Interactive Avatar Image Manipulation with Unconstrained Natural Language Instruction using Source Image Masking

shinagawa.seitaro.si8@is.naist.jp

Seitaro Shinagawa*1*2 Koichiro Yoshino*1*3
Sakriani Sakti*1 Yu Suzuki*1 Satoshi Nakamura*1*2

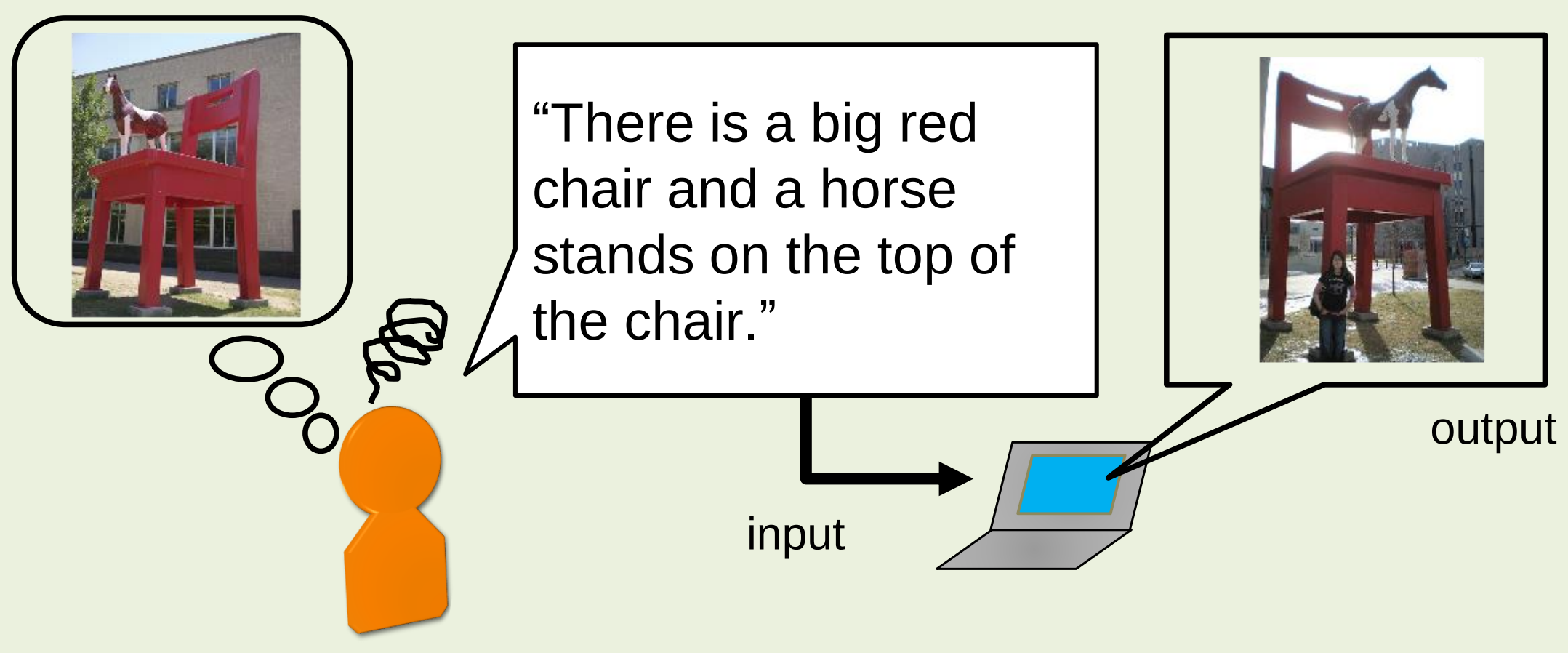
*1 Nara Institute of Science and Technology
*2 RIKEN, Center for Advanced Intelligence Project AIP
*3 PRESTO, Japan Science and Technology Agency

Summary

What is an easy way to get a desired image?

