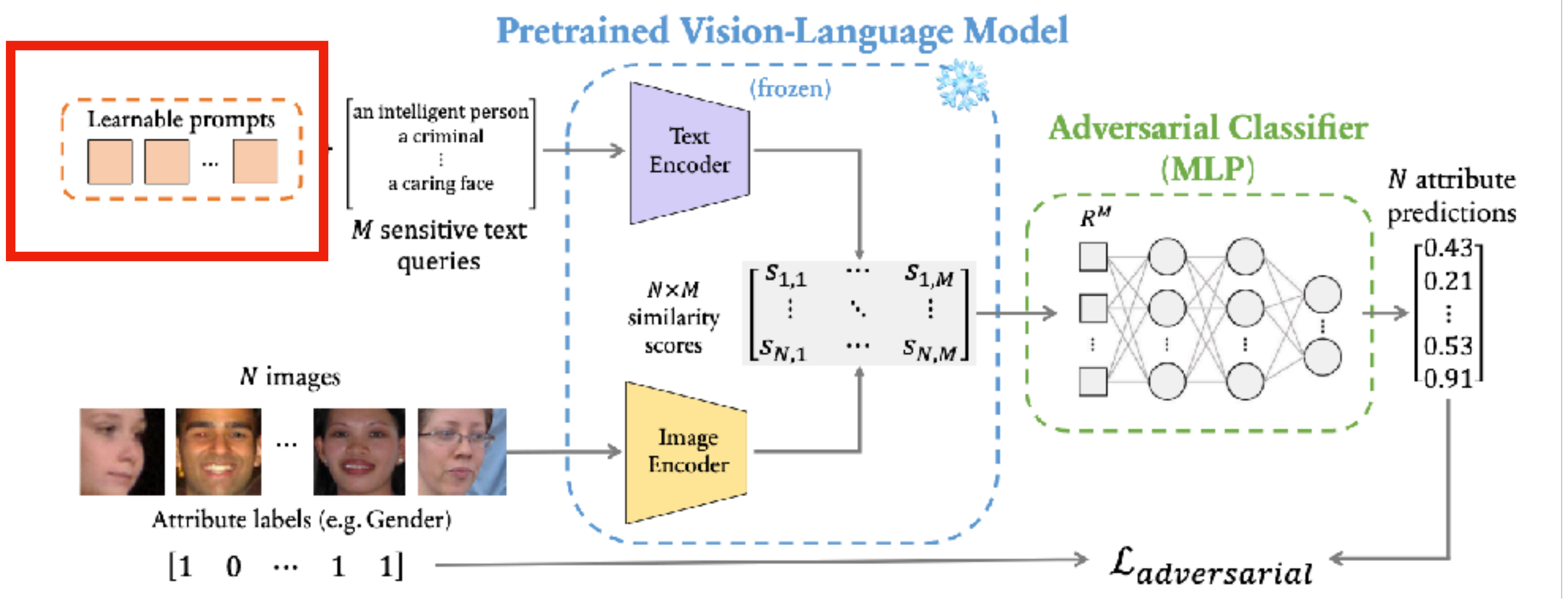
Fair Module Fine-Tuning

- Add trainable fair modules and fine-tune them
  - Learnable text prompt tokens
  - Learnable module attached after image encoder
- Training objectives: adversarial loss, etc. To mitigate bias and keep model performance.

Literature Review

Fair Module Fine-Tuning

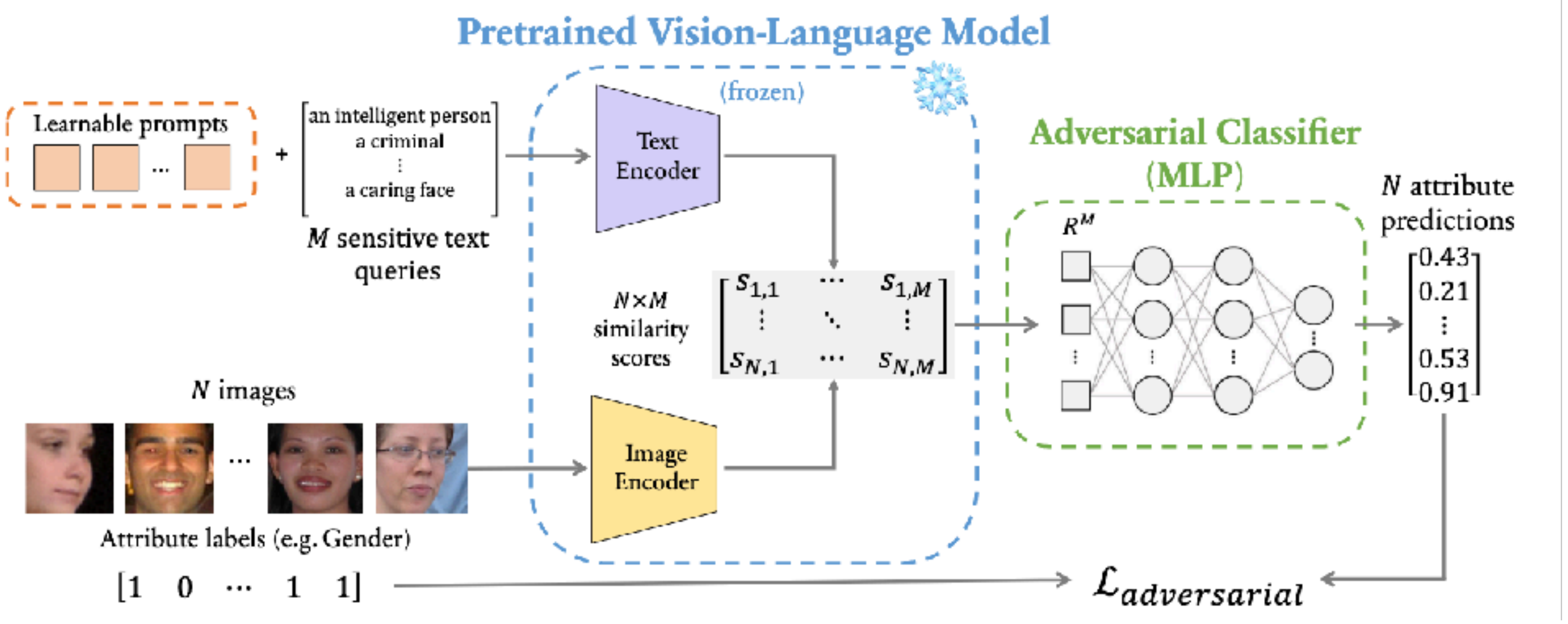
- Add trainable fair modules and fine-tune them
    - Learnable text prompt tokens
    - Learnable module attached after image encoder
- Training objectives: adversarial loss, etc. To mitigate bias and keep model performance.

