Stable Diffusion Metrics

6. February 2023

Learning from Images, Prof. Hildebrand

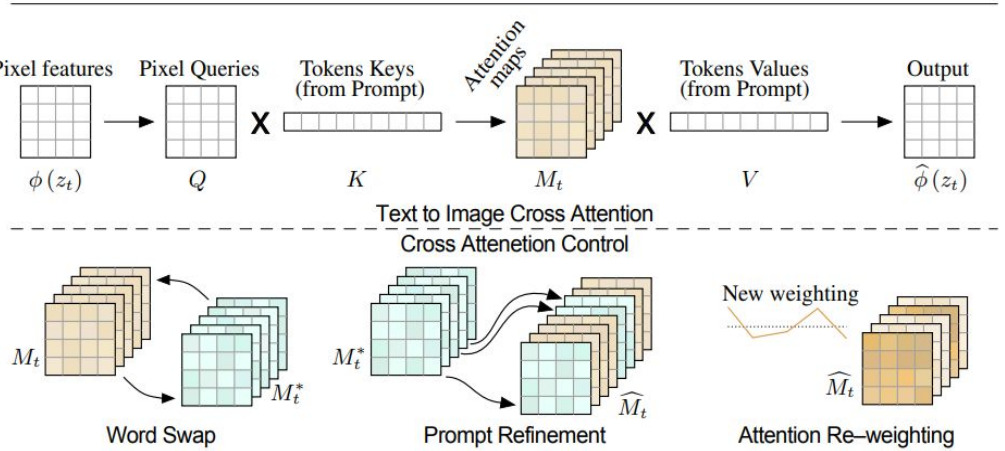
Amin Suaad, Dennis Fast, Manuel Freistein

Prompt-to-Prompt Image Editing

- Editing is challenging for generative models.
- Often small change in the text can lead to a massive change in the image.
- Cross attention layers are key to establish the relation between the image and each word of the prompt.
- Localized editing by replacing a word, global editing by adding a specification, and even controlling the extent to which a word is reflected in the image.

CROSS-ATTENTION CONTROL

- The visual and textual embeddings are combined using cross-attention layers, which generate attention maps for each token.
- When swapping a word in the prompt, we inject the source image maps Mt, overriding the target maps M∗t. When adding a refinement phrase, we inject only the maps that correspond to the unchanged part of the prompt. The corresponding attention map is reweighted to amplify or attenuate the semantic effect of a word.

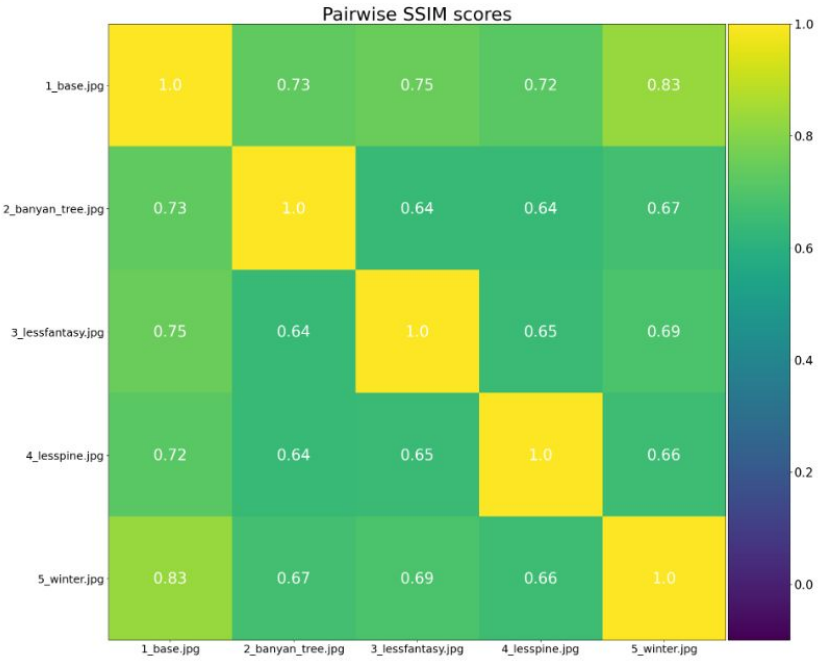


Without Cross Attention Control

- 5 generated images without cross attention control.
- Base prompt: "A fantasy landscape with a pine tree in the foreground and a red sun setting in the distance, trending on artstation"
- Structural similarity index is low.

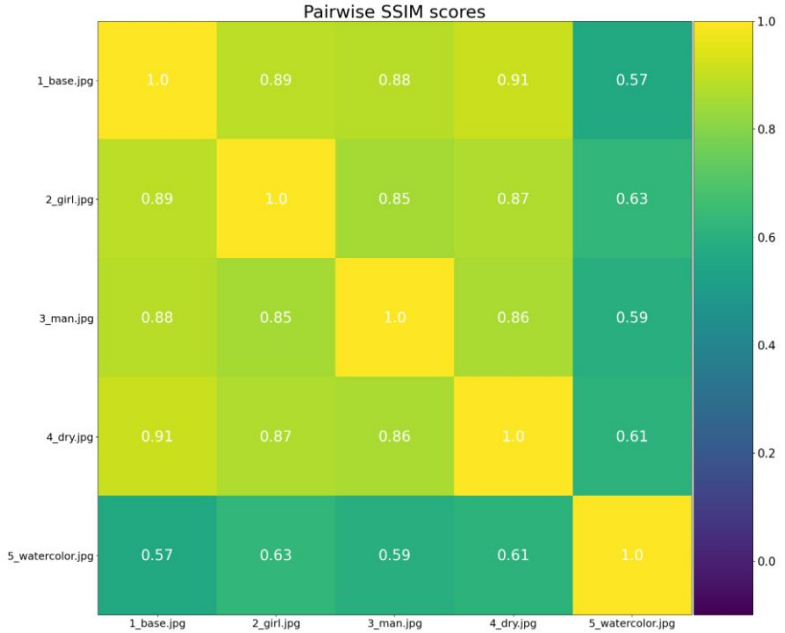
With Cross Attention Control (Example-1)

- 5 generated images with cross attention control.
- Base prompt: "A fantasy landscape with a pine tree in the foreground and a red sun setting in the distance, trending on artstation"
- Structural similarity index is high.