- **Image retrieval**: the image should be available in database
- **Hand drawing**: requires much time and drawing skills

A potential way: image generation from natural language caption (Caption2image, cap2image) [Reed et al. 2016]

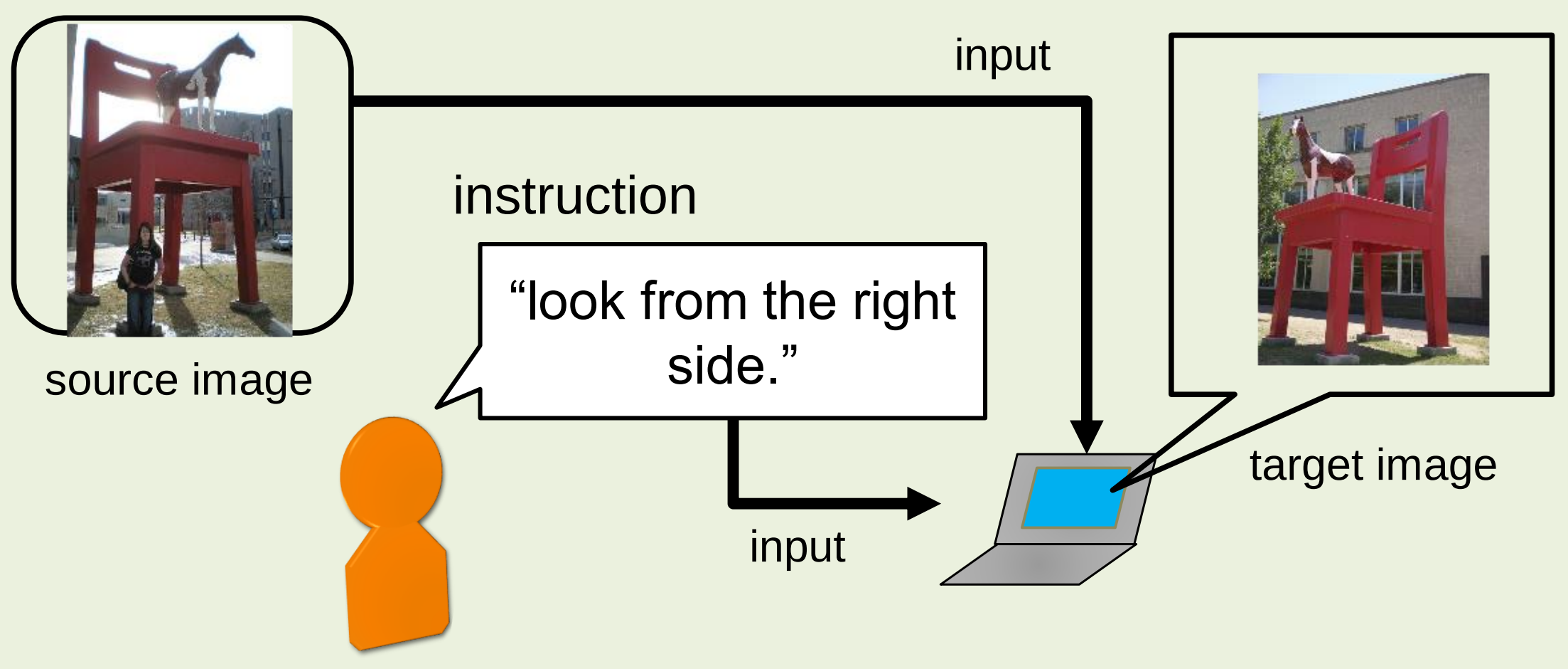

"There is a big red chair and a horse stands on the top of the chair."
input
output

However, cap2image is not good at modification

- Short text input satisfies many images
- Repetition of detailed long text input frustrates users