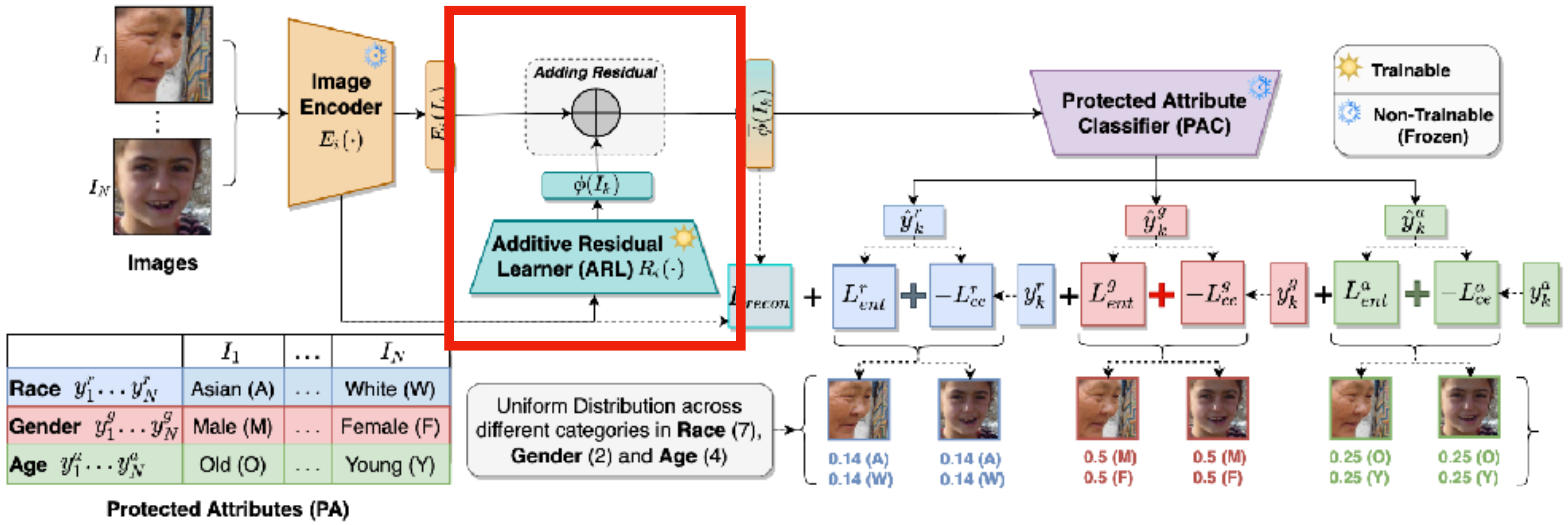


Literature Review

Fair Module Fine-Tuning

- Add trainable fair modules and fine-tune them
  - Learnable text prompt tokens
  - Learnable module attached after image encoder
- Training objectives: adversarial loss, etc. To mitigate bias and keep model performance.





Literature Review

Debiasing Coverage of Existing Techniques

- Debias **text encoder only**
  - Fair module (learnable text prompts) fine-tuning
  - Bias projection
- Debias **image encoder only**
  - Fair module (image debiasing module) fine-tuning
- Debias **both image and text encoders**
  - Feature clipping

Motivation

Limitations of Existing Methods

- **Lack of modality alignment when debiasing image/text encoders**
  - Incomplete removal of text and image biases
  - Harms V-L alignment in the original CLIP model

- Debias **text encoder only**
  - Fair module (learnable text prompts) fine-tuning
  - Bias projection
- Debias **image encoder only**
  - Fair module (image debiasing module) fine-tuning
- Debias **both image and text encoders**
  - Feature clipping

Motivation

Aim

- Unified framework for joint image and text debiasing with modality alignment
    - Study the image and text bias in CLIP
    - Remove bias from both image and text embeddings concurrently

Motivation

Study of Gender Bias in CLIP - Exp 1

- Use t-SNE to visualise the biased text/image embeddings of CLIP
- Qualitatively evaluate the bias distributions

...

A photo of a female teacher

A photo of a male teacher

...

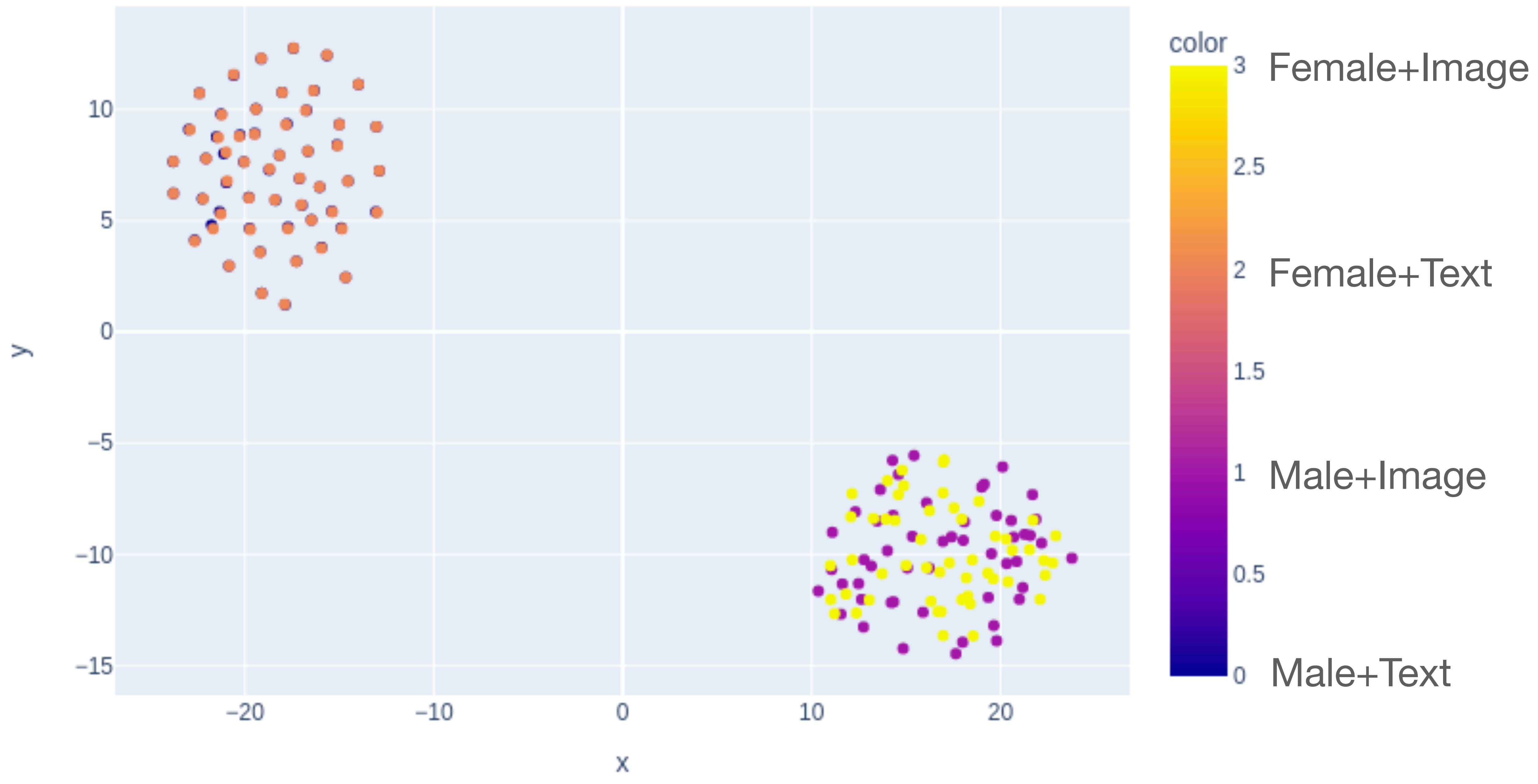
A photo of a female farmer

A photo of a male farmer

Motivation

Study of Gender Bias in CLIP - Exp 1

- **t-SNE visualisation**
  - Top left: text embeddings
  - Bottom right: image embeddings
  - Different bias distribution; more bias in the image embeddings