With Cross Attention Control (Example-2)

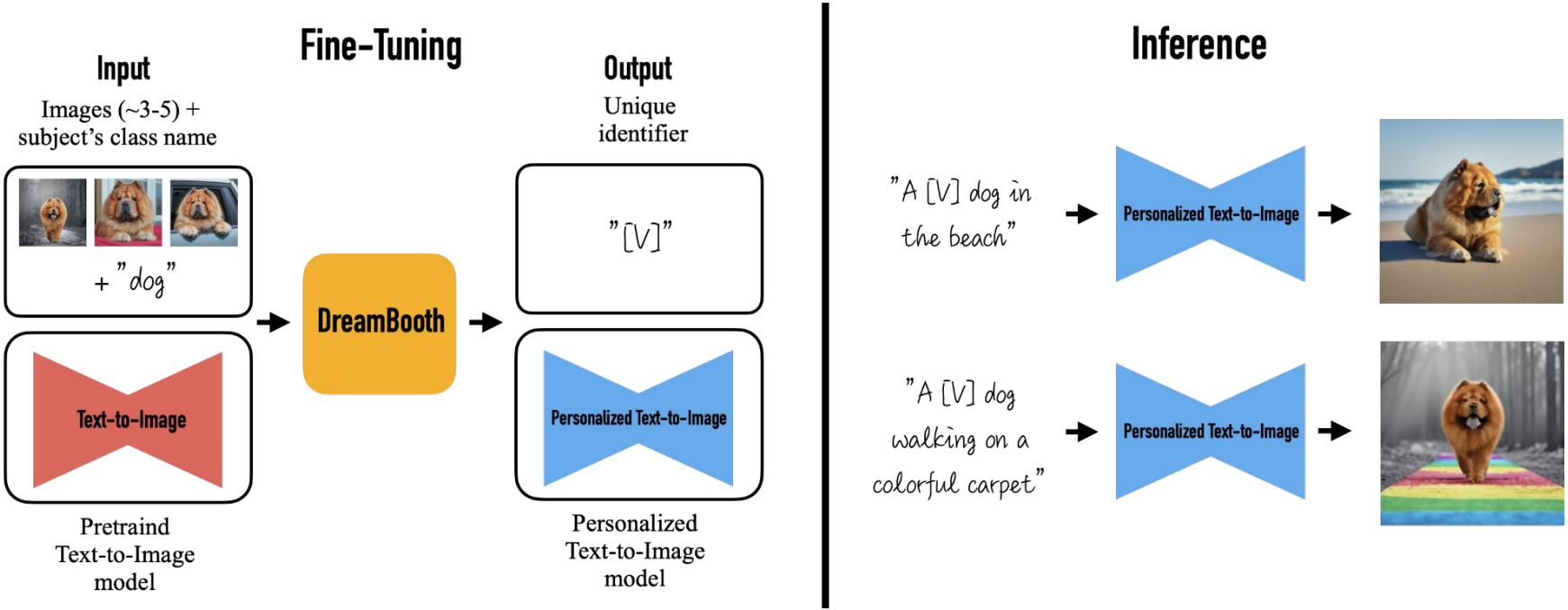
- 5 generated images with cross attention control.
- Base prompt: "A young boy playing in a field, on a hill overlooking a green valley"
- Structural similarity index is high.



Stable Diffusion as Time Machine

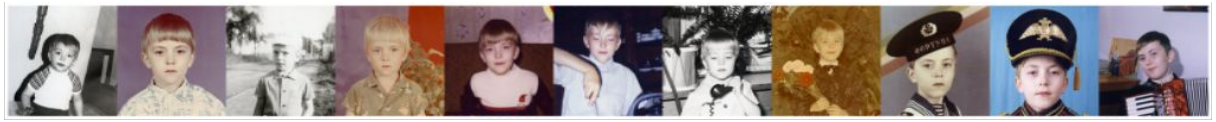
Task in a nutshell

DreamBooth (Recap)

