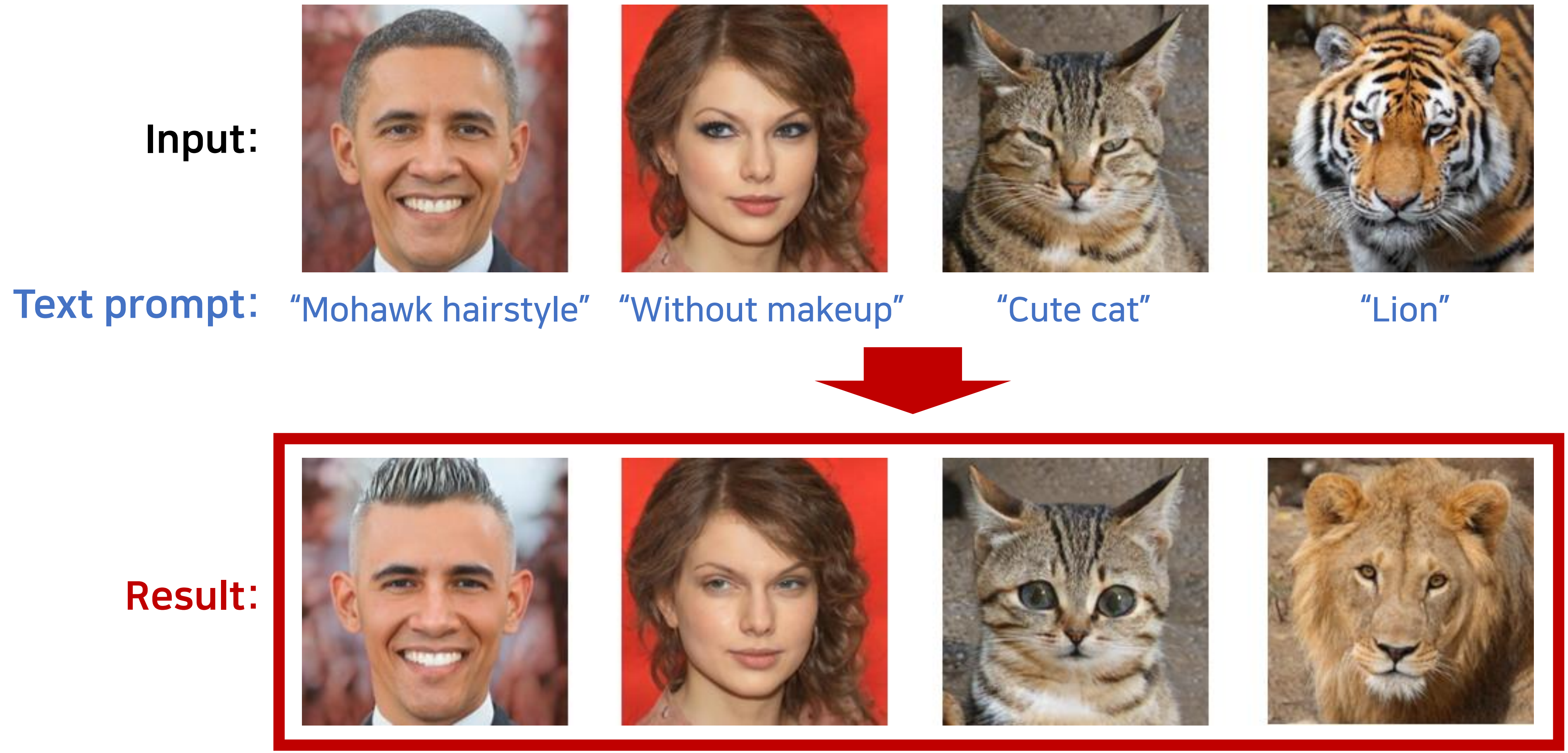
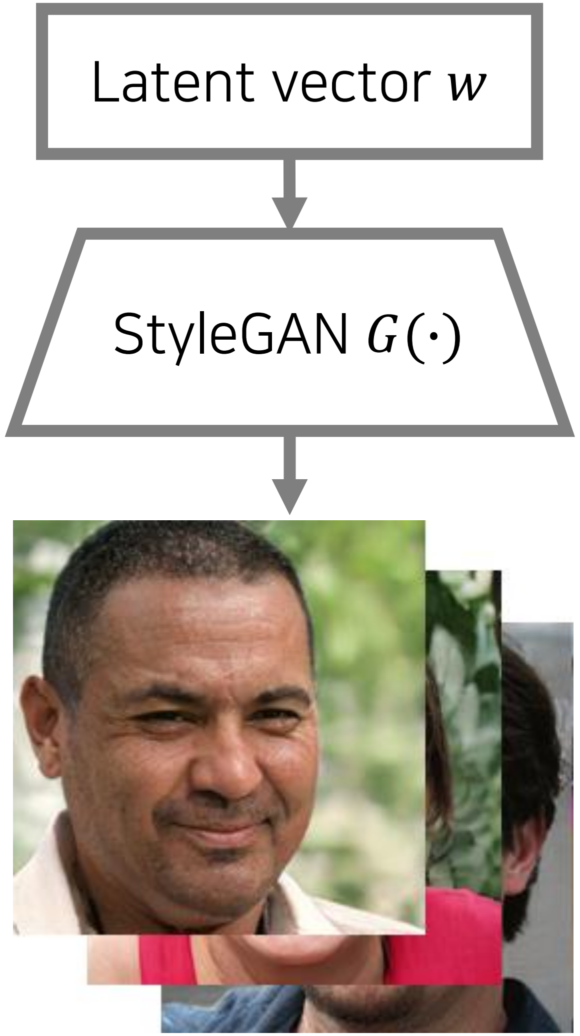
StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery



Background: StyleGAN (CVPR 2019)

- Propose an efficient architecture to generate high-quality images.
- Present a 1024 X 1024 high-quality face dataset (**FFHQ**).
- Improve the **disentanglement** of semantic features.

Source A　　Source B　　Coarse styles from source B



Background: Latent Vector Meanings of StyleGAN

Background: Face Manipulation Using StyleGAN

① Encoding step

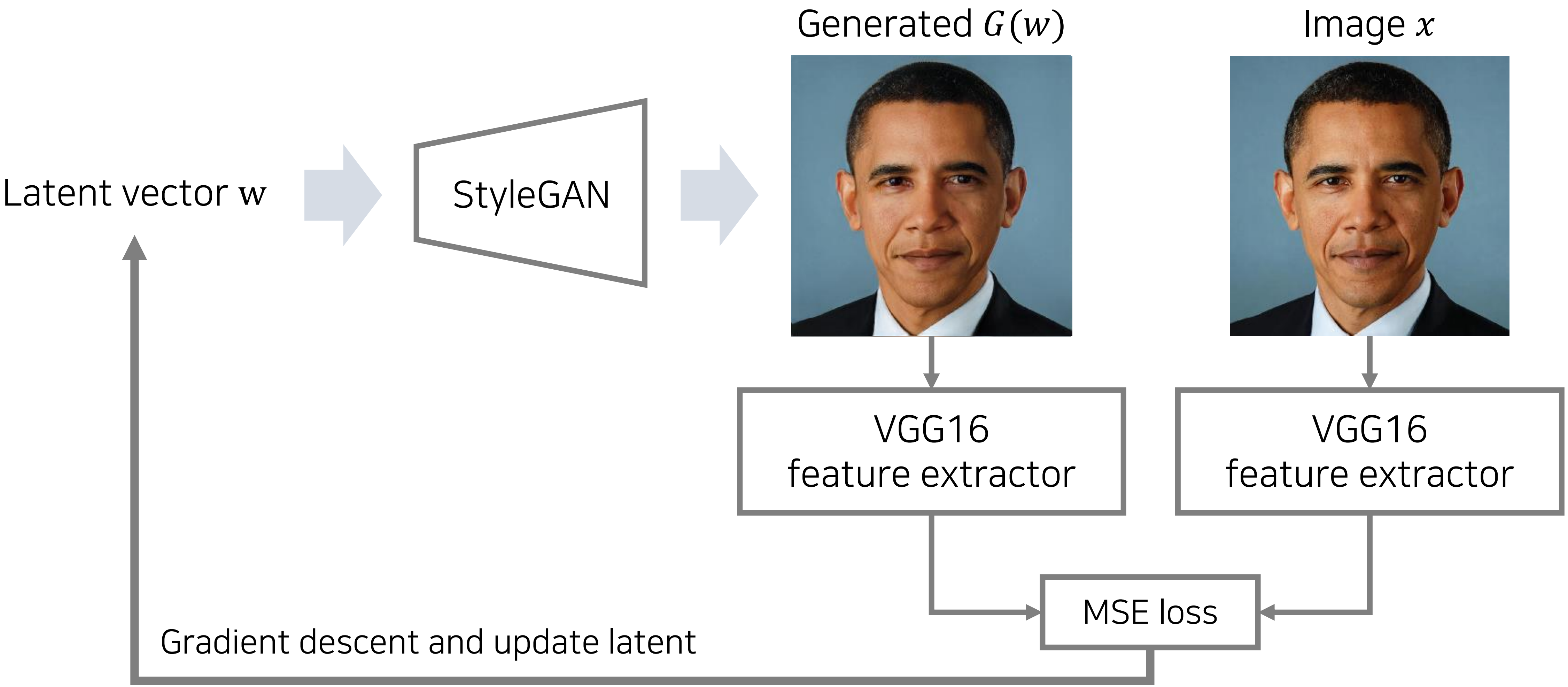
② Manipulation step

Input image x

$G(w_{encoded})$

$G(w_{manipulated})$

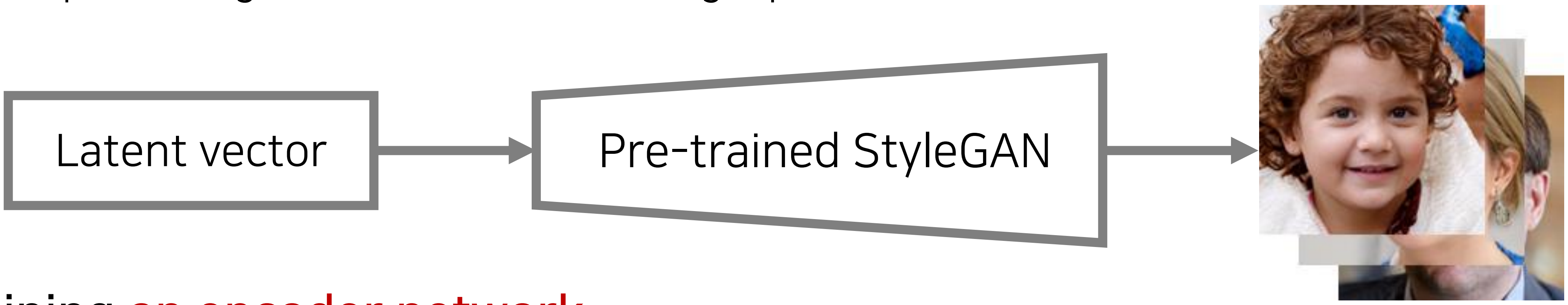
Background: StyleGAN Encoding (Inversion) ① Gradient Descent



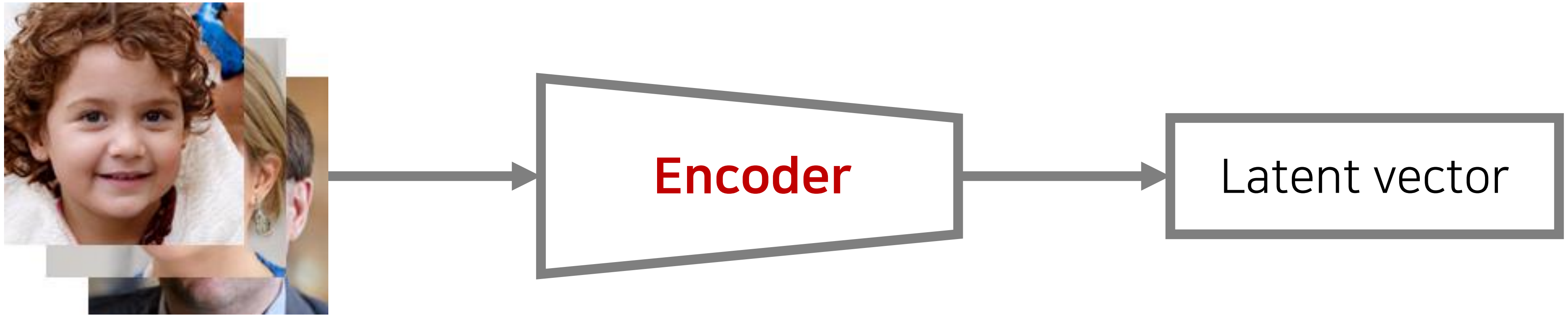
Background: StyleGAN Encoding (Inversion) ② Encoder Network