Motivation

Study of Gender Bias in CLIP - Exp 2

- **Estimate alignment of image and text bias subspaces**
  - Following the analysis in DeAR to disentangle bias information from the original image/text embedding
  - $E_i(I) = \overline{\phi}_i(I) + \phi_i(I)$ (neutral + bias)
  - $E_t(T) = \overline{\phi}_t(T) + \phi_t(T)$
  - $\phi_i(I)$ and $\phi_t(T)$ lie in the image and text bias subspaces respectively

\

\

Motivation

Study of Gender Bias in CLIP - Exp 2

- **Estimate alignment of image and text bias subspaces**
  - Sample: image-text of opposite gender but same concept
  - (Male farmer, female farmer) = $((I_m, T_m), (I_f, T_f))$
  - $E_i(I_m) = \bar{\phi}_i(I_m) + \phi_i(I_m)$ - (1)
  - $E_i(I_f) = \bar{\phi}_i(I_f) + \phi_i(I_f)$ - (2)
  - (1) - (2): $E_i(I_m) - E_i(I_f) = \phi_i(I_m) - \phi_i(I_f)$ (difference in image bias for opposite genders)
  - Similarly, $E_t(T_m) - E_t(T_f) = \phi_t(T_m) - \phi_t(T_f)$ (difference in text bias for opposite genders)
  - To check alignment of text and image subspaces, we can check whether $\phi_i(I_m) - \phi_i(I_f)$ and $\phi_t(T_m) - \phi_t(T_f)$ align with each other across different samples of $((I_m, T_m), (I_f, T_f))$

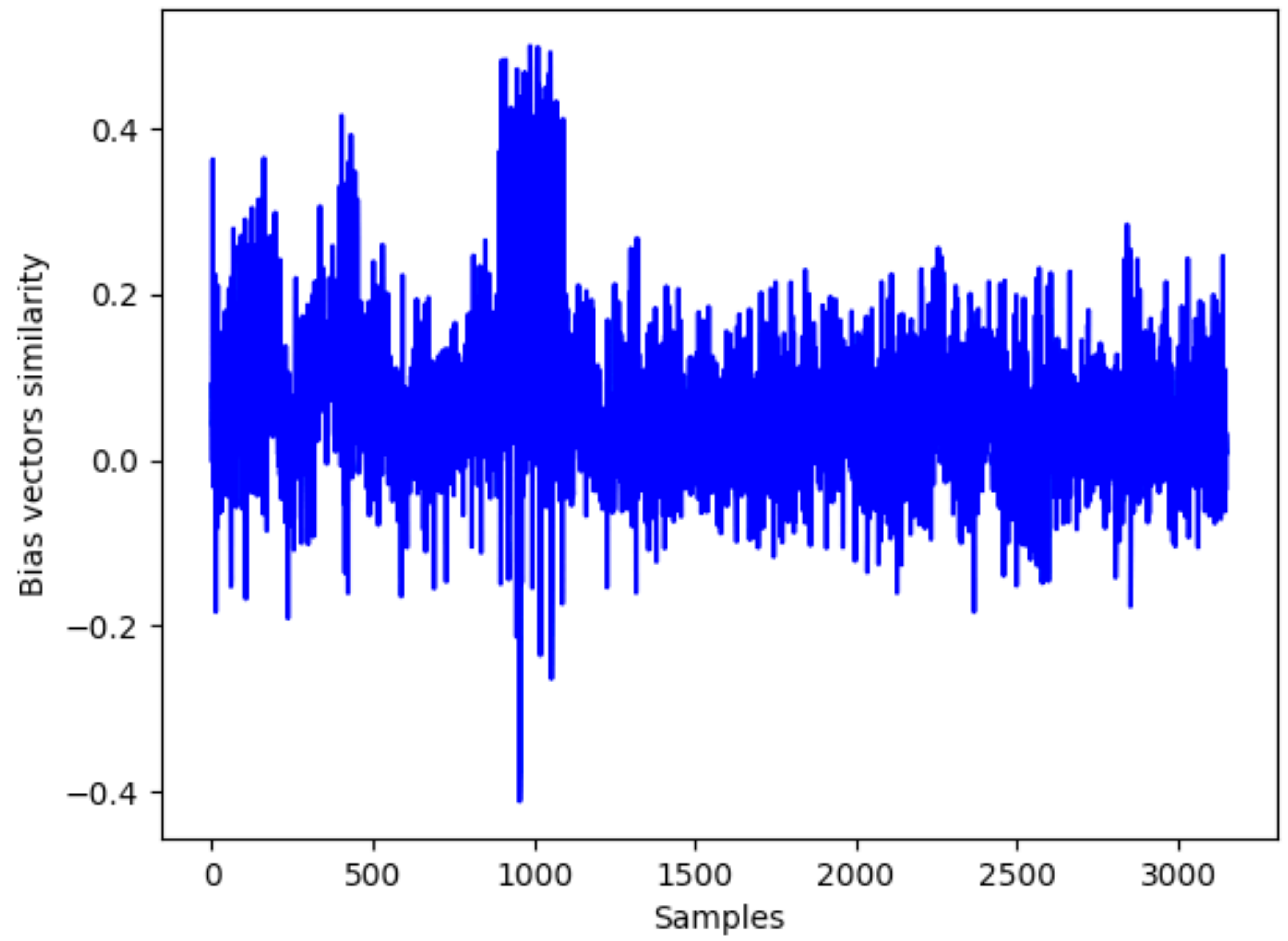
A photo of a male teacher

A photo of a female teacher

Motivation

Study of Gender Bias in CLIP - Exp 2

- **Estimate alignment of image and text bias subspaces**
  - Compare the cosine similarity between $\phi_i(I_m) - \phi_i(I_f)$ and $\phi_t(T_m) - \phi_t(T_f)$ for each sample
  - Each sample: a set of $((I_m, T_m), (I_f, T_f))$ that share the same concept (e.g. "farmer")
  - ~3k samples are used
  - Results: varied and low similarities across samples; there is no evidence that two bias subspaces are aligned.