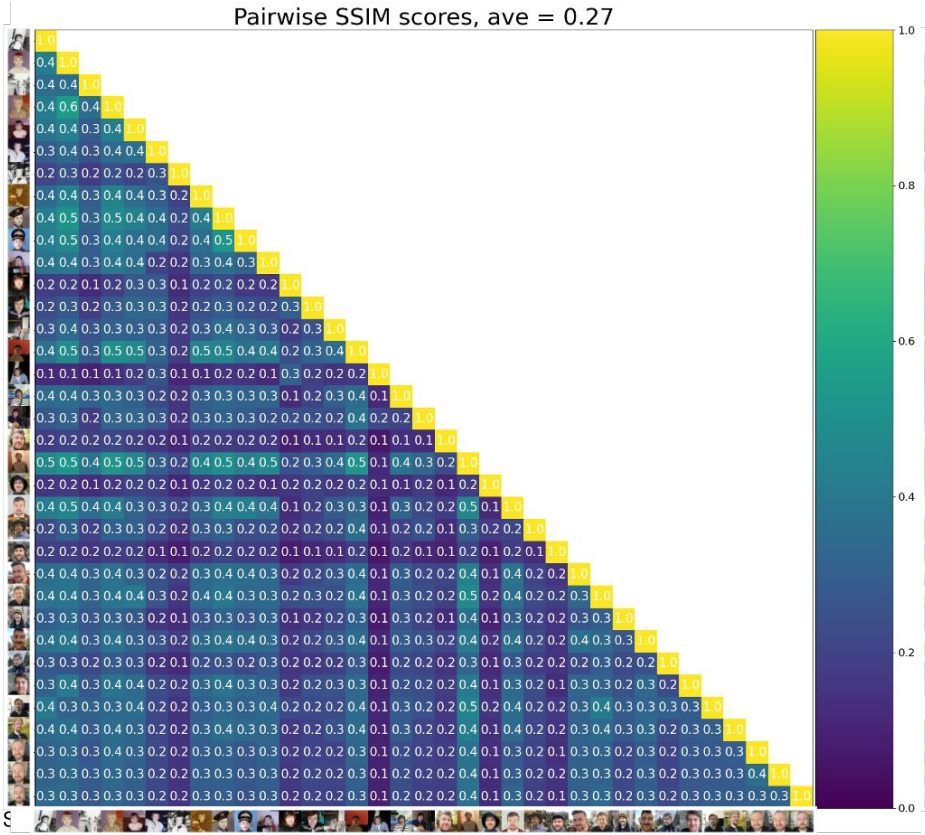


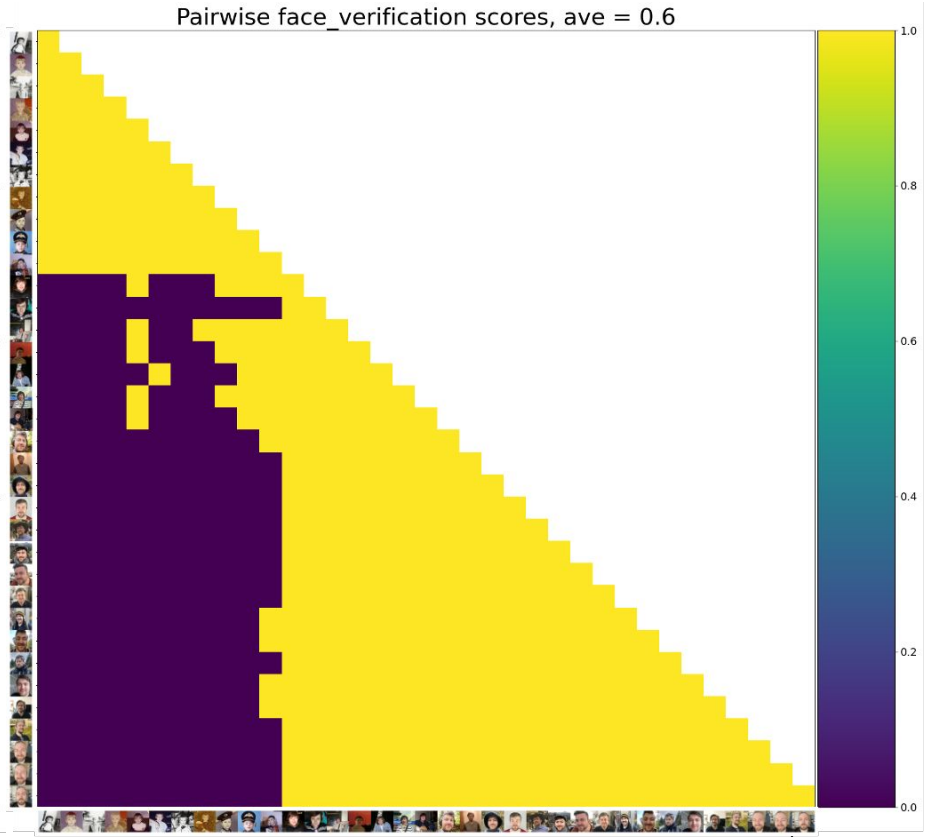
Input & Preprocessing

- **Initial training dataset**: 122 face images of the same person over the life time (36 years)
- **Preprocessing steps**:
  - assess the image variation using SSIM score (the lower the better)
  - split into 4 age classes (child, young, adult, today)
  - face verification using DeepFace python library (VGG-Face model)
- **Final training datasets:** child (11 images, age: 0 – 15), adult (24 images, age: 15 +)



Trainset Metrics





Generative Models & Output

Six models in total: (Stable Diffusion 1.5 + DreamBooth)

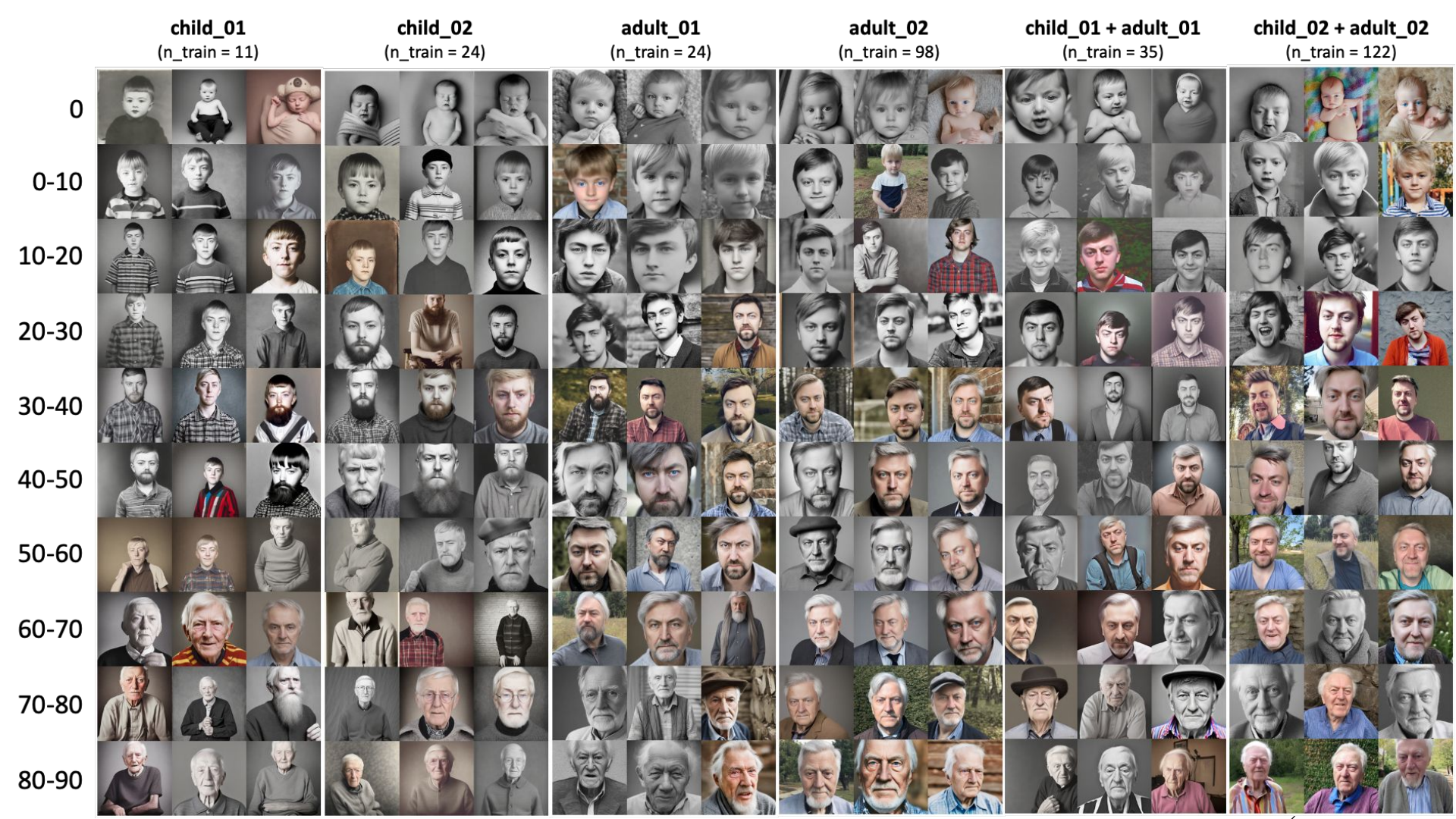
- verified faces only: 1) child, 2) adult, 3) child+adult
- all images: 4) child 5) adult 6) child+adult

