AML Mid Project Presentation
Emotion Driven Image Synthesis

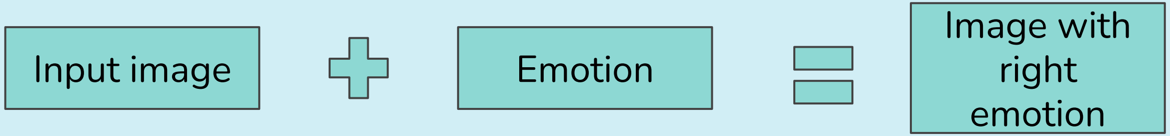
Swetha M - 23M0756
Harshvivek Kashid - 23M0762
Rashmi Kokare - 23M0785
Nikita - 23M0807

Problem Statement

Generated synthetic images lack emotions though they are realistic.

Some **challenges** are:

1. Latent diffusion models like stable diffusion **create random images** instead of starting from source image. Not able to edit images.
2. Some existing models (specially GAN) use **only style or color to change emotion**.



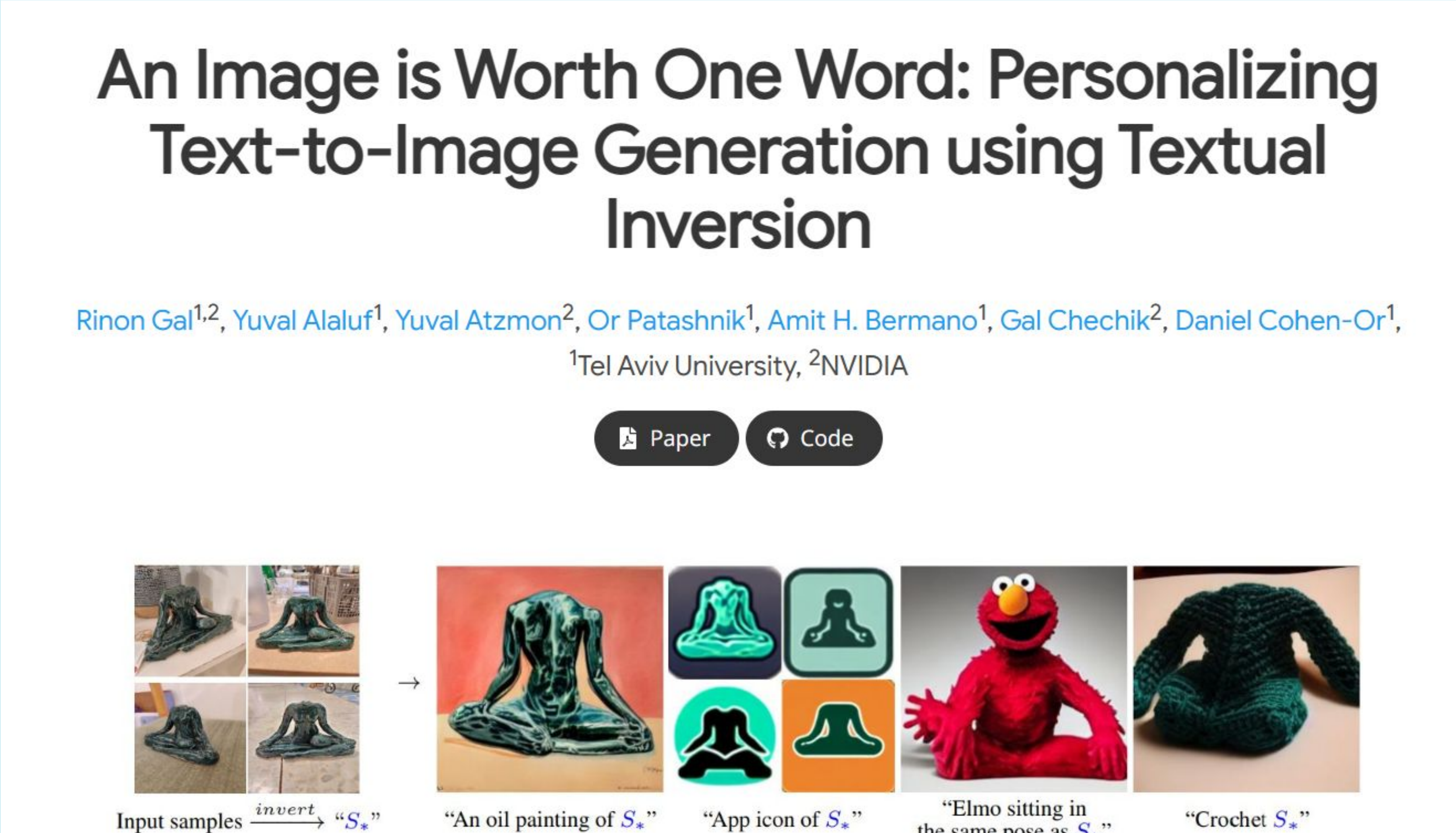
Input image

Emotion

Image with right emotion

Proposed Solution

1. To address this, we propose to use approach that integrates **Textual Inversion.**
2. Textual inversion is a few shot learning of an image concept.
3. Once the image concept is learnt, a token is added to the vocabulary which can be used in the text-to-image applications



An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion

Rinon Gal[1,2], Yuval Alaluf[1], Yuval Atzmon[2], Or Patashnik[1], Amit H. Bermano[1], Gal Chechik[2], Daniel Cohen-Or[1],

[1]Tel Aviv University, [2]NVIDIA

Paper
Code
Input samples $\xrightarrow{invert}$ “S_*”
→
“An oil painting of S_*”
“App icon of S_*”
“Elmo sitting in the same pose as S_*”
“Crochet S_*”

Existing methods

Methods: Stable diffusion based, StyleGAN, VAE based and Emogen, Asymmetric reverse process in diffusion

- Emogen can generate images with a certain emotion - but not add or include emotion to an input image.

Textual inversion based:

- **LoRA** - Learn a text for a given image which can be used in prompts for future images (personalised image editing).
- **Dreambooth** - Subject-Driven Generation which few shot trains an image and uses it to generate multiple contexts of the same subject.

Literature survey

Paper	**Main idea**
EmoGen: Emotional Image Content Generation with Text-to-Image Diffusion Models	Map the emotion space to CLIP space to generate a more diverse set of images