A photo of a female teacher

A photo of a male teacher

Motivation

Study of Gender Bias in CLIP - Exp 3

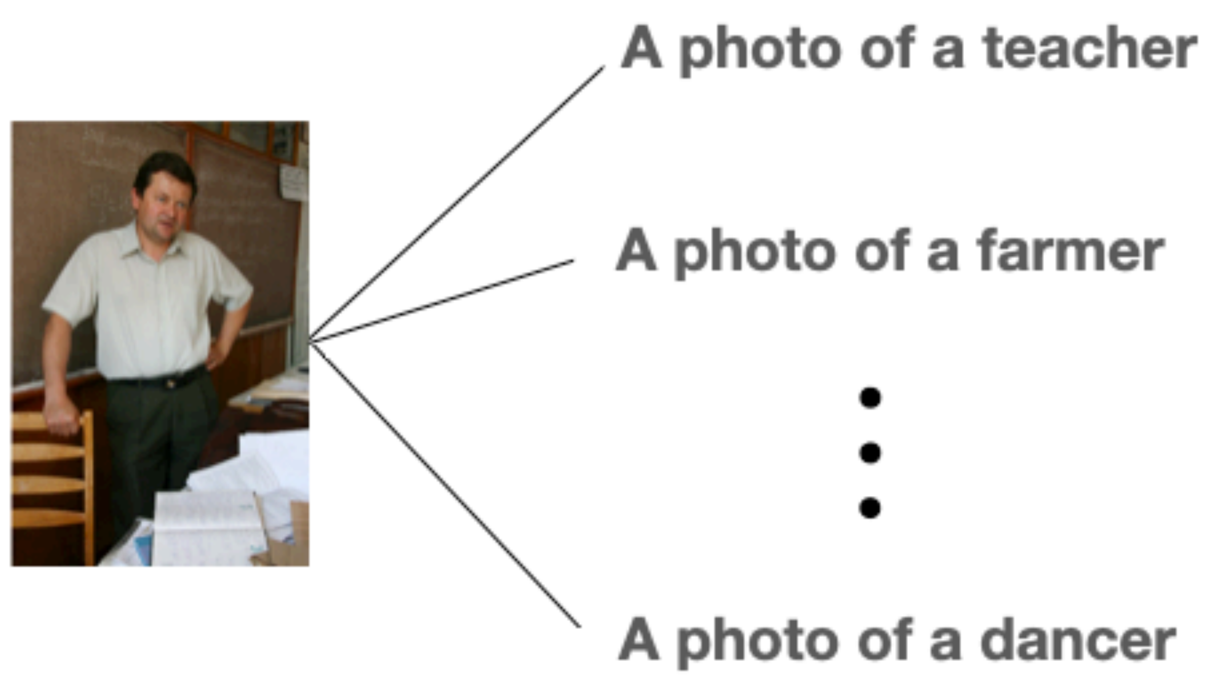
- **Bias from Cross-Modal Interaction**
  - Test 1: Image concept matched to text concepts
  - Test 2: Text concept matched to image concepts
  - Stage 1: Test 1 and Test 2
  - Stage 2: Test 1 and Test 2 with gender information

Motivation

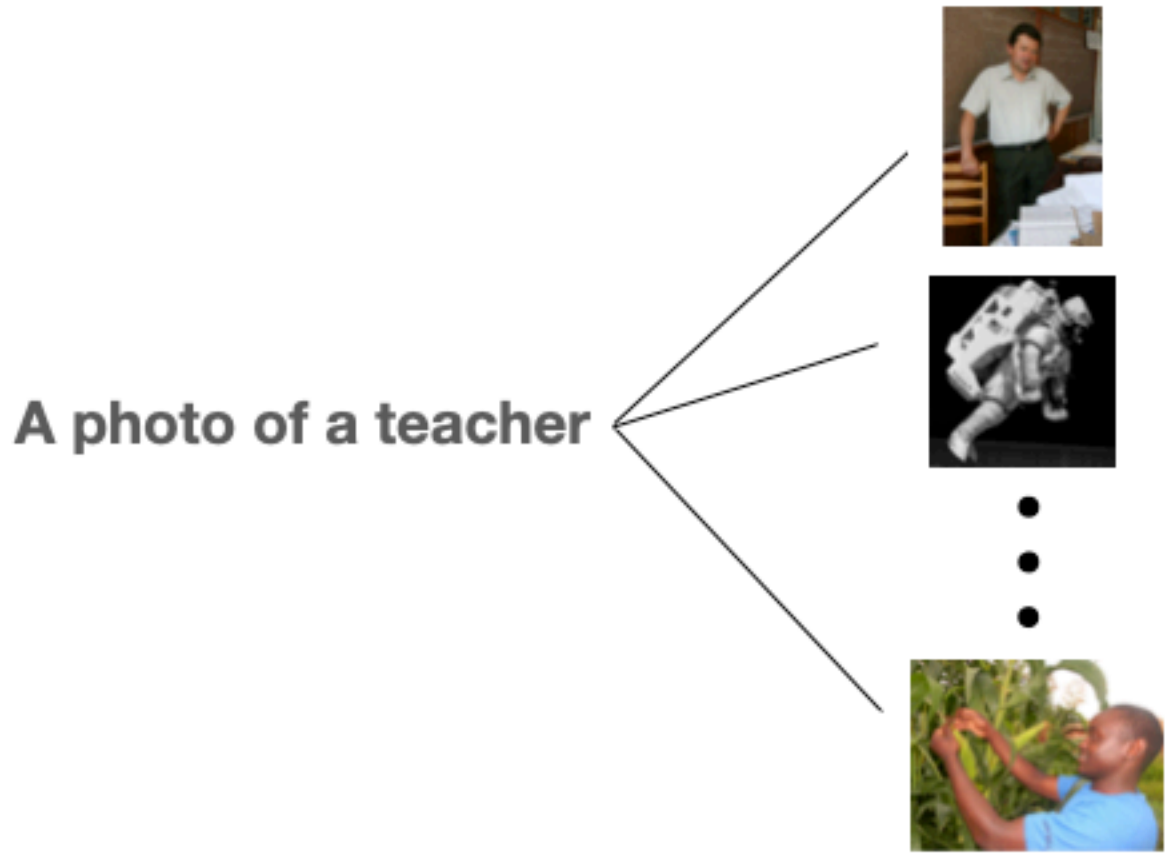
Study of Gender Bias in CLIP - Exp 3

Matching an image to text prompts

Stage 1



Matching a text to image prompts



Motivation

Study of Gender Bias in CLIP - Exp 3

Matching an image to text prompts

Stage 1

A photo of a teacher (Correct match)

A photo of a farmer

⋮

A photo of a dancer

Matching a text to image prompts

A photo of a teacher

(Correct match)