1) Latent vector dataset generation

- Prepare a large number of (latent, image) pairs.

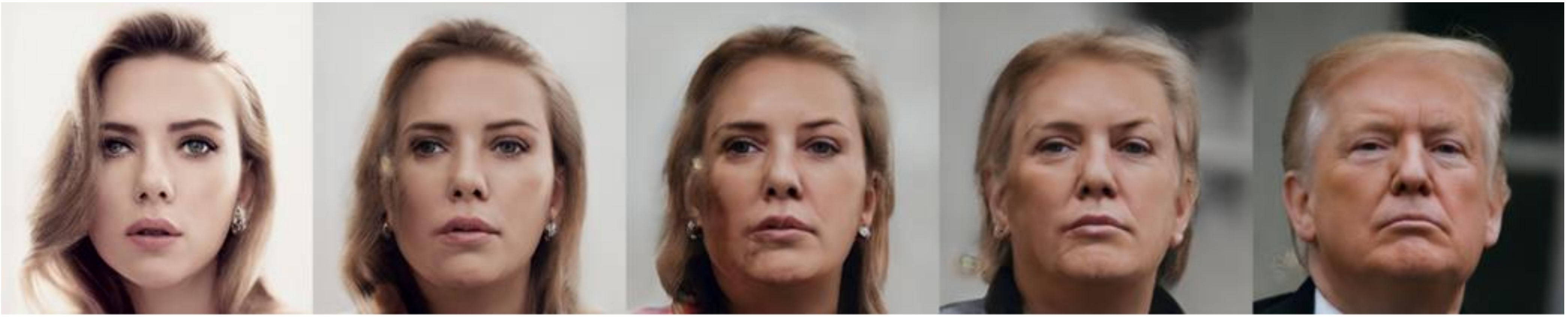


2) Training an encoder network



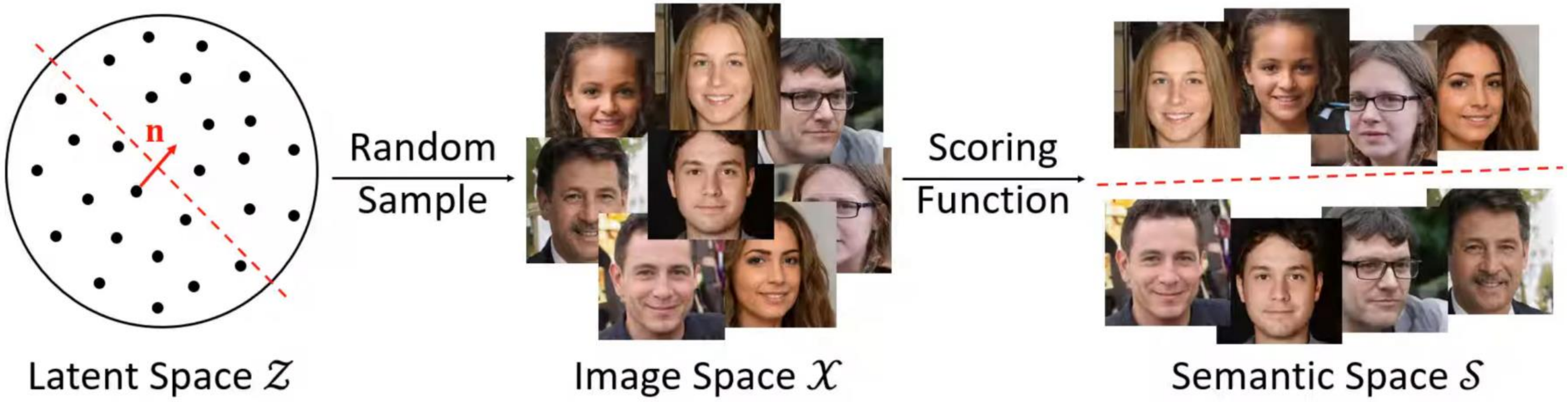
Background: Face Manipulation Examples ① Latent Interpolation