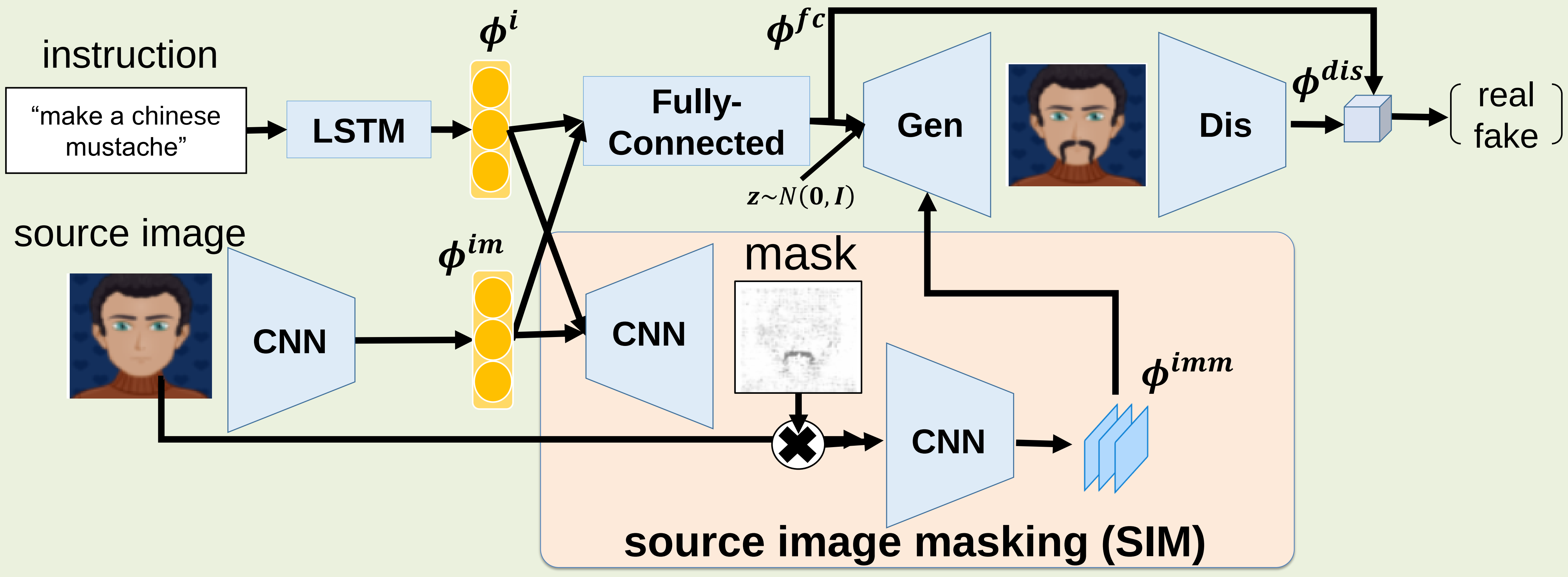
Main idea: **Image Manipulation with Instruction (IMI)**

natural language instruction represents the difference between source image and target image



- IMI make cap2imge interactive toward improving usability
- Source image masking (SIM) mitigates the unintentional change in generated images generated by IMI model