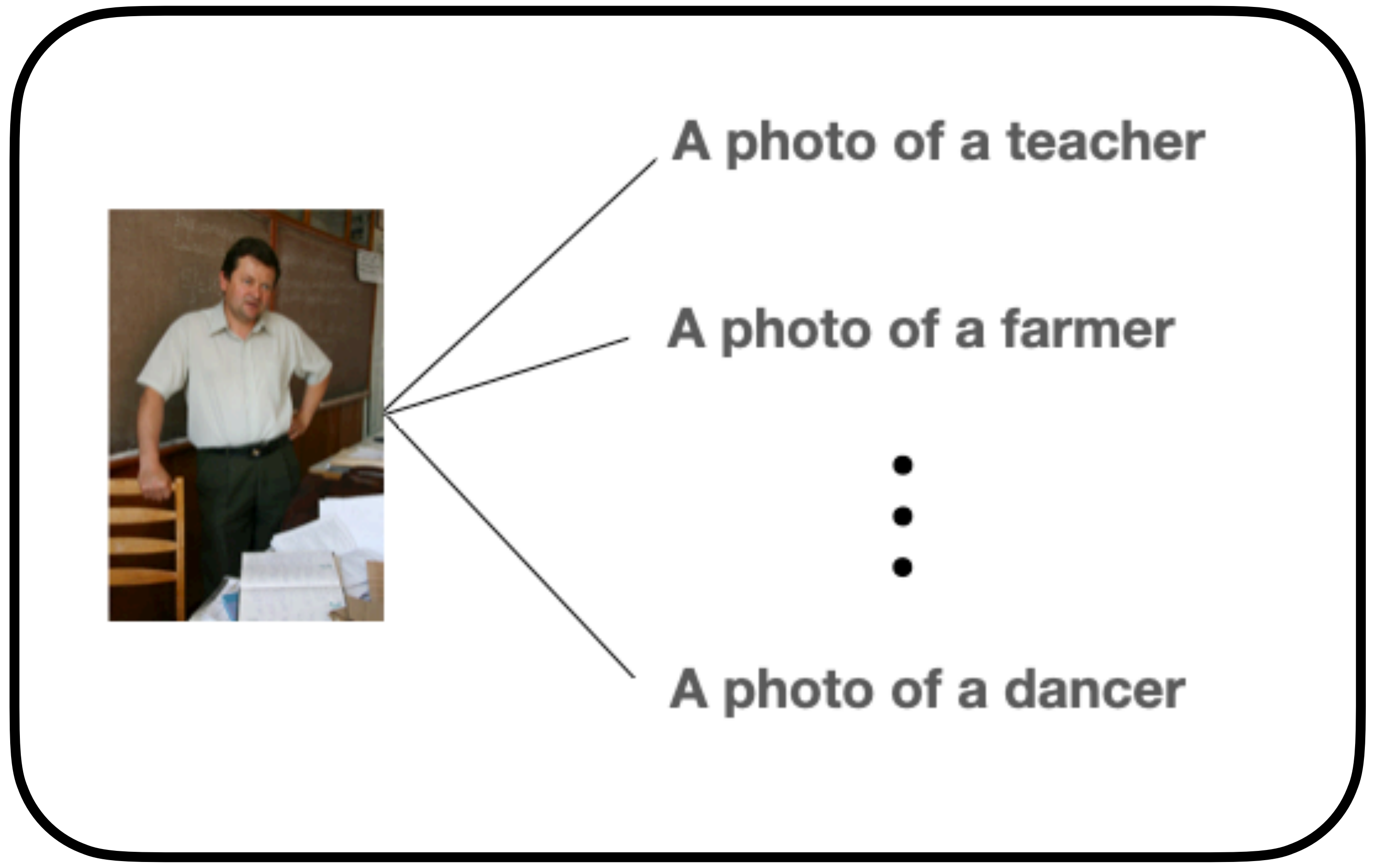
⋮

Motivation

Study of Gender Bias in CLIP - Exp 3

Matching an image to text prompts

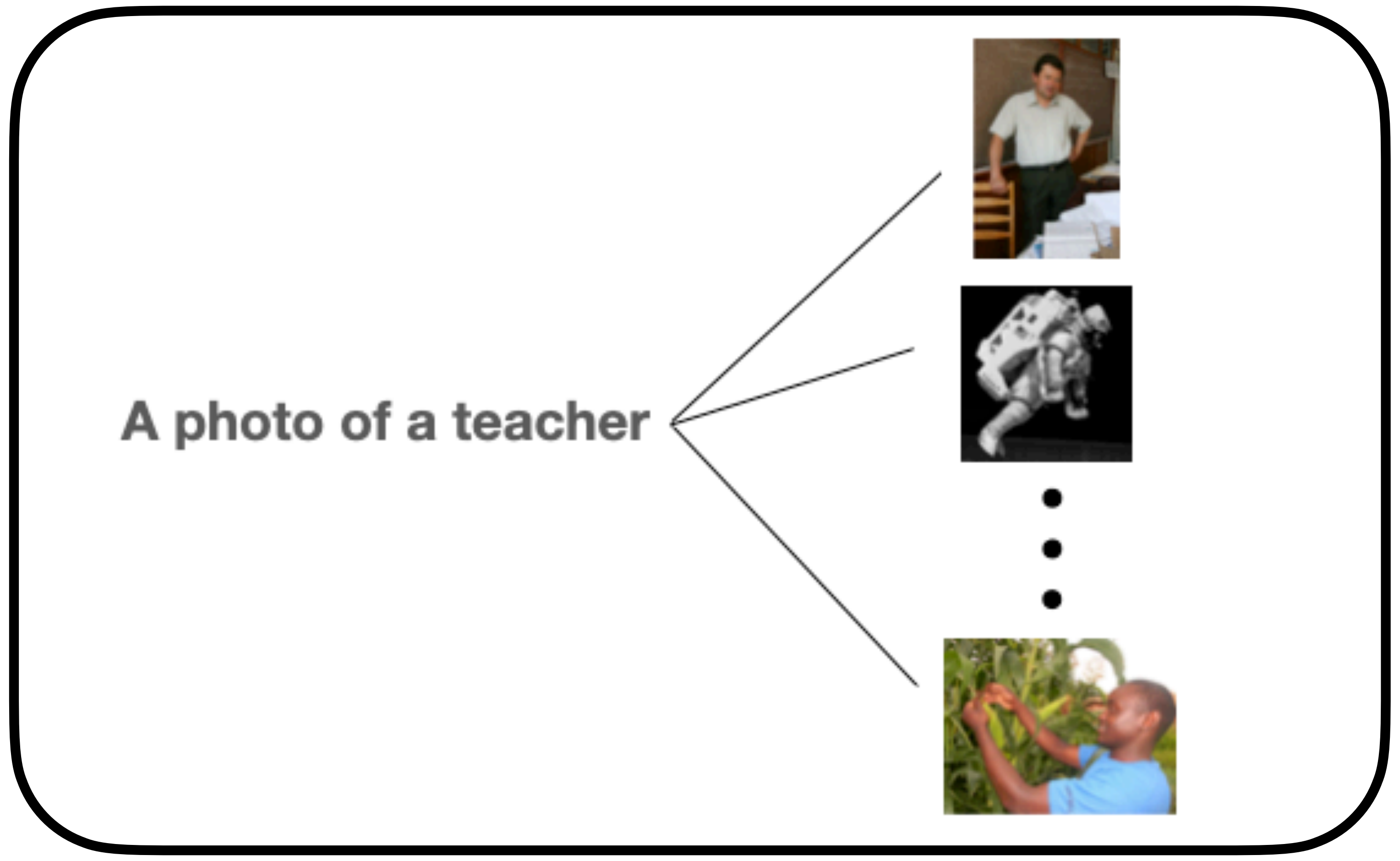
Stage 1



$\times 416$

Correct matching rate: 52.6%

Matching a text to image prompts

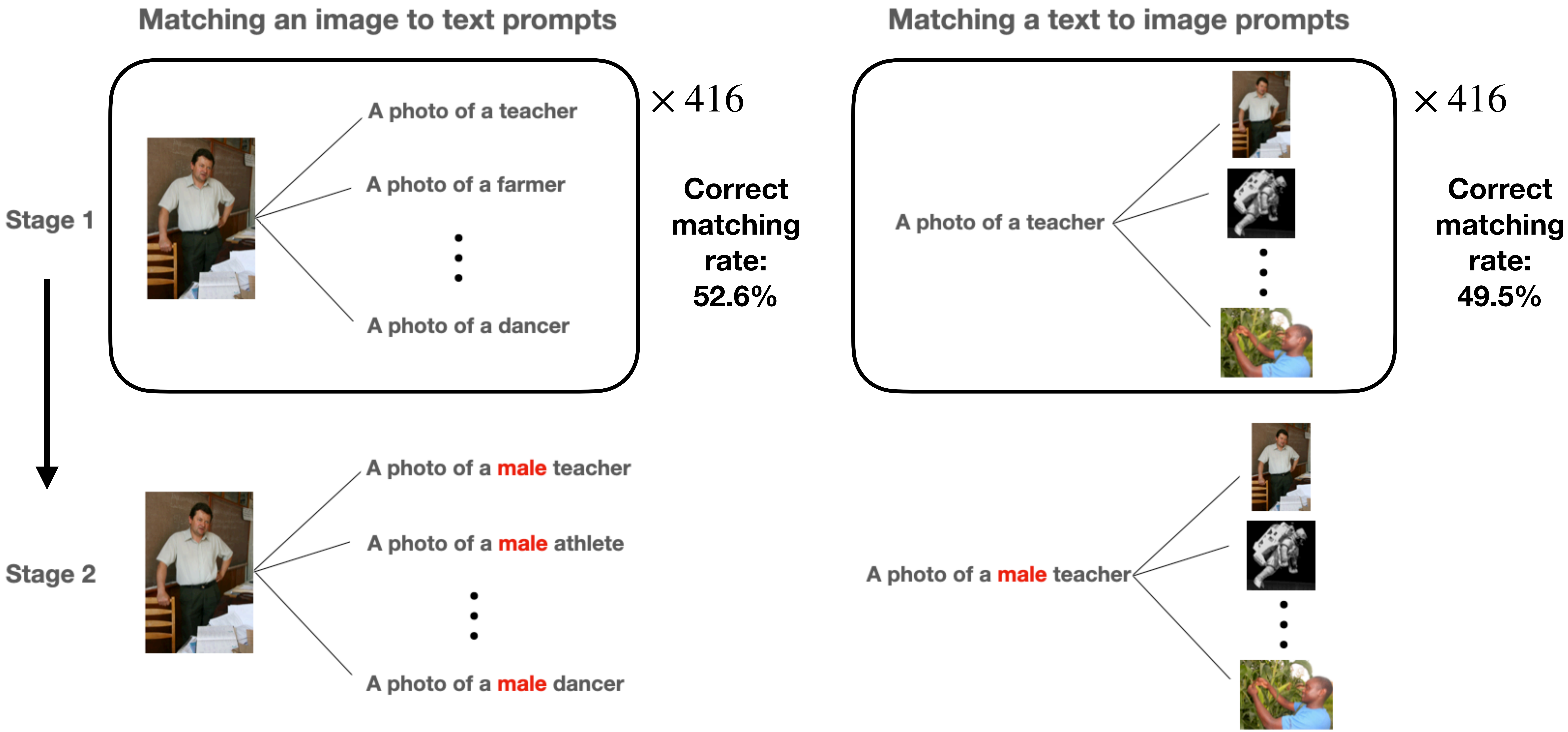


