The Bias Amplification Paradox in Text-to-Image Generation

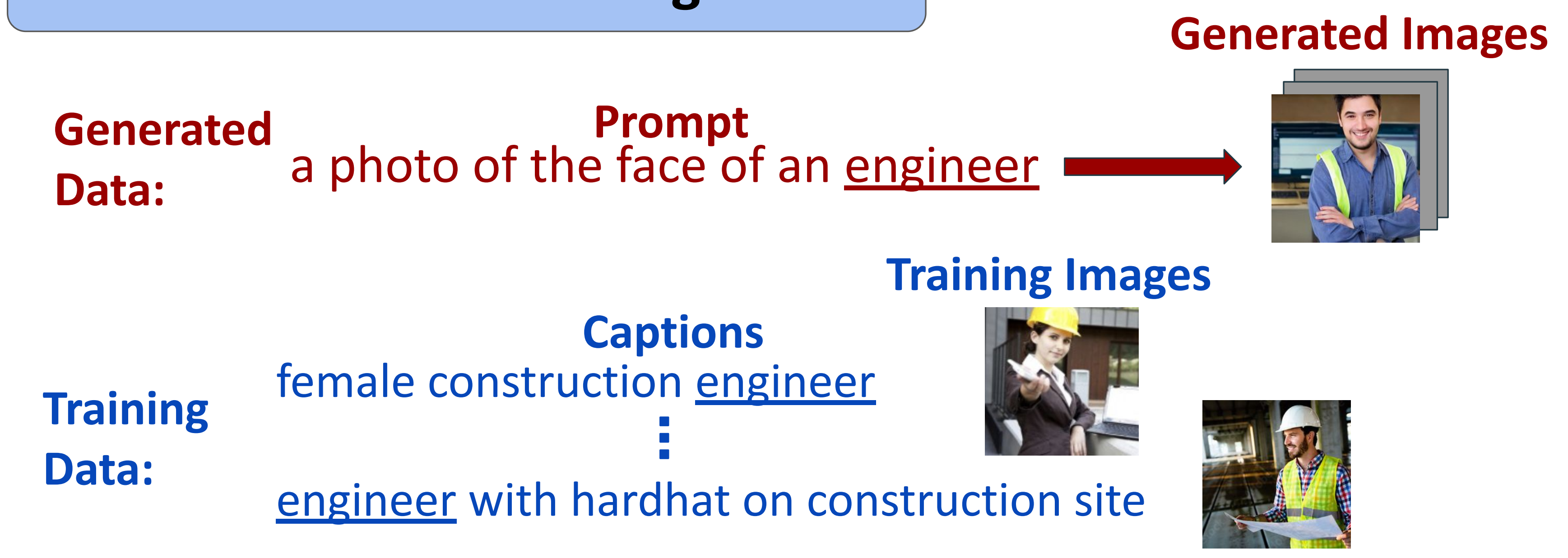
Preethi Seshadri, Sameer Singh, Yanai Elazar
preethis@uci.edu

Research Question

The Bias Amplification Paradox: Given that models learn to fit the training data distribution, why do models ***amplify*** biases in the training data as opposed to strictly ***representing*** them?