Baseline (w/o SIM) & proposed (w/ SIM) model



matching aware loss

(ϕ^{f},ϕ^{dis}) is ...
matched → real
mismatched → fake

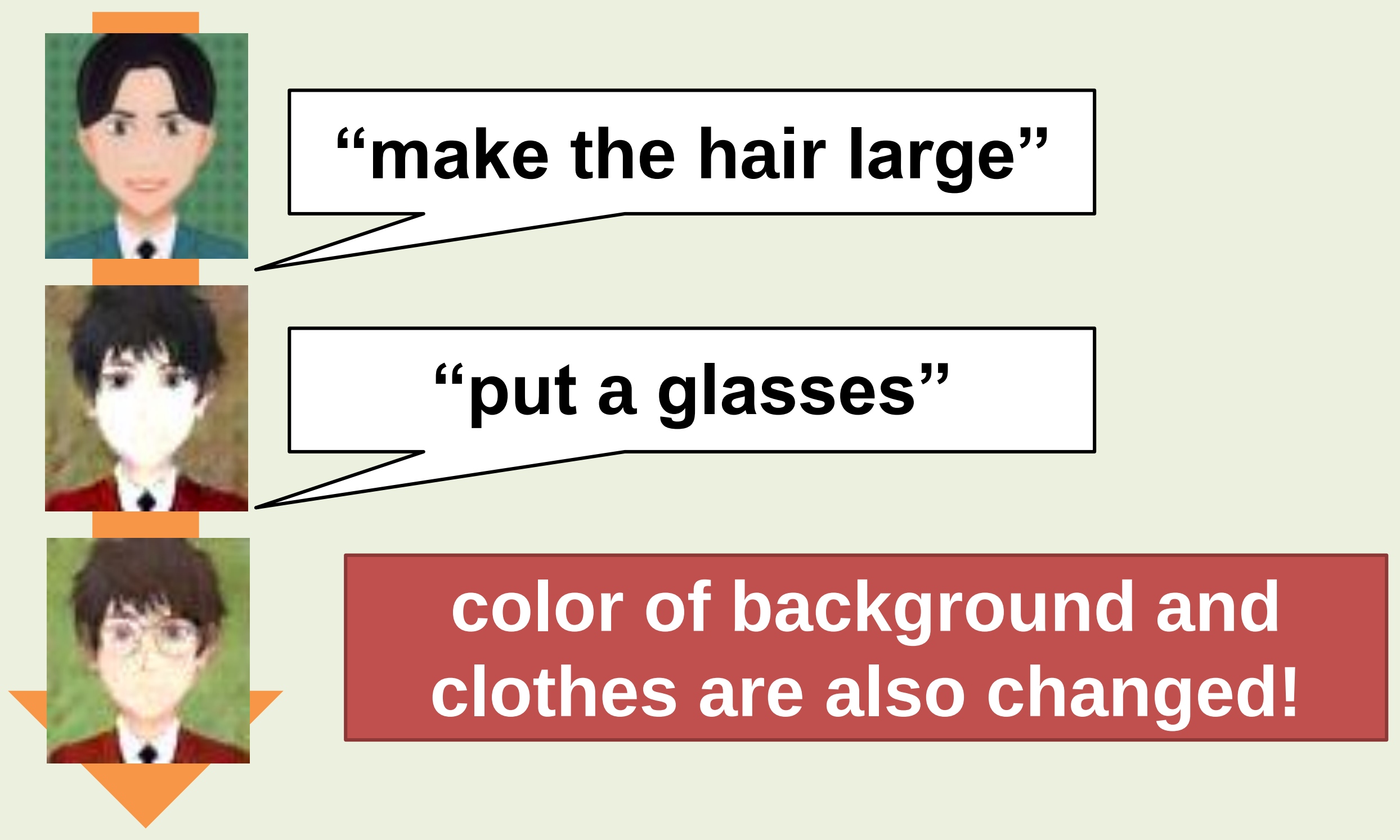
mask loss

= MSE

generated mask
target image
source image

Why source image masking?

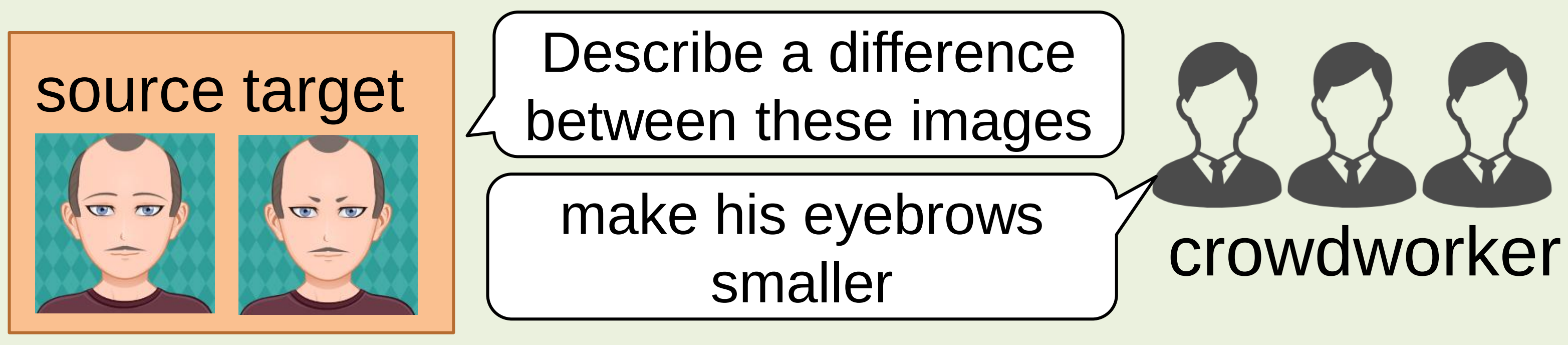
A naive model suffers from **not mentioned change**