$\times 416$

Correct matching rate: 49.5%

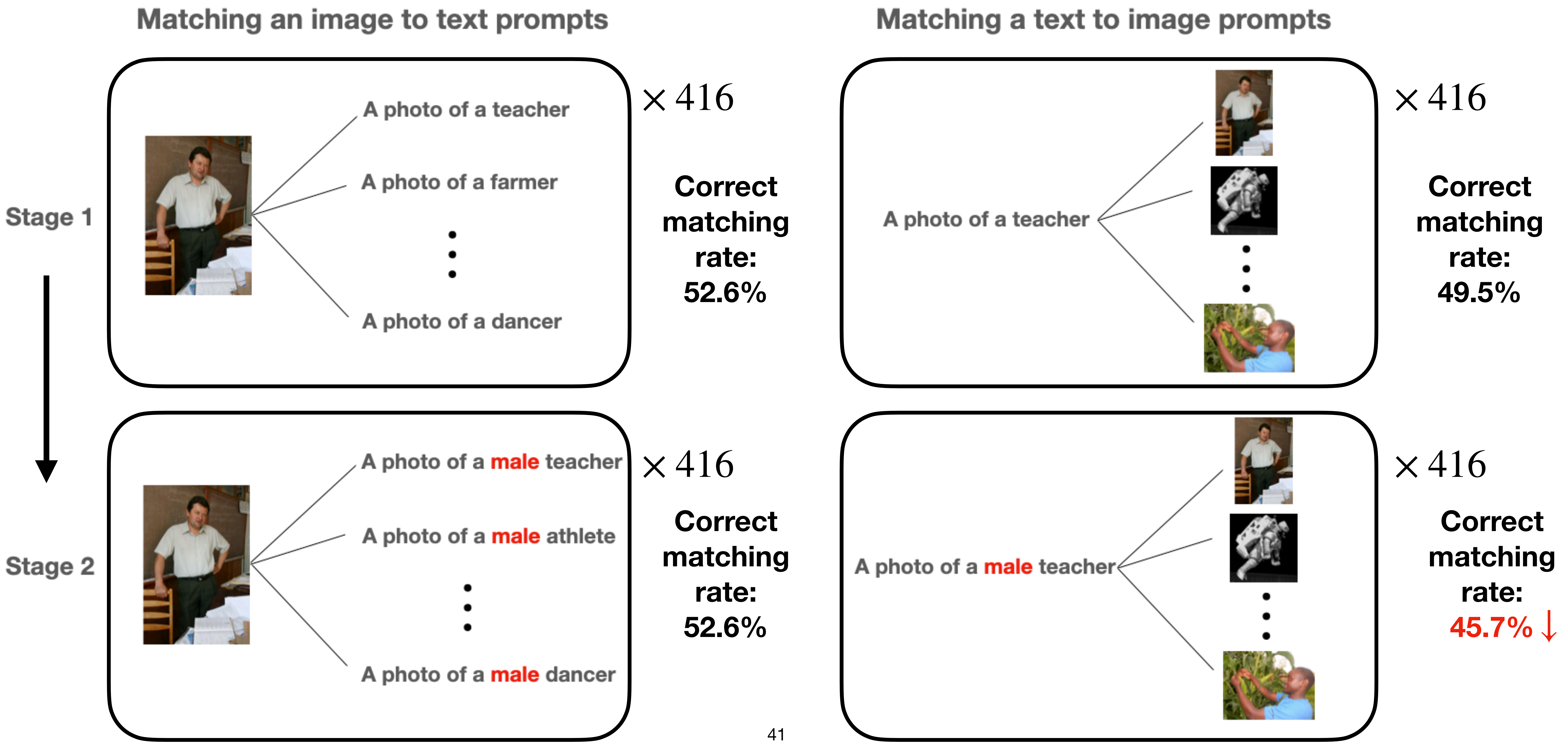
Motivation

Study of Gender Bias in CLIP - Exp 3



Motivation

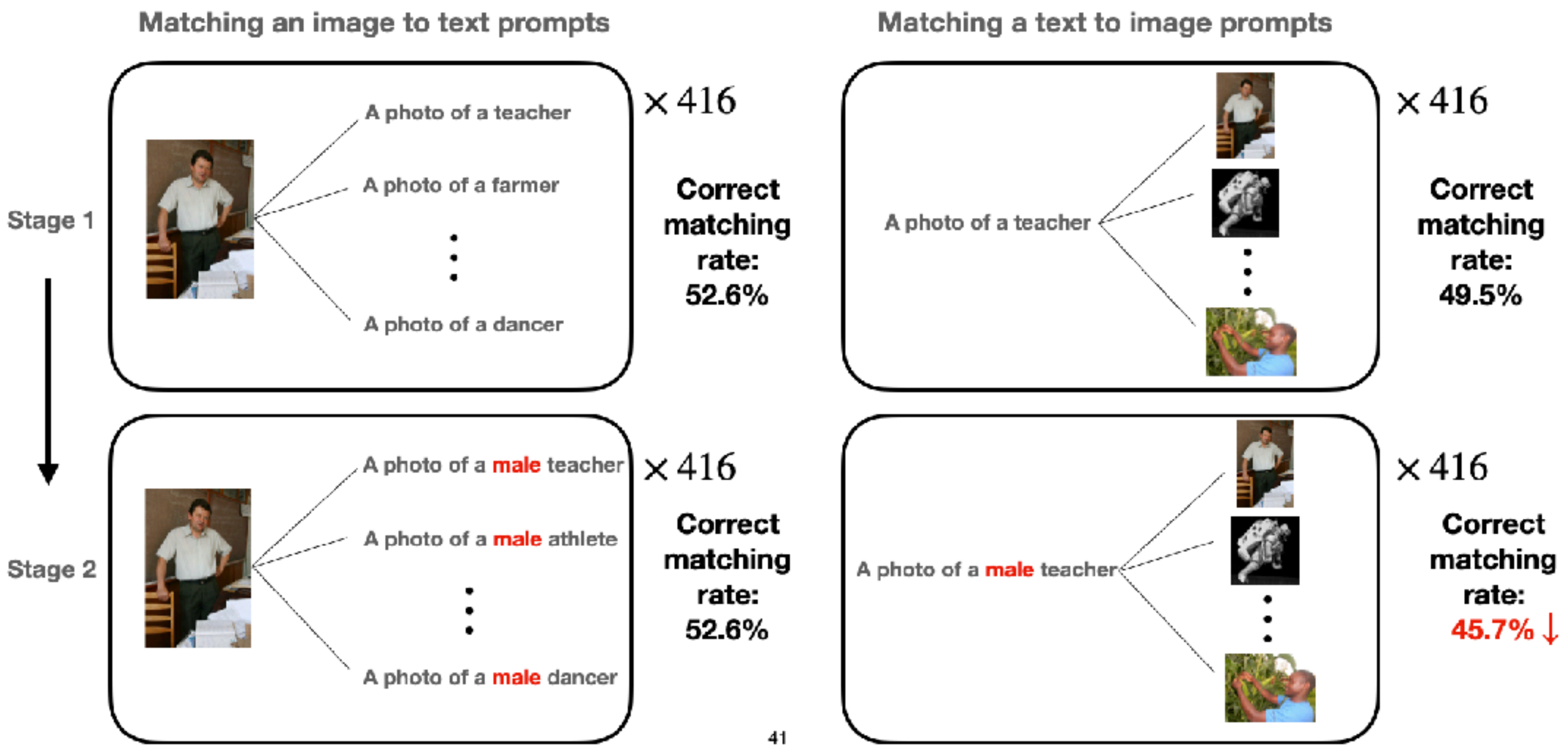
Study of Gender Bias in CLIP - Exp 3



Motivation

Study of Gender Bias in CLIP - Exp 3