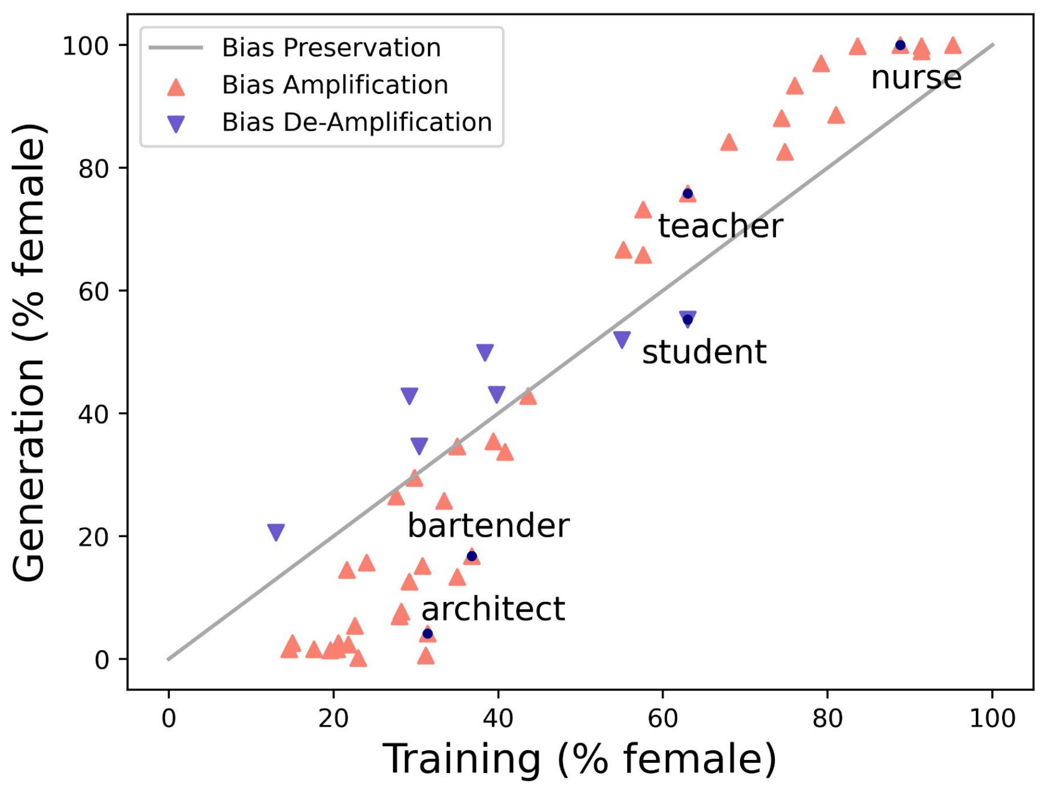
* We focus on gender-occupation associations in text-to-image generation using Stable Diffusion and its training dataset, LAION.

Generated and Training Data


Generated Images
Generated Data:
Prompt
a photo of the face of an engineer
Training Images
Captions
Training Data:
female construction engineer
⋮
engineer with hardhat on construction site

After obtaining training and generated images, we compare the % female in generated vs. training data for each occupation.

