AI Model Training

Charlotte Li, Ian Lei, Sonya Dymova, Chris Cao, Kiran Sheth, Shreyasi Periketi

Problem

- Generative AI is revolutionary, but nowhere near perfect
- Flawed training data → problematic outcomes
- Tools produce content projecting a limited worldview

Can we train the AI to create art that is more editorial in nature?

Project Scope

- Use Stable Diffusion to create images used for editorial purposes
- Train models to create more fair and balanced images more representative of society
- Determine how to prompt AI to create these types of images

LoRA – V7

Training

- Social Worker: More variety in activities
- Fast food Worker: More male representation
- Construction Worker: More female representation

Observations

- Even one training cycle can have significant impacts
- For some professions (e.g. social worker) there have been more conceptual learning than for others (e.g. construction worker)
    - Different captions?

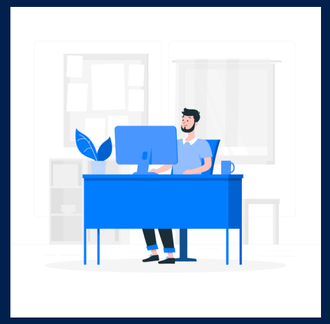
LoRA – V8

Training (WORKER_TEST_V8)

- Social Worker: Removed Environmental descriptions.
- Fast food Worker: Removed gendered language from the description.
- Construction Worker: Edited captions to not include words with strong gender connotations (such as “hard, tractor, etc”).

Observations

- Removing adjectives really helped with the biases in construction workers.
- Faces are distorted when we changed the description of the environment.

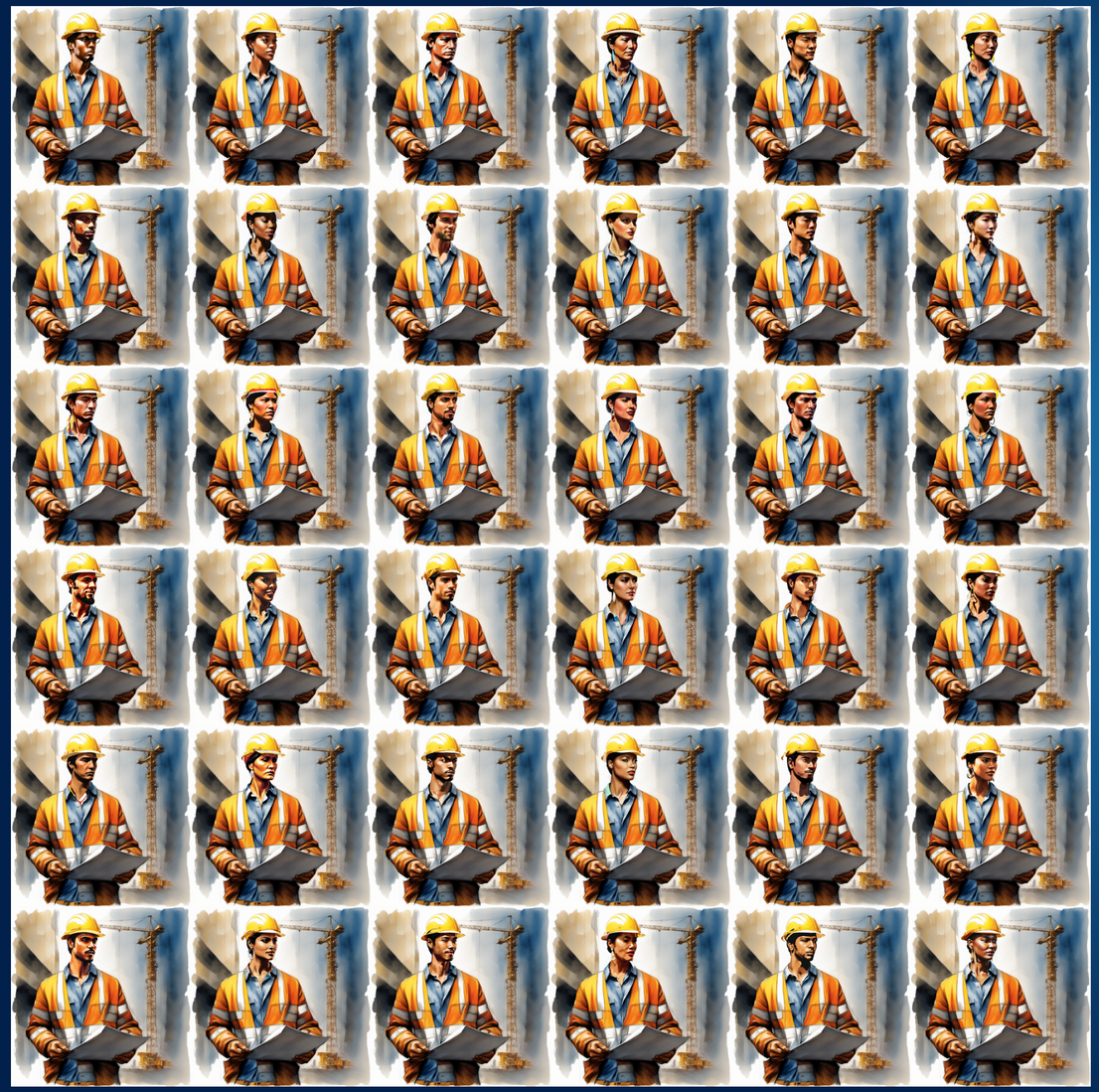


Chris

IP – Adapter

IP: Image Prompting

- Using stable diffusion to identify image features
- Using stable diffusion to generate mask
- Feeding additional image to replace identified feature



Style Variances

Tested for how aesthetic style affects image output

- Applied IP-adapter onto workers generated in a certain artist's style

Takeaway:

- Bias reflected in particular styles

Dynamic Prompts

Weighted Prompt: A {58.9::caucasian|13.6::african american|1.3::native american|6.3::asian|0.3::pacific islander|19.1::latino|3::mixed race} construction worker, illustration

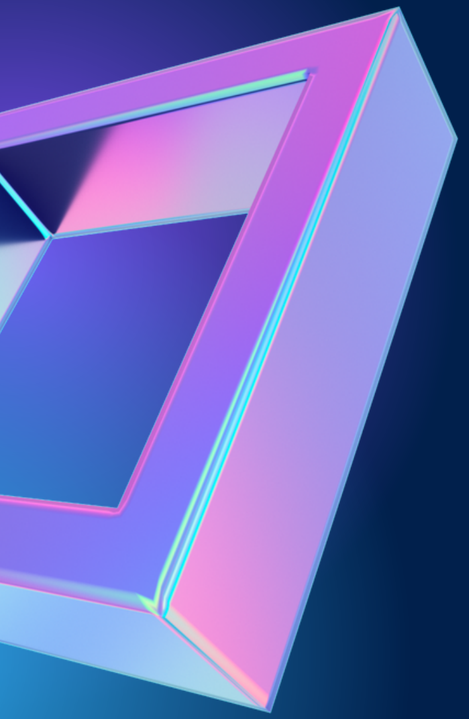
weighted

unweighted

Takeaway: weighted dynamic prompts produce results more similar to the ethnicity distributions in the U.S. Census. This might be a viable way for prompt engineering.

Learning Outcomes

- Just the starting point (negative prompting, Sola etc.)
- Sensitivity is key when training models!
  - Artist style
  - Adjectives
  - Environment
- “Fair and balanced” is subjective when it comes to training models



Thank You