We hypothesized covering source image with mask preserves not mentioned part in the instruction

Experiments and Discussion

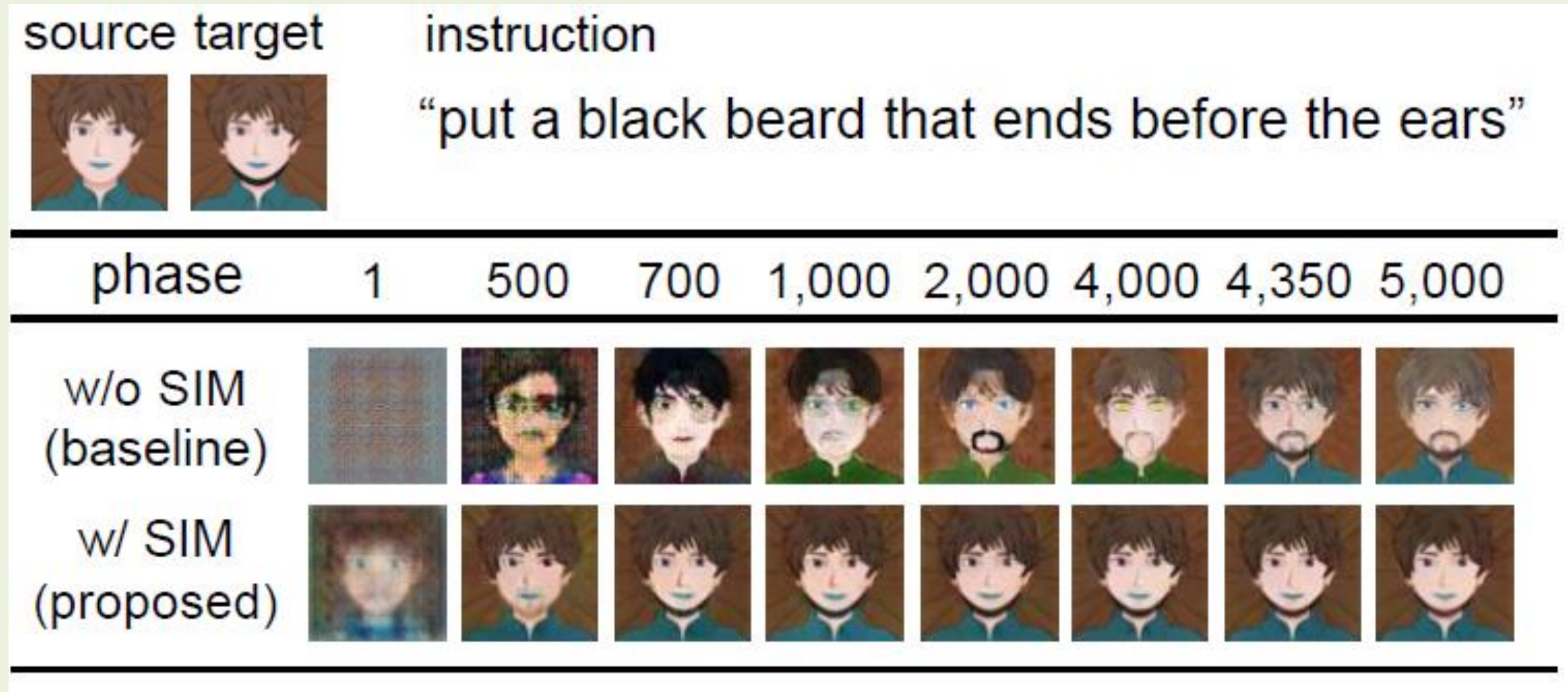
Data Collection


source target
Describe a difference between these images
make his eyebrows smaller
crowdworker