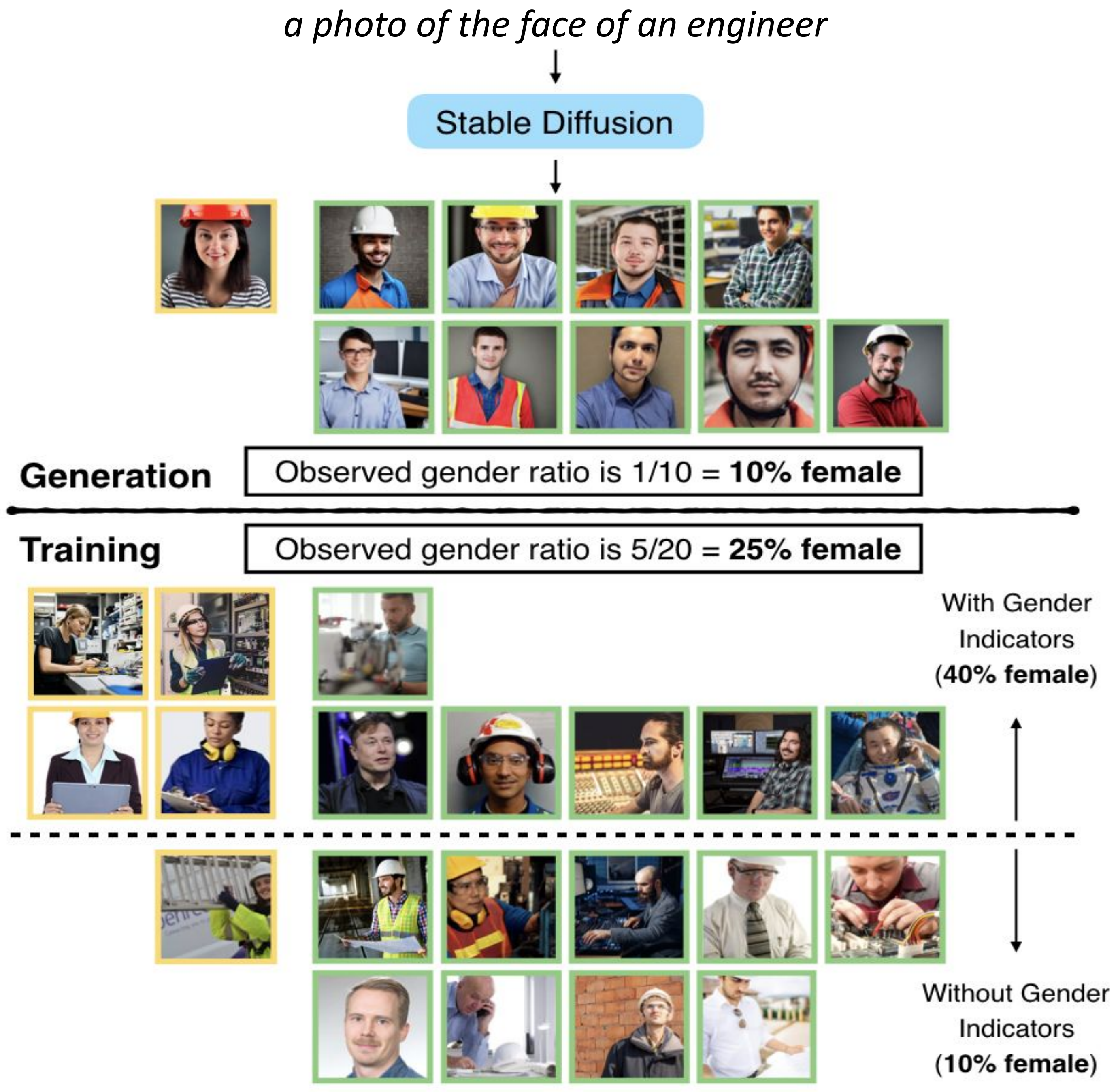
Bias Amplification in Generation

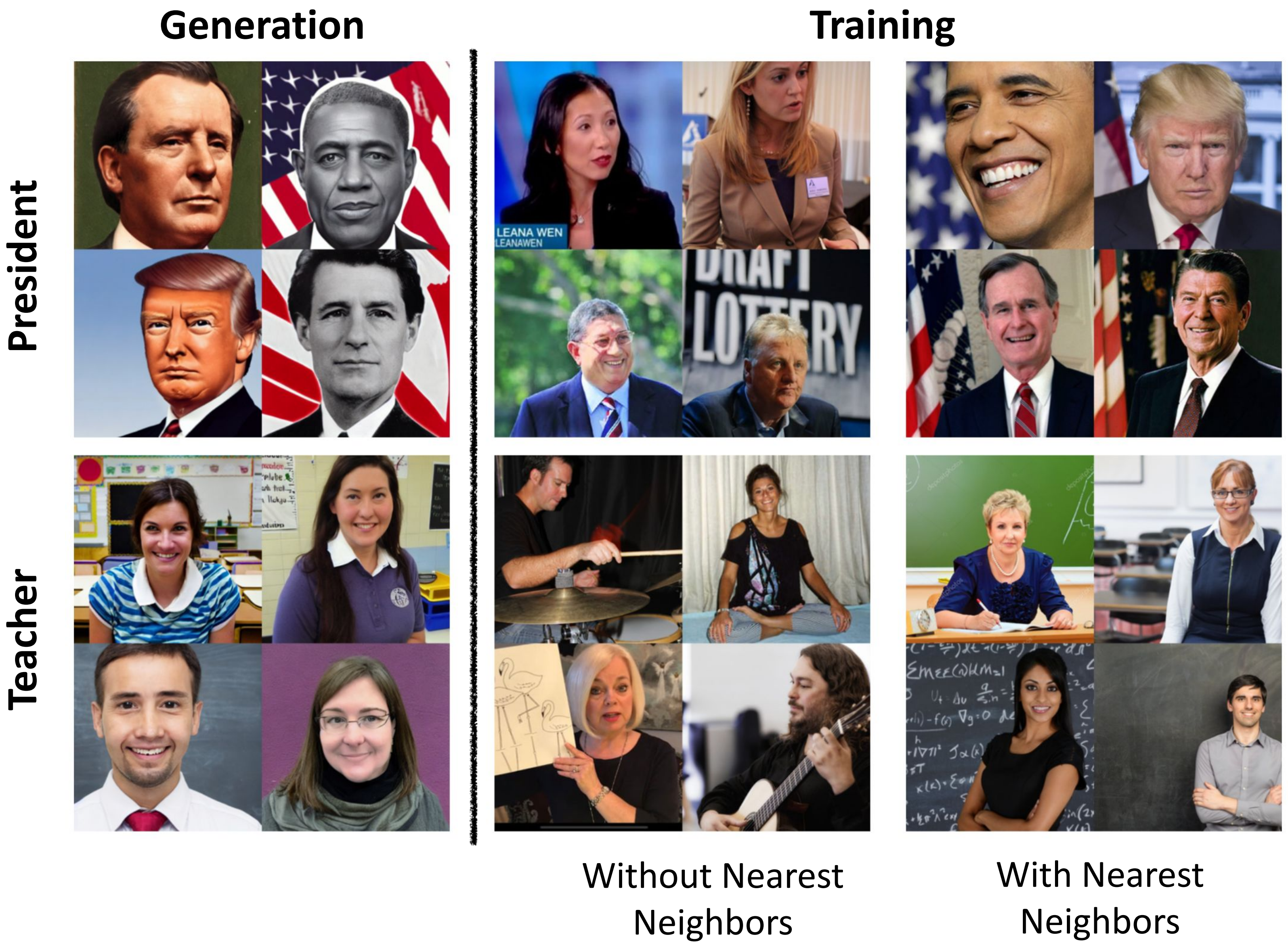


Bias is amplified considerably across occupations!
What happens when we dig deeper into the data?

Investigating Discrepancies

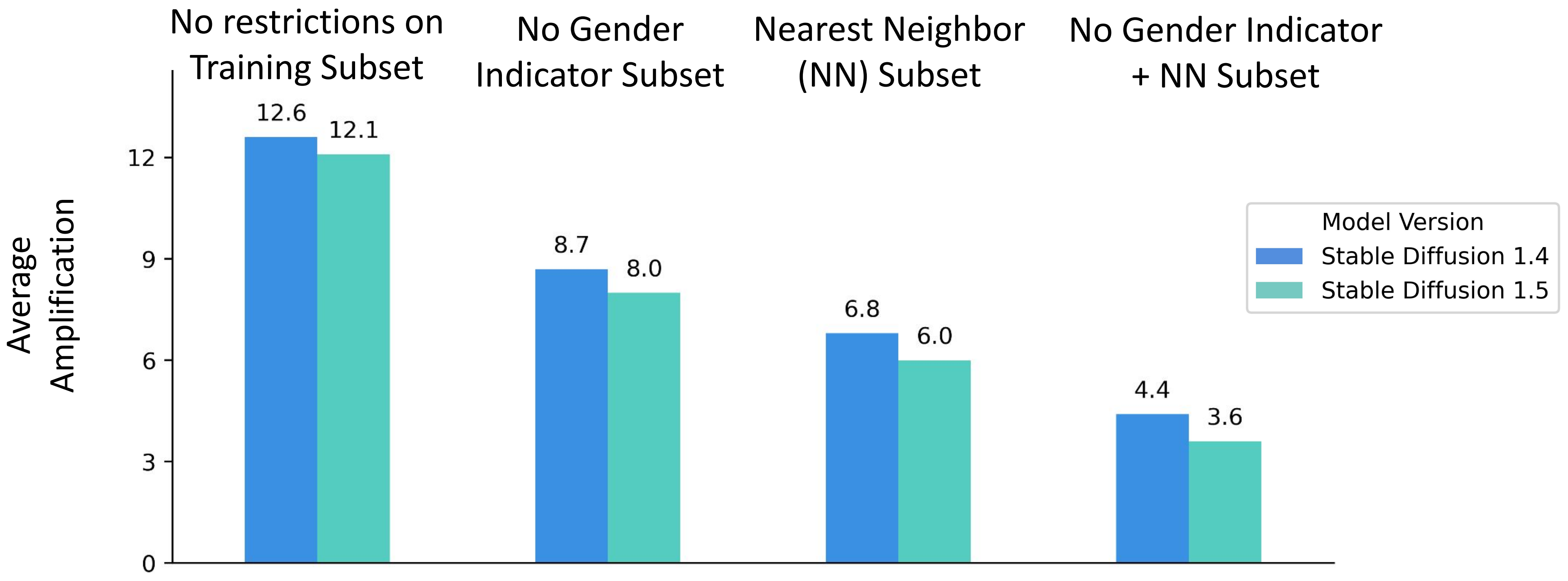
Training captions often contain (1) **explicit gender indicators** and (2) additional context, which may **implicitly convey gender information.** In contrast, the prompts we use exclude this information by design. What if we select captions that resemble prompts?