Textual inversion	Learn visual contexts and use them in future prompts
StyleGAN	Color and style transfer to change the emotion

Code Survey

Code	Module usable	Link
Asyrp	**Greater part of learning from bottleneck of UNET can be used.** Textual inversion to be incorporated	asyrp
StyleGAN	Helped study latent structure manipulation	StyleGAn
Textual inversion	Hugging face also has starters for textual inversion but not useful for custom data	Textual inversion
Emogen	Helped understand emotion space vs CLIP	Emogen

Datasets Explored

1. **Finding Emo** - https://arxiv.org/pdf/2402.01355.pdf https://gitlab.com/EAVISE/lme/findingemo/-/tree/main?ref_type=heads
   a. 25k images
   b. Plutchik's discrete Wheel of Emotions (PWoE)
   c. **Very useful because it has complex backgrounds and realistic images**
2. **Emoset** - https://github.com/JingyuanYY/EmoSet/blob/main/EmoSet.py
   a. With 8 emotion categories (amusement, anger, awe, contentment, disgust, excitement, fear, and sadness
   b. 118K images
   c. **Not very useful dataset**

Implementation Done

- We have implemented **Textual inversion** to add emotions to existing images.
- We have studied the codebase of existing **GAN, VAE based methods** for emotion induction.
- We have studied the hugging face models for **CLIP, Stable diffusion, text to image pipelines** etc. and tested small experiments.
- We have trained custom data on **ASYRP repo** and noted down observations.

Observations and Preliminary Results

- **Text inversion** for existing concept libraries work fine but requires effort for custom data.
- QUALITY measured from a Emotion classifier
- **ASYRP** - Not able to reproduce the results when training on custom datasets.
- **Hugging face models** generate random images not starting from source image.
- **StyleGAN and VAE based models explored:** Has some limitations like distorted, blurry images and inability to handle data different from training data.