- We can interpolate between encoded latent a and encoded latent b.

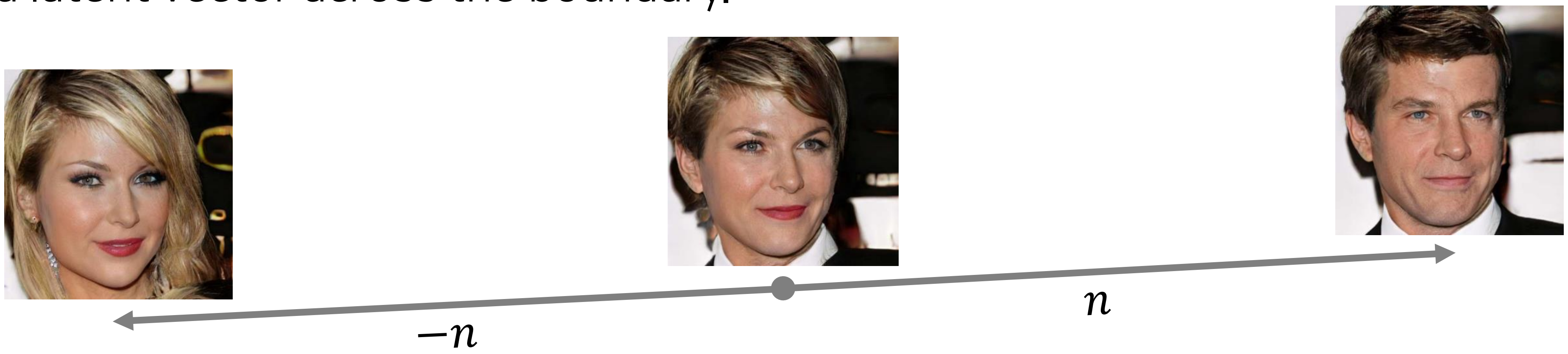


Background: Face Manipulation Examples ② Learned Direction

1) Learns a boundary of an attribute (such as gender, age).