Experimental settings

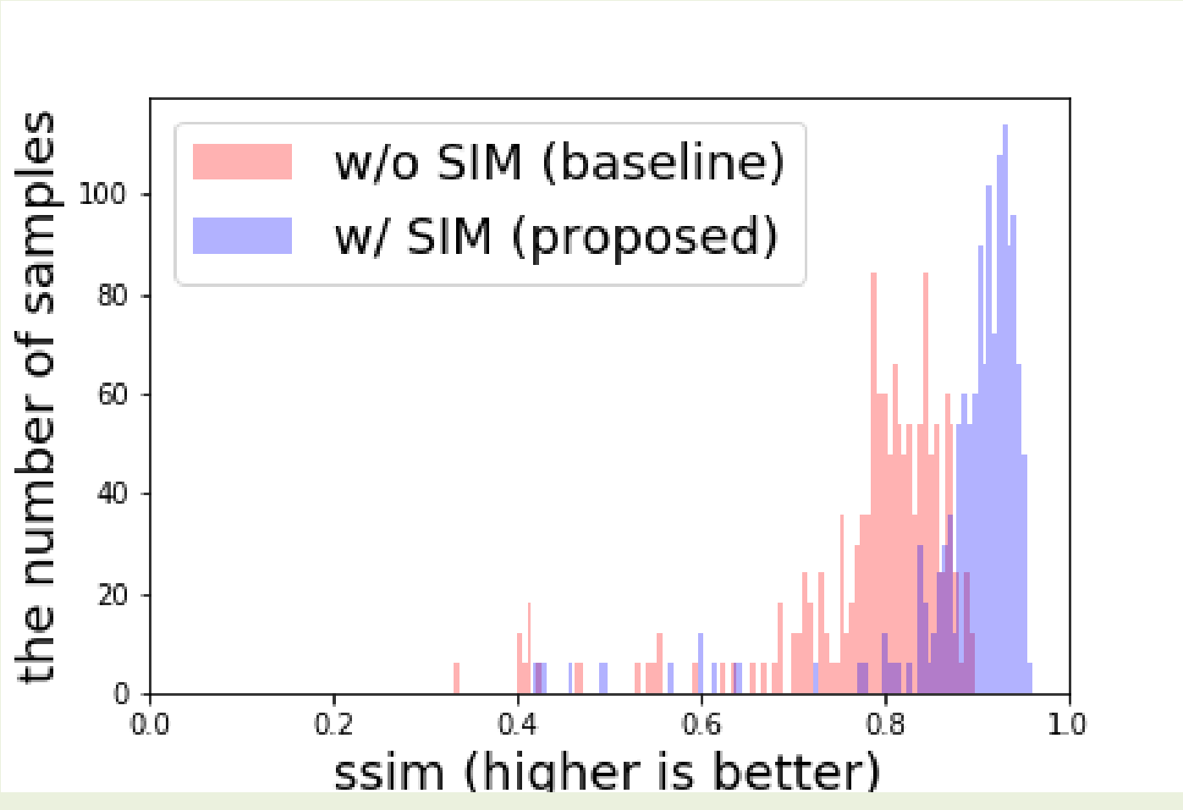
train:val:test = 4,296:230:230, random: 161,065
optimizer: Adam($\alpha = 2.0\times10^{-4}, \beta = 0.5$)
hidden: $\phi^{i}, \phi^{fc}: 128$, $\phi^{im}: 1024$, $\phi^{imm}: 512\times4\times4$
batch size: 64
vocabulary size: 1892
other option: feature matching loss to stabilize training