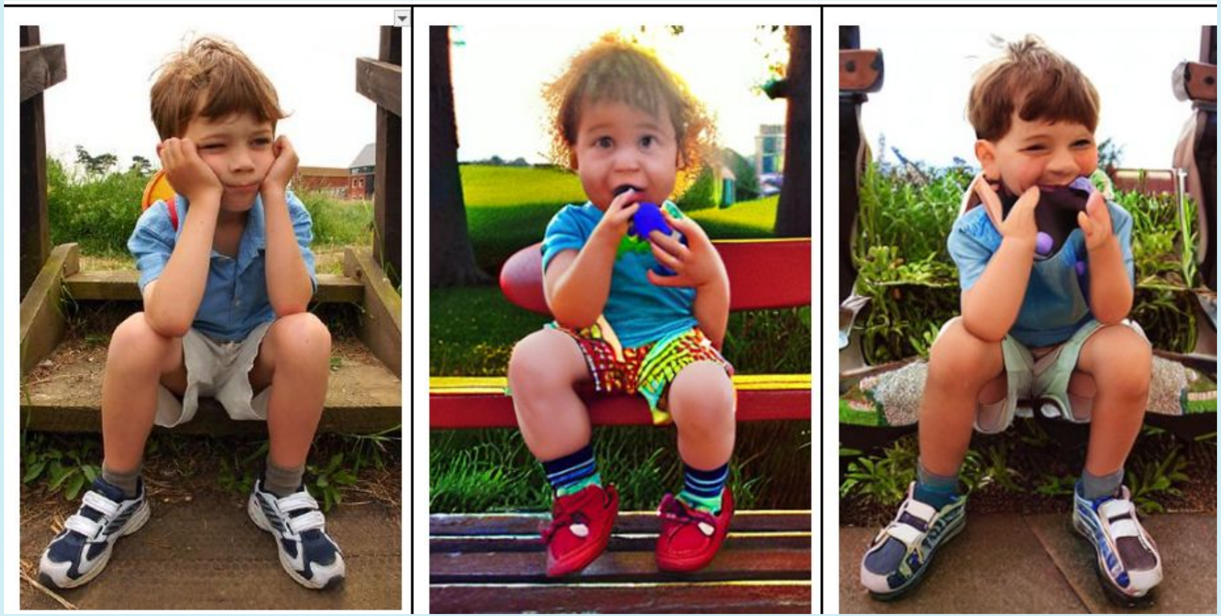
Custom trained

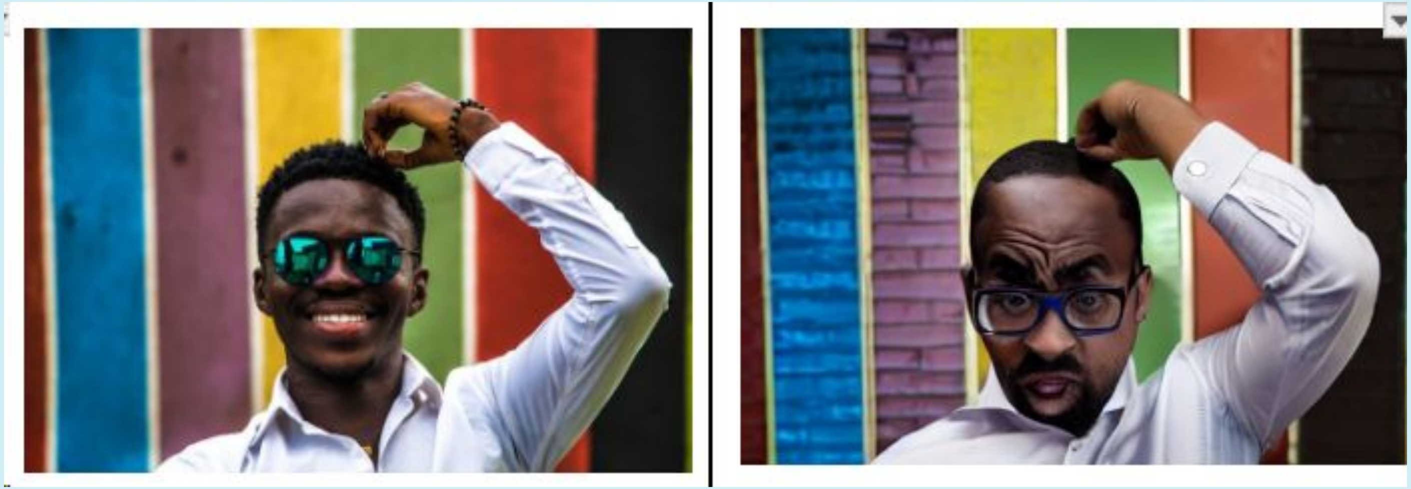
Concepts already available

Textual inversion output