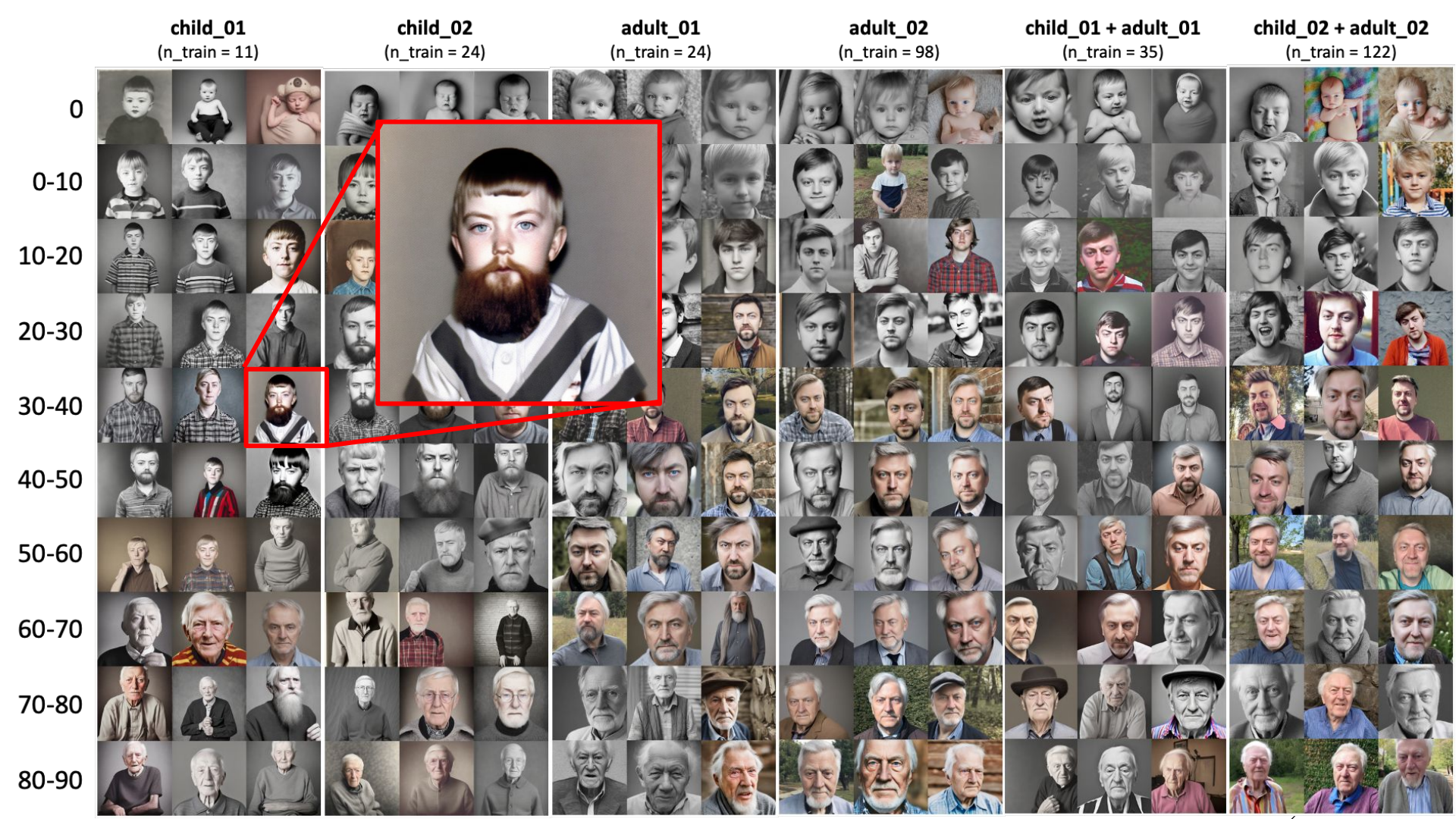
Prompt: 'A photo of {model} as {age}, expressive face, highly detailed, sharp focus, natural light'

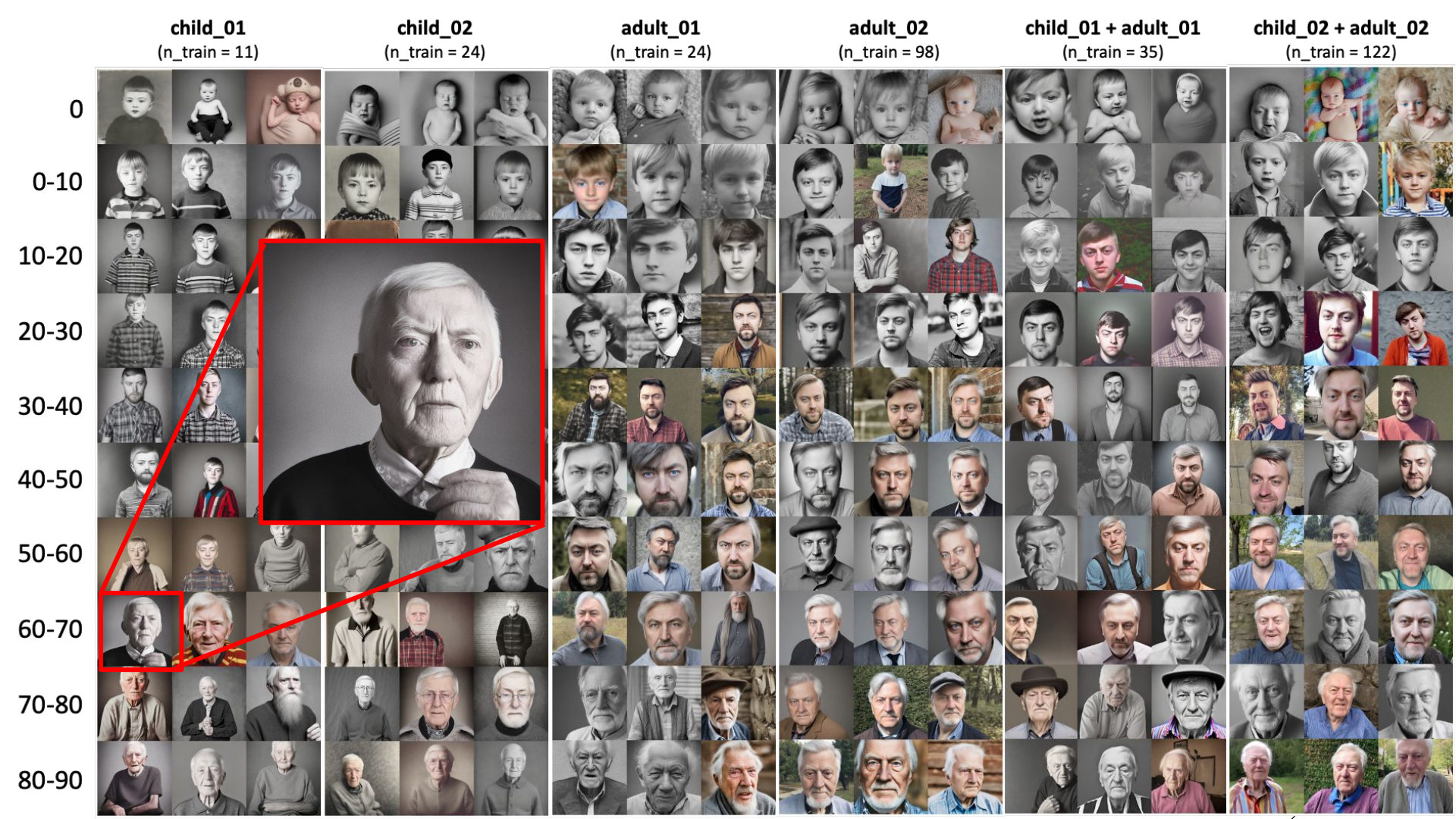
age_list = ['a baby', 'a five years old child', 'a fifteen years old teenager', 'a young man in his twenties',