Generated images between w/ and w/o SIM


source target
instruction
"put a black beard that ends before the ears"
phase
1
500
700
1,000
2,000
4,000
4,350
5,000
w/o SIM (baseline)
w/ SIM (proposed)

w/ SIM model can generate a similar image to the target in early time

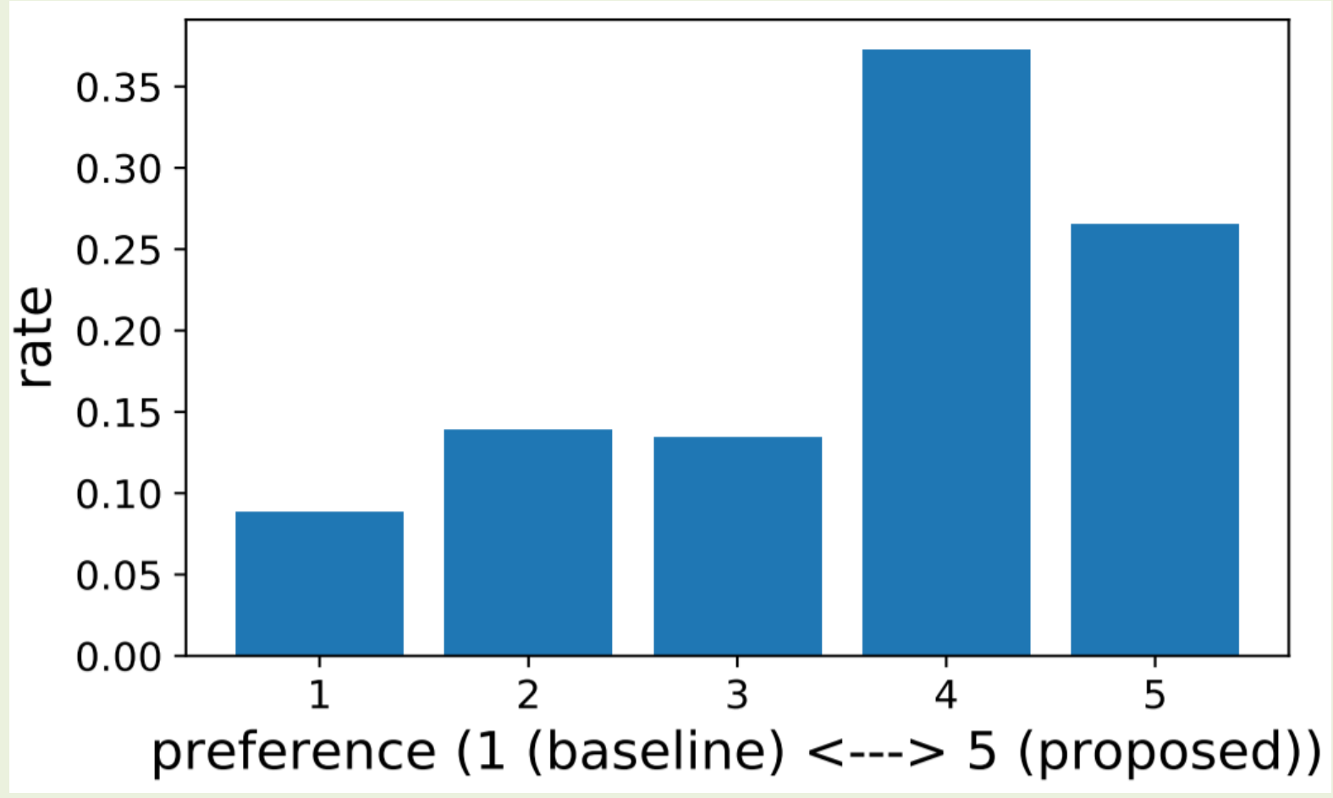
Objective Evaluation



SSIM histogram between generated and target image with whole test set

- **Score of w/ SIM is higher than that of w/o SIM**

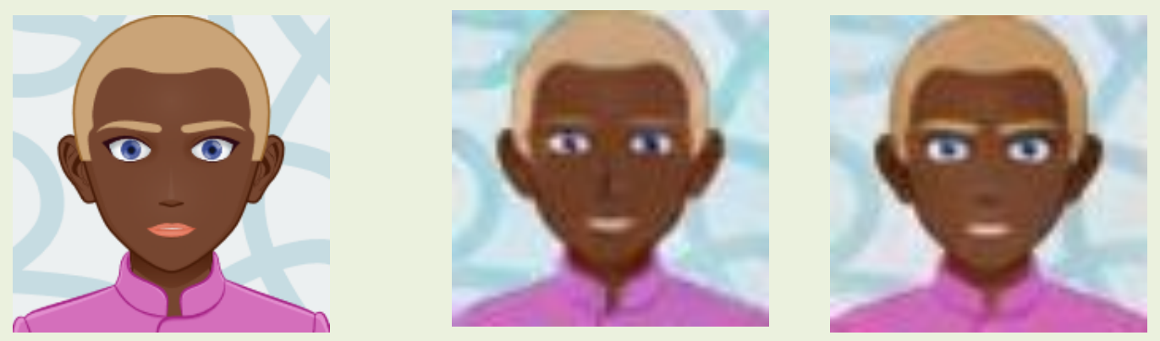