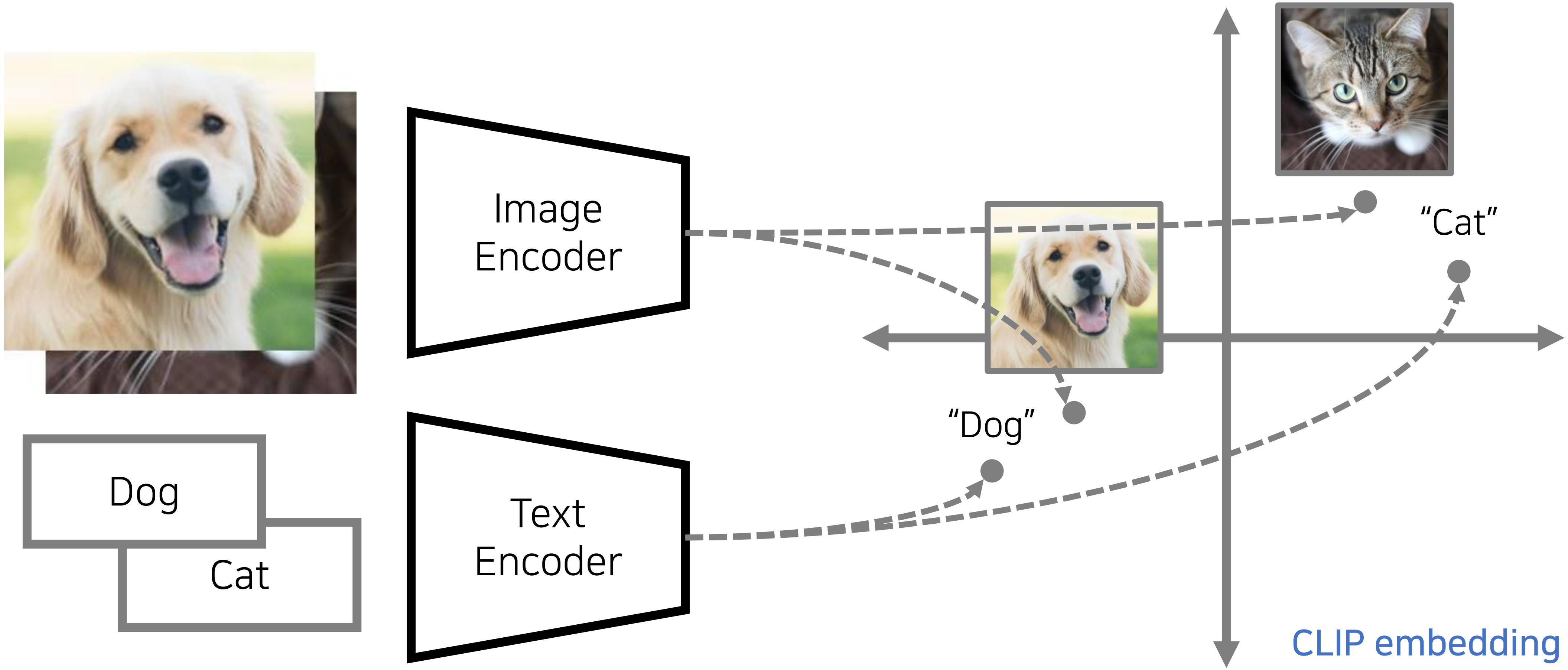


2) Update a latent vector across the boundary.



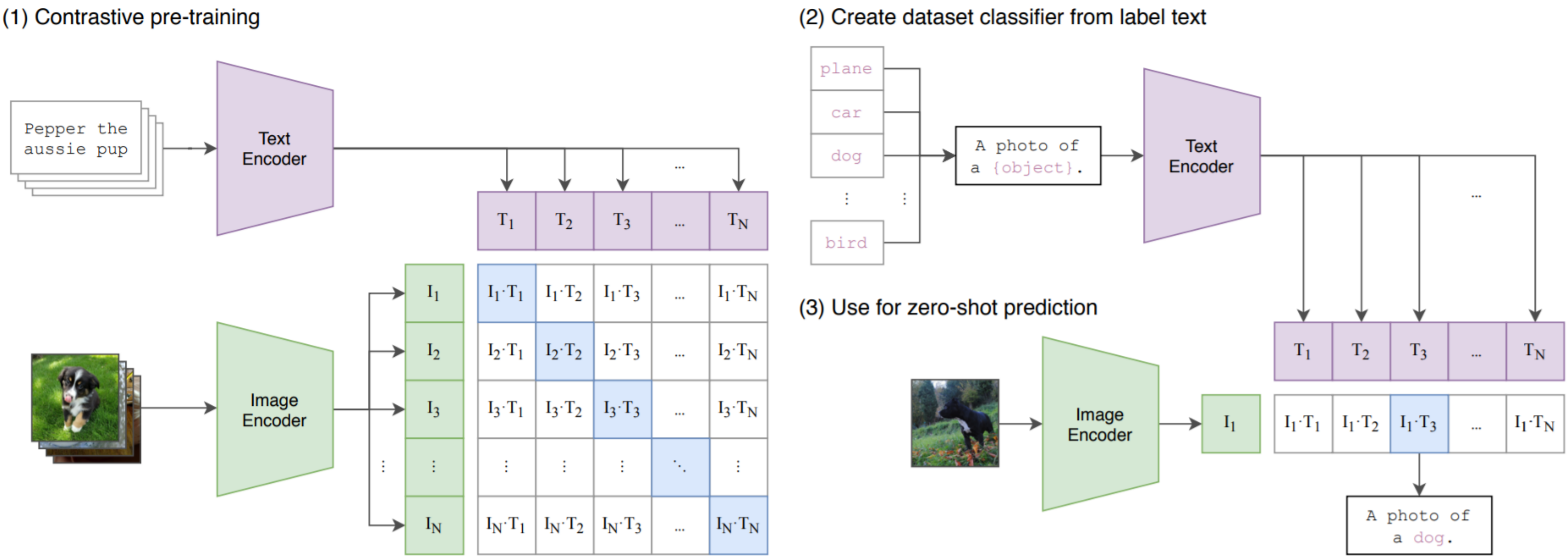
Background: CLIP (Contrastive Language-Image Pre-training)