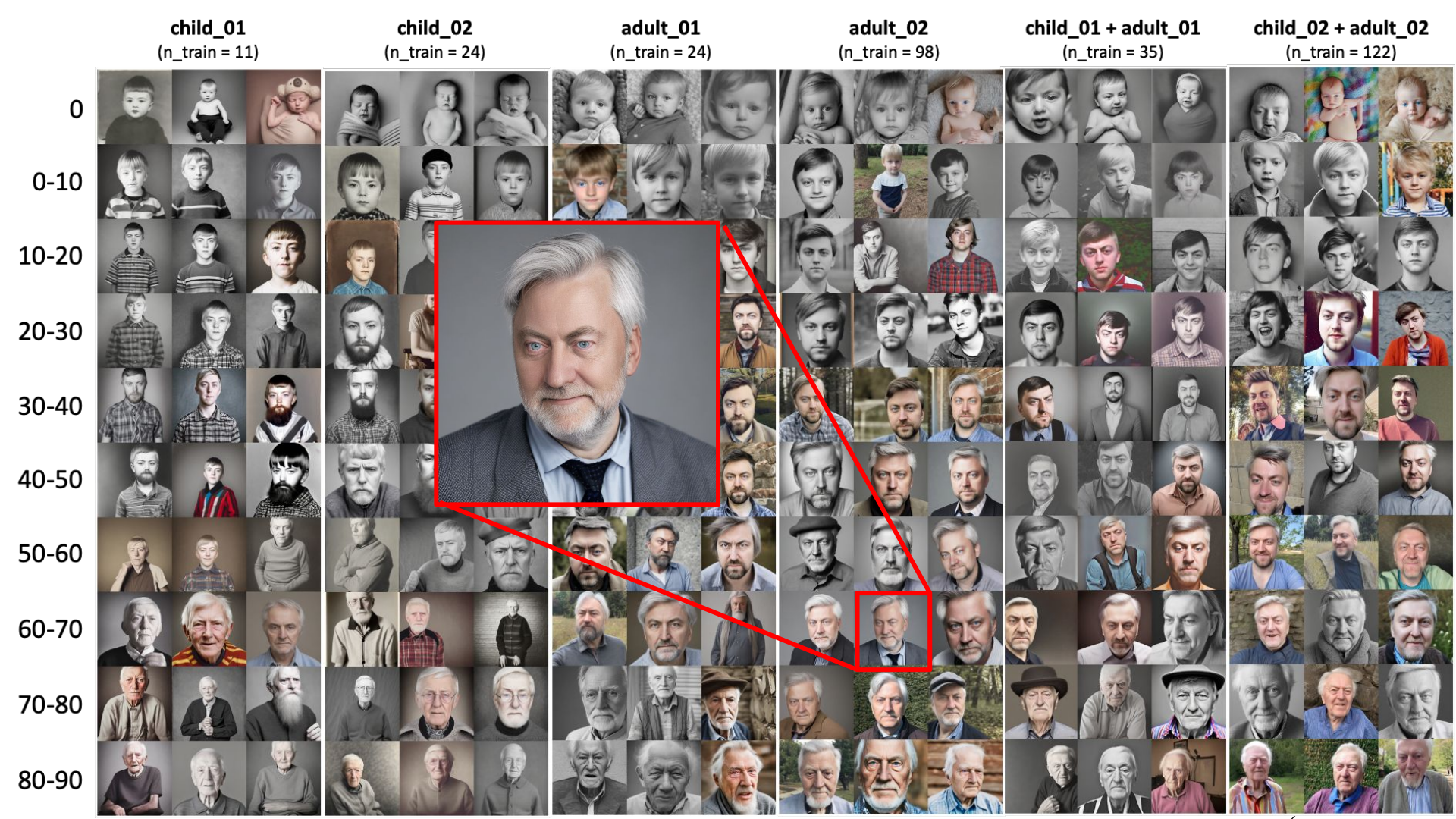
'an adult man in his thirties', 'an exhausted adult man in his forties with grayish hair',