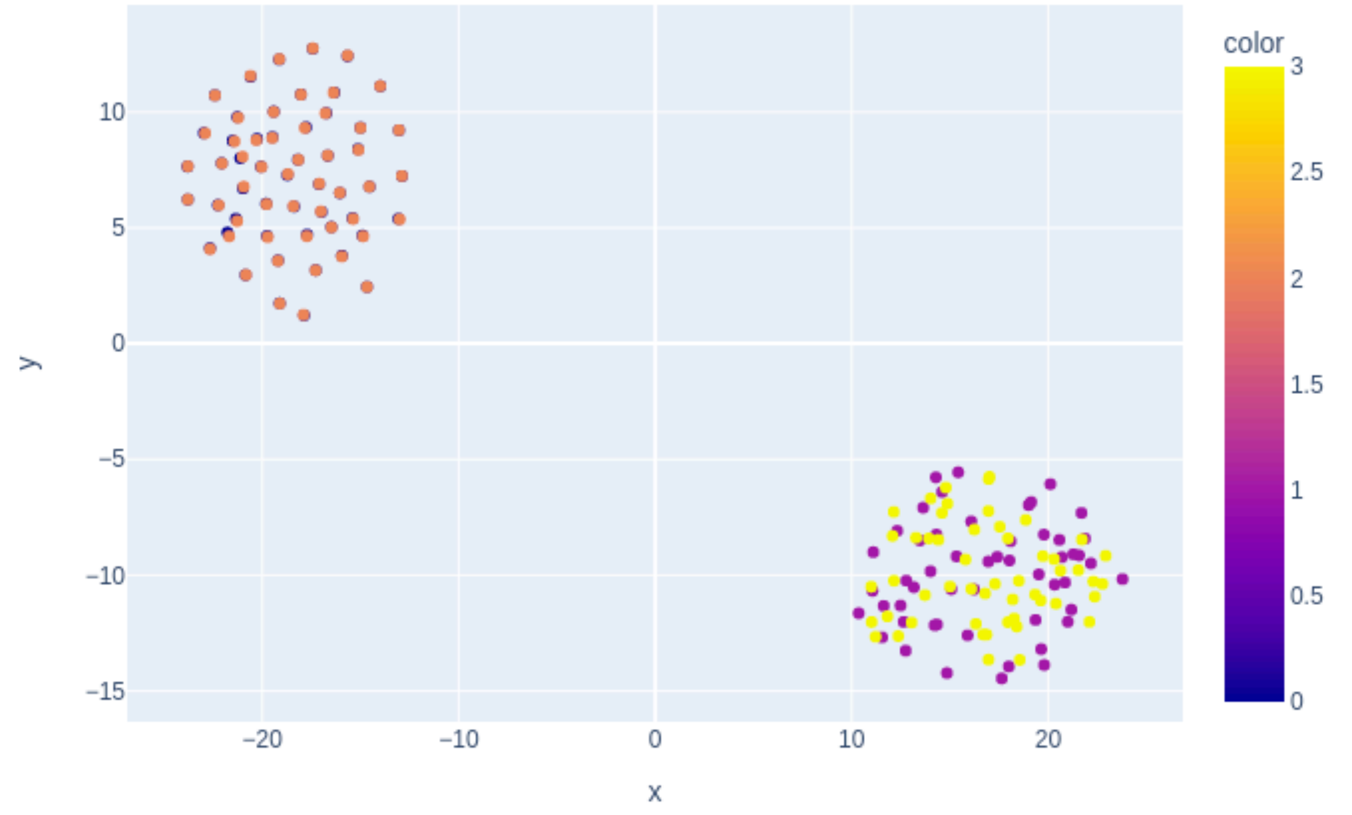
- **Bias from Cross-Modal Interaction**
  - More gender association leads to lower correct matching rate of **a text** to **image prompts**
  - Gender bias in certain images, closer to the text prompt with gender information
  - **More significant bias in image embeddings**