- CLIP jointly trains an image encoder and a text encoder using a large dataset.

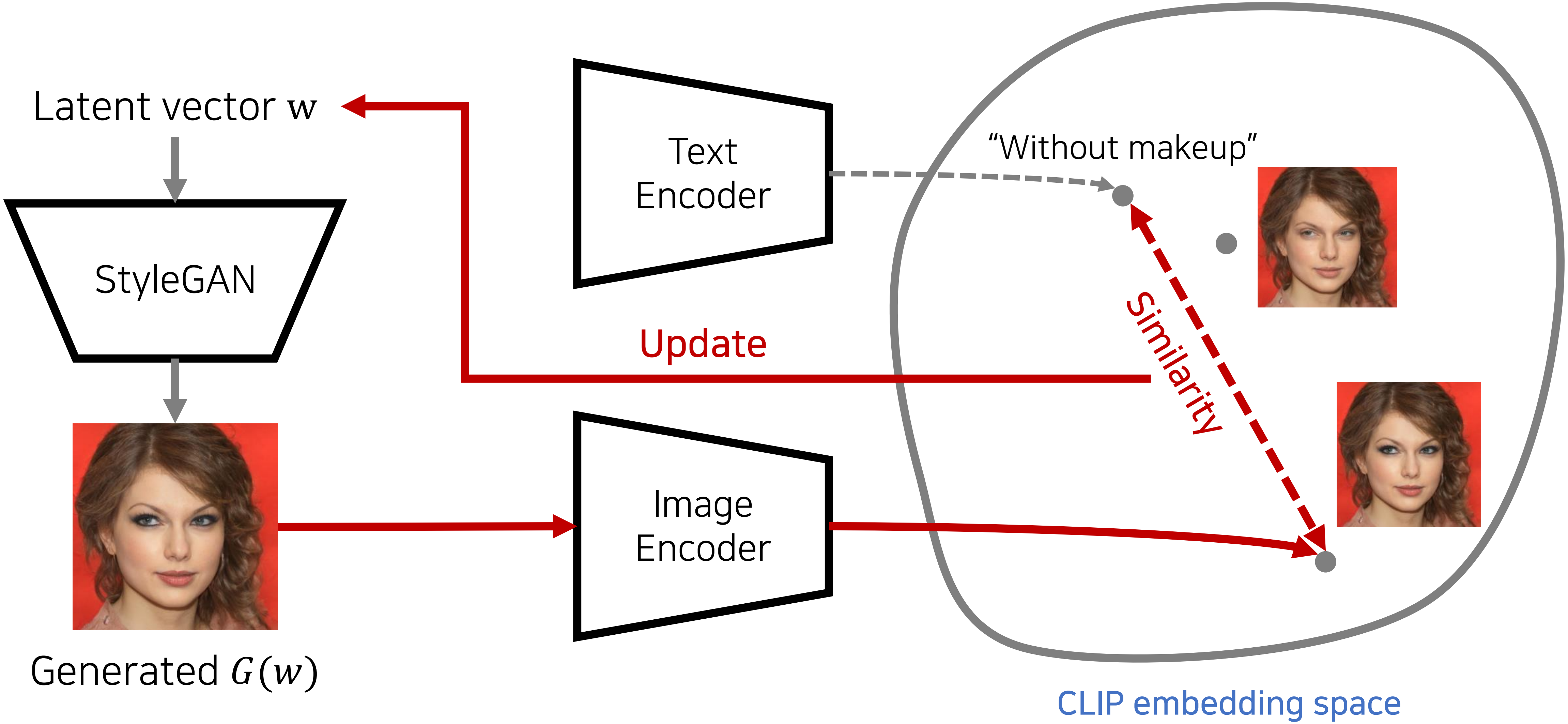


Background: CLIP (Contrastive Language-Image Pre-training)

- CLIP jointly trains an image encoder and a text encoder using a large dataset.