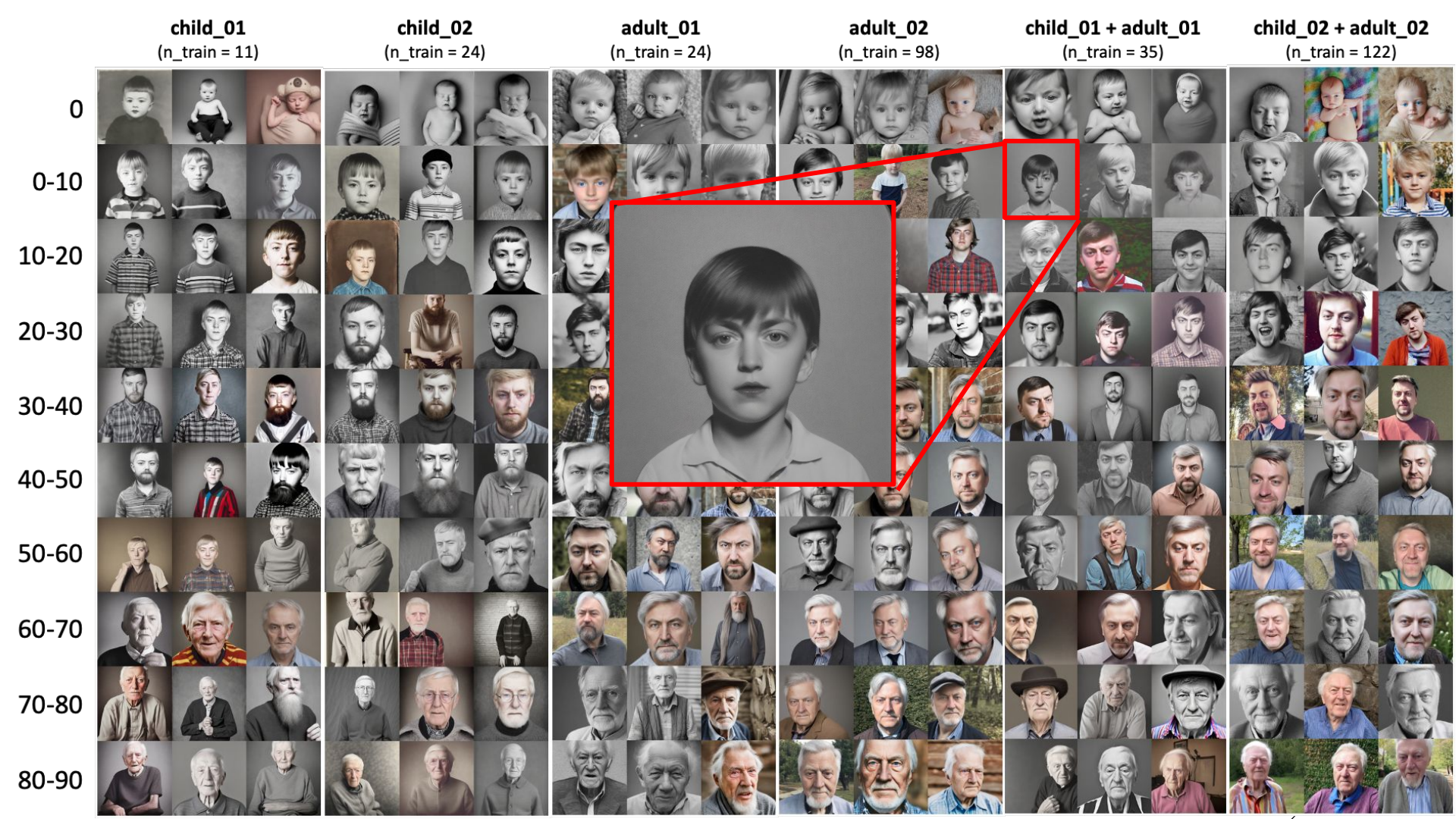
'an adult man in his fifties with gray hair', 'a senior in his sixties', 'a senior in his seventies', 'an elderly senior in his eighties']

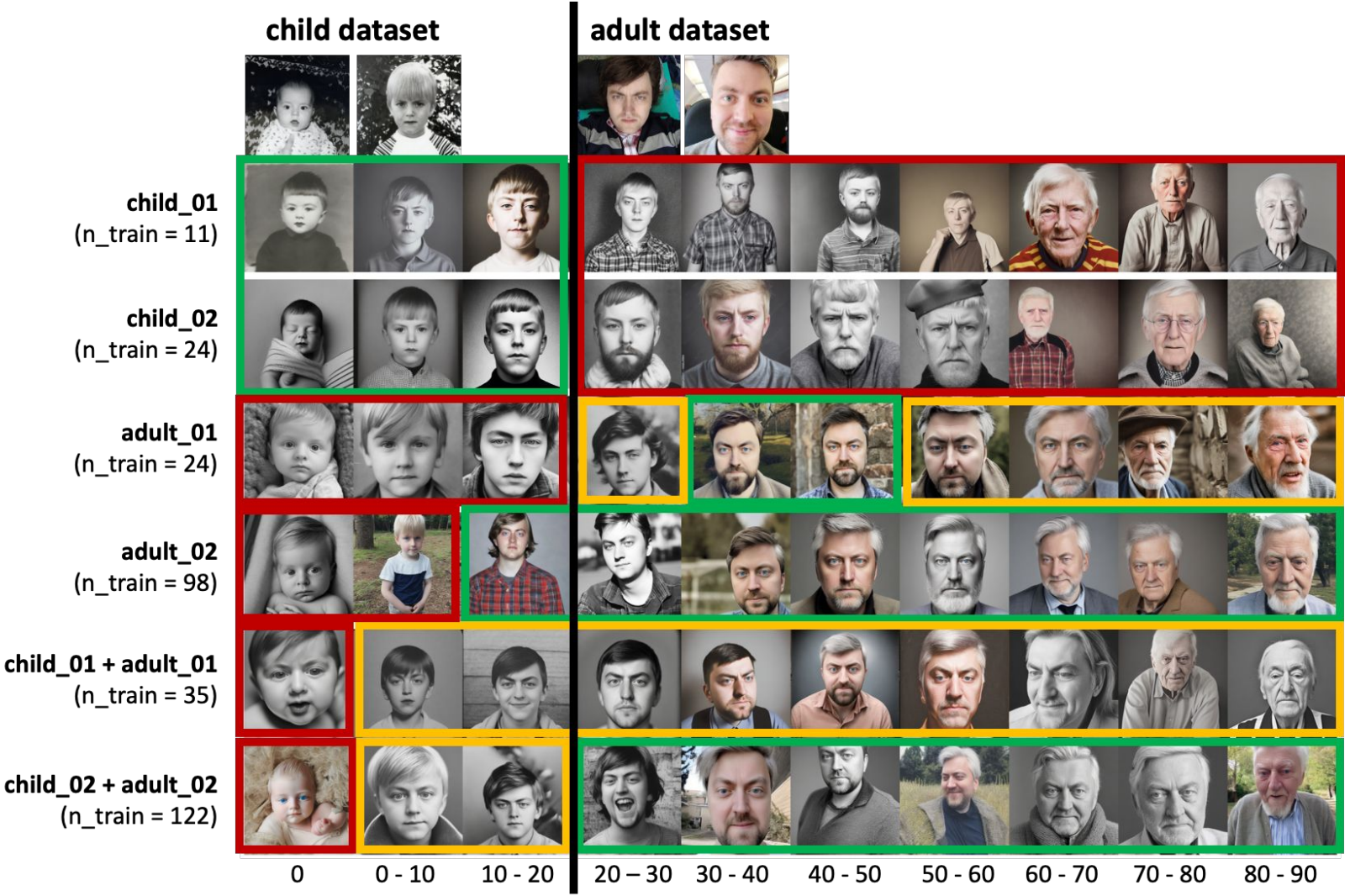

child_01
(n_train = 11)
child_02
(n_train = 24)
adult_01
(n_train = 24)
adult_02
(n_train = 98)
child_01 + adult_01
(n_train = 35)
child_02 + adult_02
(n_train = 122)
0
0-10
10-20
20-30
30-40
40-50
50-60
60-70
70-80
80-90