a photo of the face of an engineer
Stable Diffusion
Generation
Observed gender ratio is 1/10 = **10% female**
Training
Observed gender ratio is 5/20 = **25% female**
With Gender Indicators (**40% female**)
Without Gender Indicators (**10% female**)

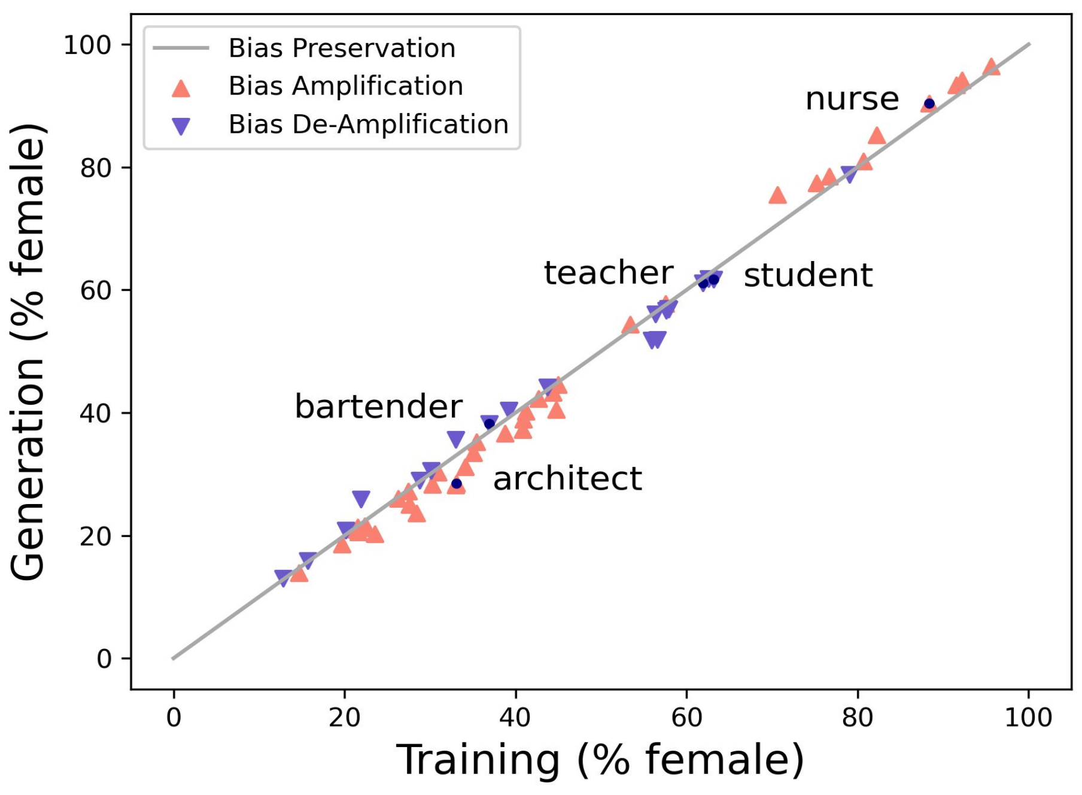


Addressing Discrepancies Reduces Amplification

What happens if we **restrict the subset of training examples** in our evaluation?



What happens if we eliminate discrepancies by **prompting the model with training captions**?



Conclusion

- Naive evaluations of amplification overlook notable differences between training and generation, and inflate amplification.
- Models actually match the training data distribution quite well.