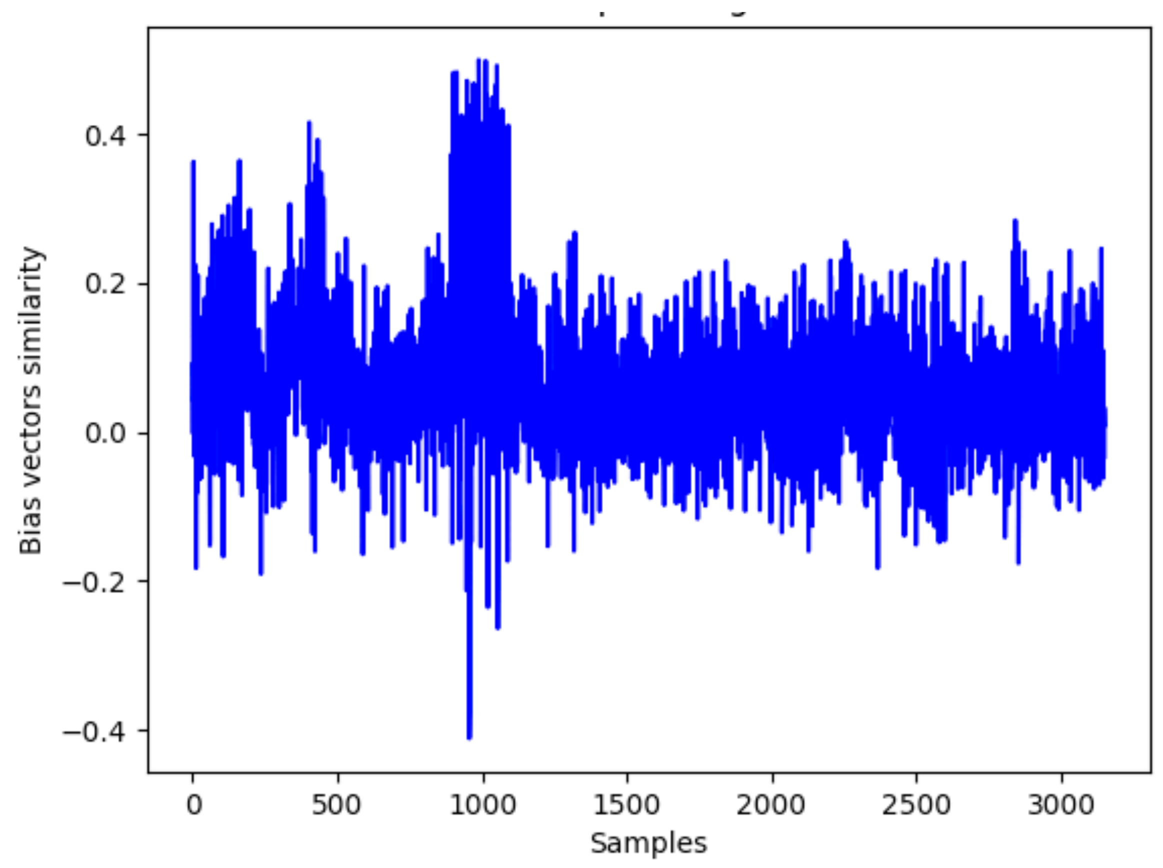


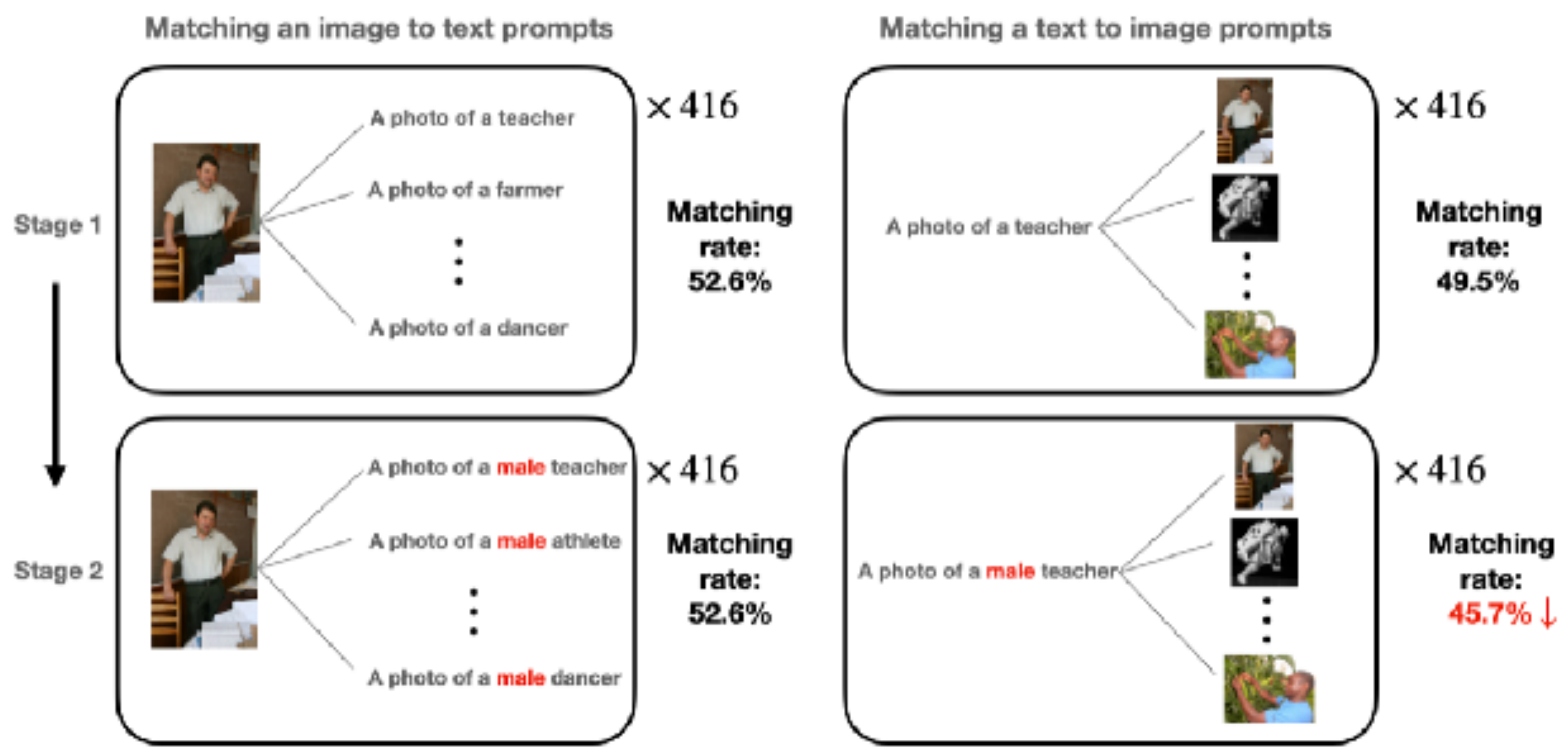
Motivation

Study of Gender Bias in CLIP

- **Summary**
  - Image biases seem to be more significant (based on Exp 1 and 3)
  - Image and text biases may manifest differently, as there is no evidence showing that their bias subspaces are aligned (Exp 2)







Proposed Method

Ideas (not finalised)

- **Bias subspaces alignment**
    - Learnable module to transform image bias subspace to text bias subspace (or the other way)
- **Joint V-L debiasing**
    - Debias text embeddings/image embeddings
    - The other modality will be debiased at the same time
- Currently on gender, later extend to race/age…

Proposed Method

Evaluation Metrics

- Fairness
  - Retrieval-based metrics
  - Fairness in generative models: images generated by Stable Diffusion with our debiased CLIP text encoder
- V-L task performance
  - Zero-shot classification
  - Zero-shot retrieval

Proposed Method

Evaluation Metrics

- Fairness
    - Retrieval-based metrics
    - Fairness in generative models: images generated by Stable Diffusion with our debiased CLIP text encoder
- V-L task performance
    - Zero-shot classification
    - Zero-shot retrieval

Proposed Method

Evaluation Metrics

- Fairness
  - Retrieval-based metrics
  - Fairness in generative models: images generated by Stable Diffusion with our debiased CLIP text encoder
- V-L task performance
  - Zero-shot classification
  - Zero-shot retrieval

Thank you for listening!