Preliminary Results

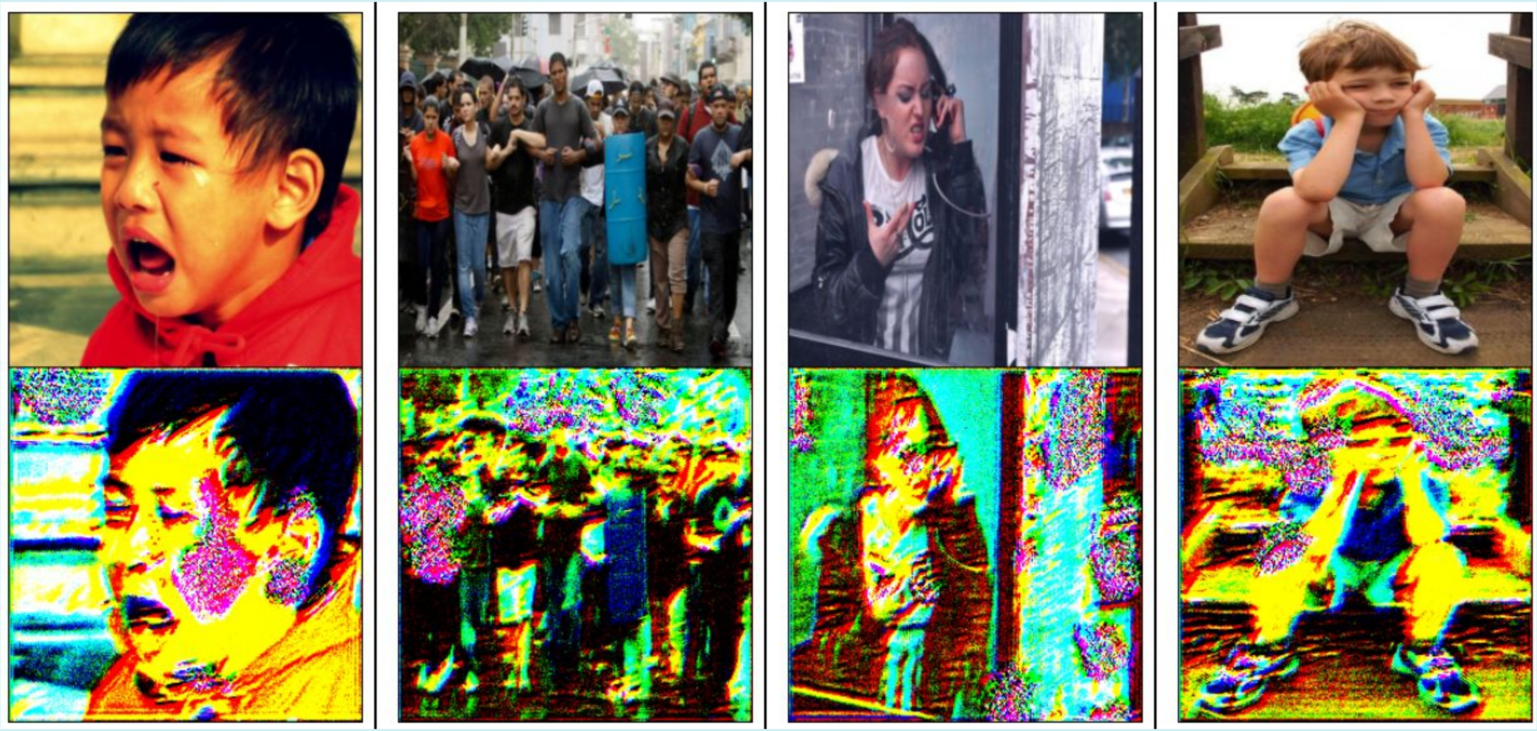
- **Hugging face models** & basic text inversion. (parameters - more noise or less noise)
-