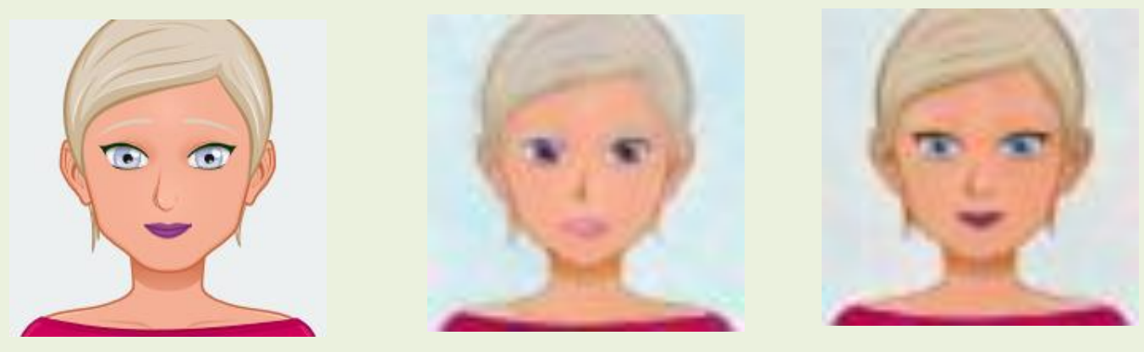
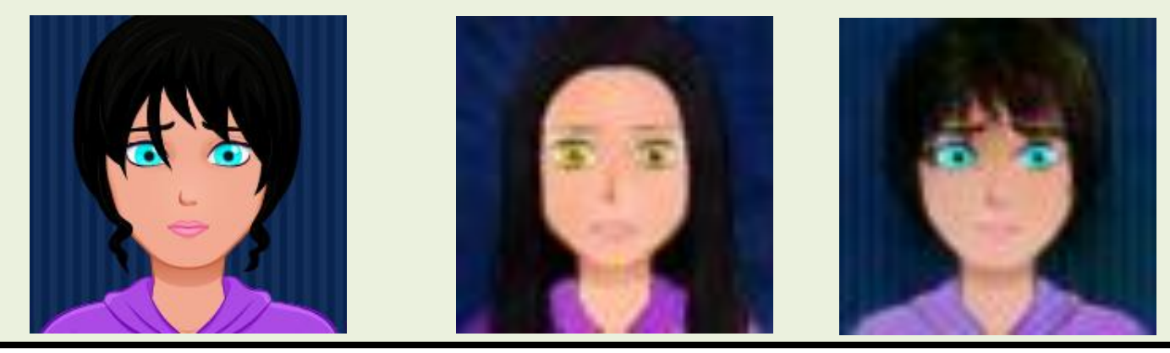
Subjective Evaluation



Crowdworker evaluated the preference of generated images in 5-grade between w/o and w/ SIM

- with test set, 3 evaluation on each sample, considering order effect: 230x3x2 = 1380 in total
- **Over 60%, w/ SIM was preferred**

Case study

source A B	given instruction	preference
	make his mouth thicker	B is much better
	make the upper earlobes flare outwards	A and B are equal
	make her hair longer and smooth	A is much better