child_01
(n_train = 11)
child_02
(n_train = 24)
adult_01
(n_train = 24)
adult_02
(n_train = 98)
child_01 + adult_01
(n_train = 35)
child_02 + adult_02
(n_train = 122)
0
0-10
10-20
20-30
30-40
40-50
50-60
60-70
70-80
80-90


child_01
(n_train = 11)
child_02
(n_train = 24)
adult_01
(n_train = 24)
adult_02
(n_train = 98)
child_01 + adult_01
(n_train = 35)
child_02 + adult_02
(n_train = 122)
0
0-10
10-20
20-30
30-40
40-50
50-60
60-70
70-80
80-90

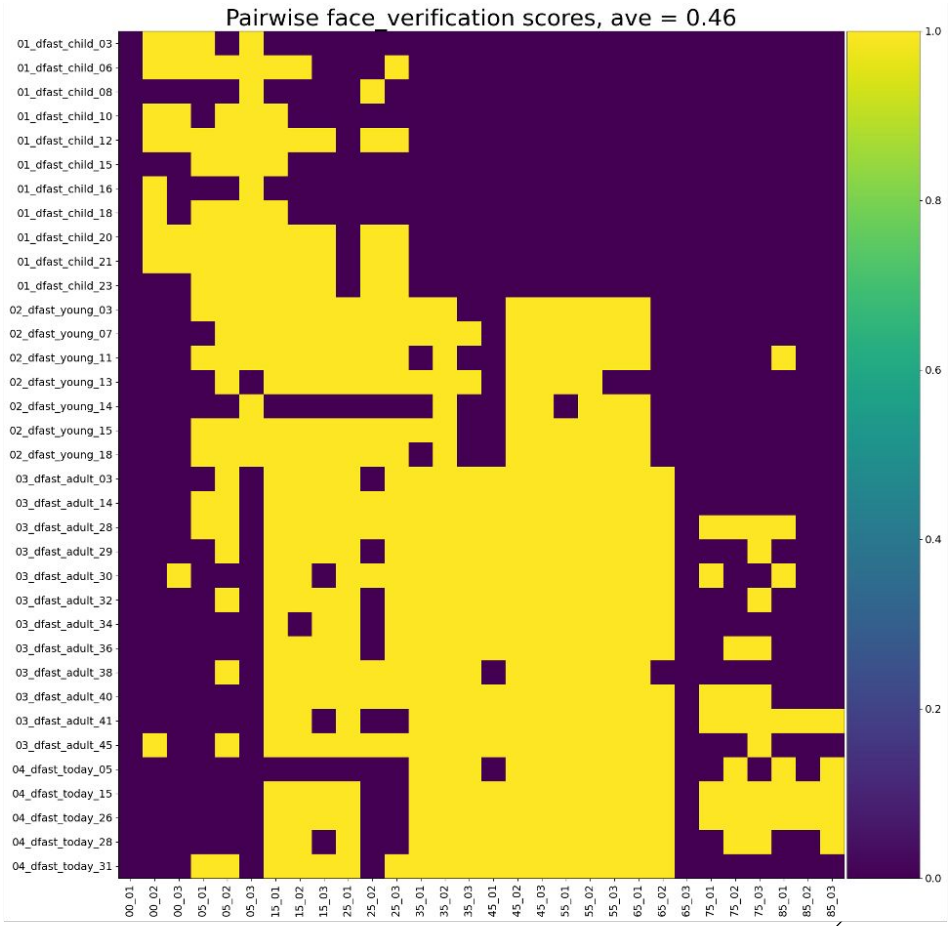

child_01
(n_train = 11)
child_02
(n_train = 24)
adult_01
(n_train = 24)
adult_02
(n_train = 98)
child_01 + adult_01
(n_train = 35)
child_02 + adult_02
(n_train = 122)
0
0-10
10-20
20-30
30-40
40-50
50-60
60-70
70-80
80-90


child_01
(n_train = 11)
child_02
(n_train = 24)
adult_01
(n_train = 24)
adult_02
(n_train = 98)
child_01 + adult_01
(n_train = 35)
child_02 + adult_02
(n_train = 122)
0
0-10
10-20
20-30
30-40
40-50
50-60
60-70
70-80
80-90


child dataset
adult dataset
child_01
(n_train = 11)
child_02
(n_train = 24)
adult_01
(n_train = 24)
adult_02
(n_train = 98)
child_01 + adult_01
(n_train = 35)
child_02 + adult_02
(n_train = 122)
0
0 - 10
10 - 20
20 – 30
30 - 40
40 - 50
50 - 60
60 - 70
70 - 80
80 - 90

Evaluation Metrics

Data	Pairwise SSIM (averaged)	Face_verification (averaged)
Train images	0.29	0.64
Generated images	**0.39**	0.43
Train images / generated images	0.32	0.35



Overall best model: adult_02

Possible applications

- **Family resemblance** (curious parents, images of both parents / relatives at different ages)
- **Aging simulation** (police investigations, predict the aging process of the suspect)
- **Facial reconstruction** (forensic investigations, likeness of a person based on old photos)
- **Film and video game characters** (entertainment, digital characters based on reference photos)
- **Marketing and advertising** (generate images of customers for advertisements and promotions)

How well does Textual Inversion recreate style?

Textual Inversion teaches Stable Diffusion new 'words' ("objects" or "styles") in the embedding space of the pipeline's text encoder which can be used within text prompts to achieve very fine-grained control of the resulting images.

Conventional Metrics such as Structural Similarity Index Measure or Fréchet Inception Distance compare generated images to original undistorted images.

Reconstruction and Editability were used to gauge Textual Inversion's ability to recreate style by the authors of the original paper *An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion*. They compared the mean spatial distance between 64 generated images and the training images ("Reconstruction") as well as comparing the mean spatial distances of different prompts with the placeholder to the same prompt without the spaceholder ("Editability") using cosine similarity.

REAL, FAKE OR GENERATED?