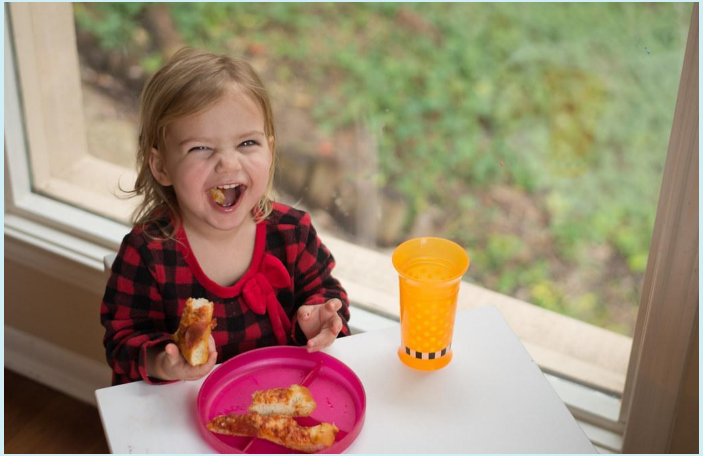


Difficulty moving from very happy to sad/angry

ASYRP (training on custom data)



Not comparable to Stability AI API



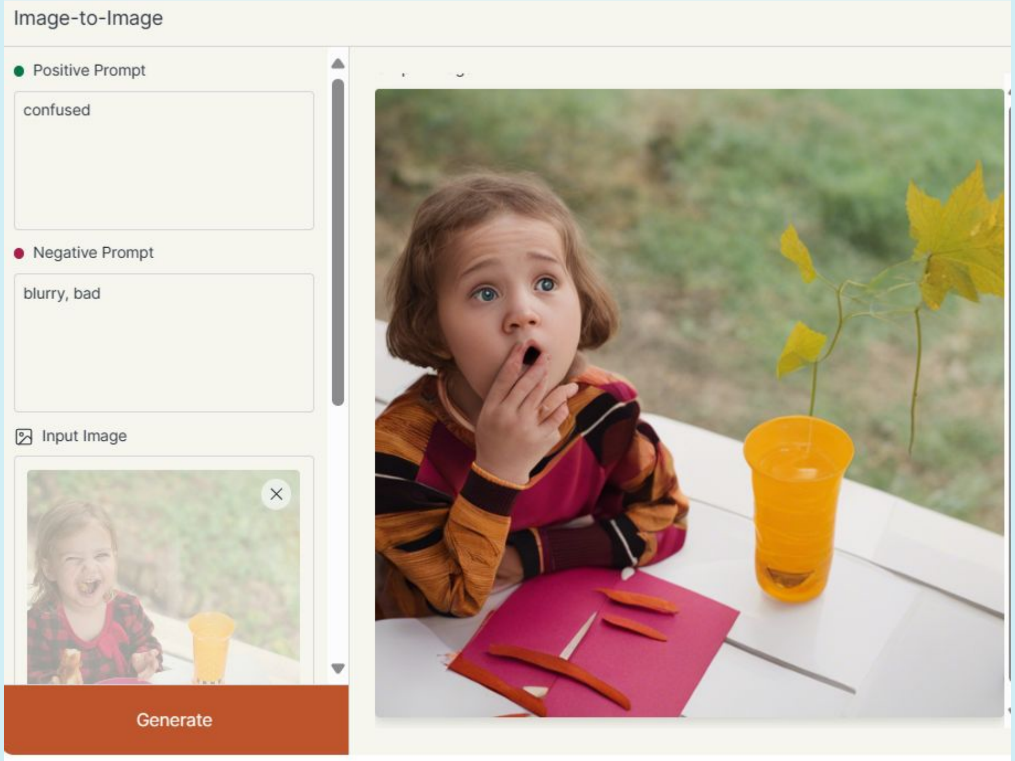


Image-to-Image

● Positive Prompt

confused

● Negative Prompt

blurry, bad

Input Image

Challenges Encountered

- **Random generation** happens since diffusion has learnt denoising from **random seed.**

- **Blurring of images** and making the features worse than original image.

- Not interpret **complex backgrounds of people** and expects **portrait only** input images.

- Strong degree of change (very happy to angry) is getting distorted outputs

Thank You!