Paul Cézanne (1839-1906) **Dataset**

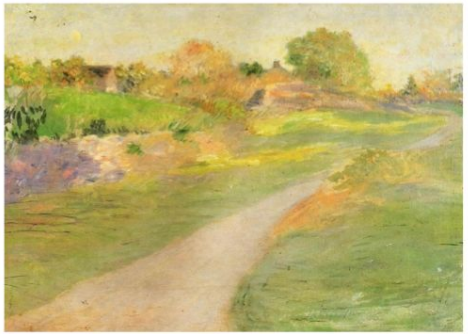
89 authentic Cézanne landscape paintings

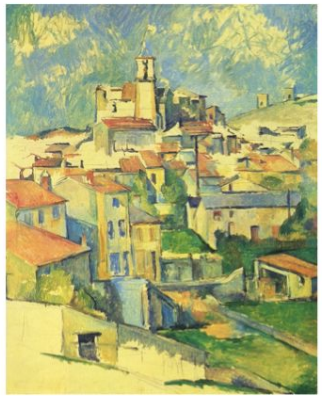
68 hand-painted replicas and forgeries of authentic Cézanne landscape paintings

88 Stable Diffusion 1.5 generated Cézanne landscape paintings with guidance scale 8 (w/o textual inversion)

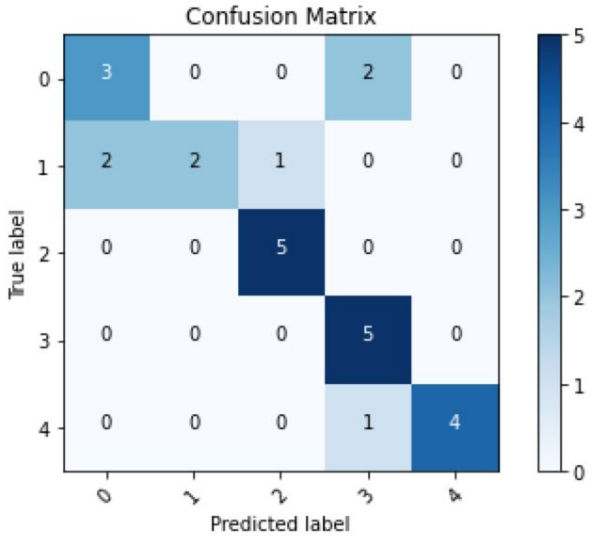
93 Stable Diffusion 1.5 generated Cézanne landscape paintings with guidance scale 0-1 (w/o textual inversion)

94 Impressionist landscape paintings from WikiArt dataset

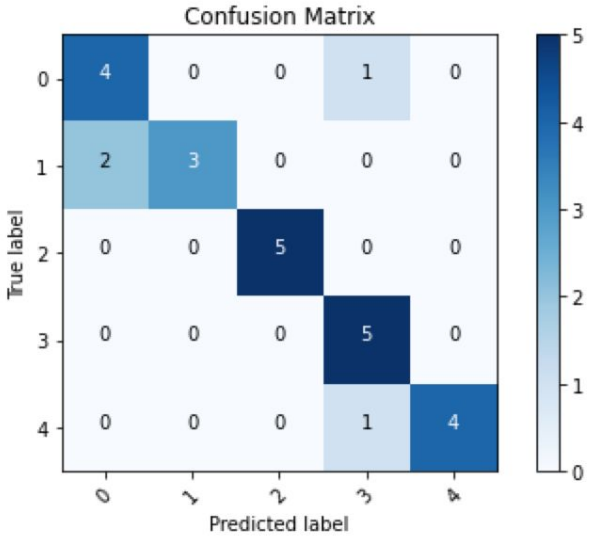




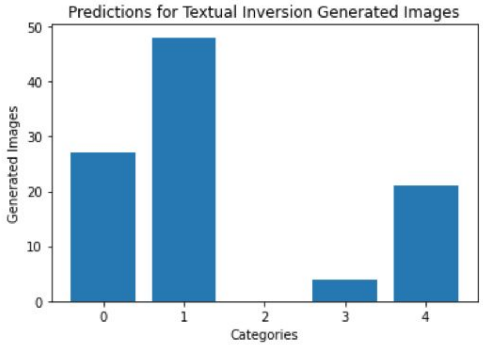
EfficientNet B7
(66.7M parameters)



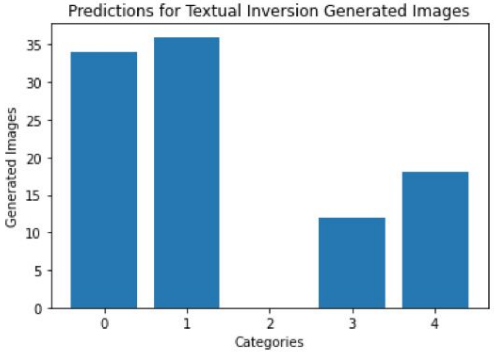
EfficientNet V2L
(119.0M parameters)



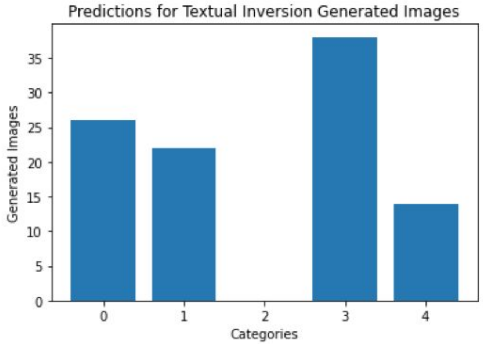
"painting in the style of **<Cezanne>**"



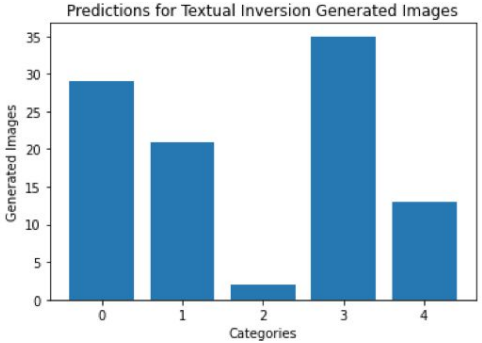
”painting of the Provence in the style of **<Cezanne>**"



”landscape painting in the style of **<Cezanne>**"



”painting of Mont Saint Victoire in the style of **<Cezanne>**"



Improvements

-more finetuned textual inversion training

-more training data

-wider array of prompts

-hyperparameter tuning

-combining textual inversion and Dreambooth

-better CNN training/testing dataset

-curation

-diversity

-size

-using image tiles instead of full images

-CNN model choice and hyperparameter tuning

Thank you for your attention!