StyleCLIP (StyleGAN + CLIP): Main Idea



StyleCLIP Introduction

- StyleCLIP is a text-based interface for StyleGAN image manipulation.
- Propose three methods that do not require such manual effort.
    1. Introduce an **optimization method** that utilizes a CLIP-based loss.
    2. Introduce a **latent mapper** that infers a text-guided latent manipulation.
    3. Present a method for mapping text prompts to input-agnostic **global directions**.

	pre-proc.	train time	infer. time	input image dependent	latent space
optimizer	–	–	98 sec	yes	$\mathcal{W}+$
mapper	–	10 – 12h	75 ms	yes	$\mathcal{W}+$
global dir.	4h	–	72 ms	no	$\mathcal{S}$

StyleCLIP Method ① Latent Optimization

- **Latent optimization**: a simple approach for leveraging CLIP to guide image manipulation.

$$\underset{w \in \mathcal{W}+}{\arg\min}\ \underbrace{D_{\text{CLIP}}(G(w), t)}_{\text{For manipulation}} + \underbrace{\lambda_{\text{L2}} \left\| w - w_s \right\|_2 + \lambda_{\text{ID}} \mathcal{L}_{\text{ID}}(w)}_{\text{For similarity to the input image}}$$

$$\mathcal{L}_{\text{ID}}(w) = 1 - \langle R(G(w_s)), R(G(w)) \rangle$$

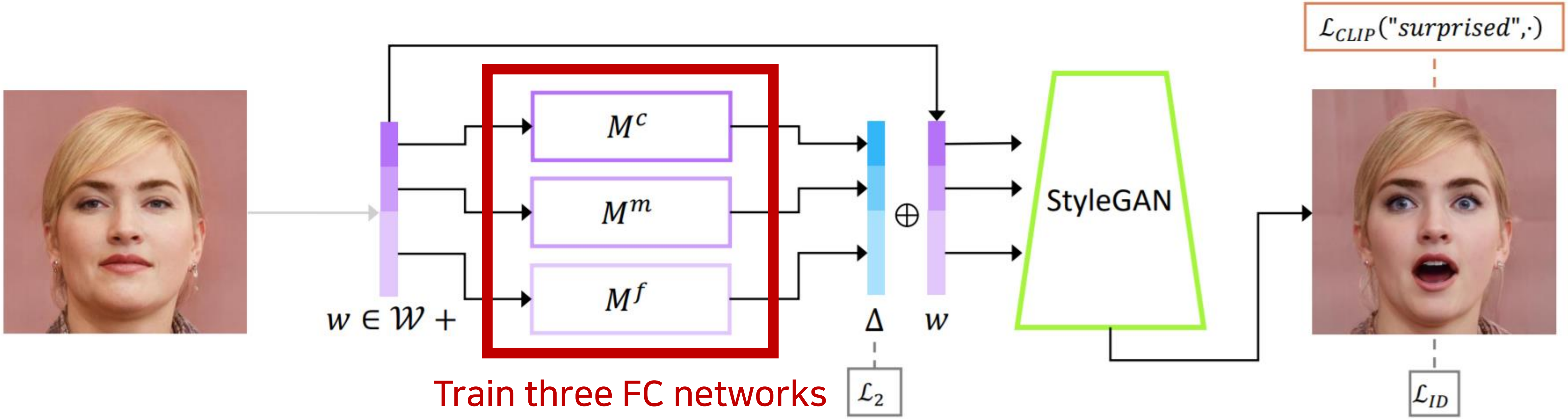
R: Pretrained $ArcFace$ network

D_{clip}: Cosine distance between the CLIP embeddings

The optimization method requires **200 – 300 iterations** that spend several minutes.

StyleCLIP Method ② Latent Mapper

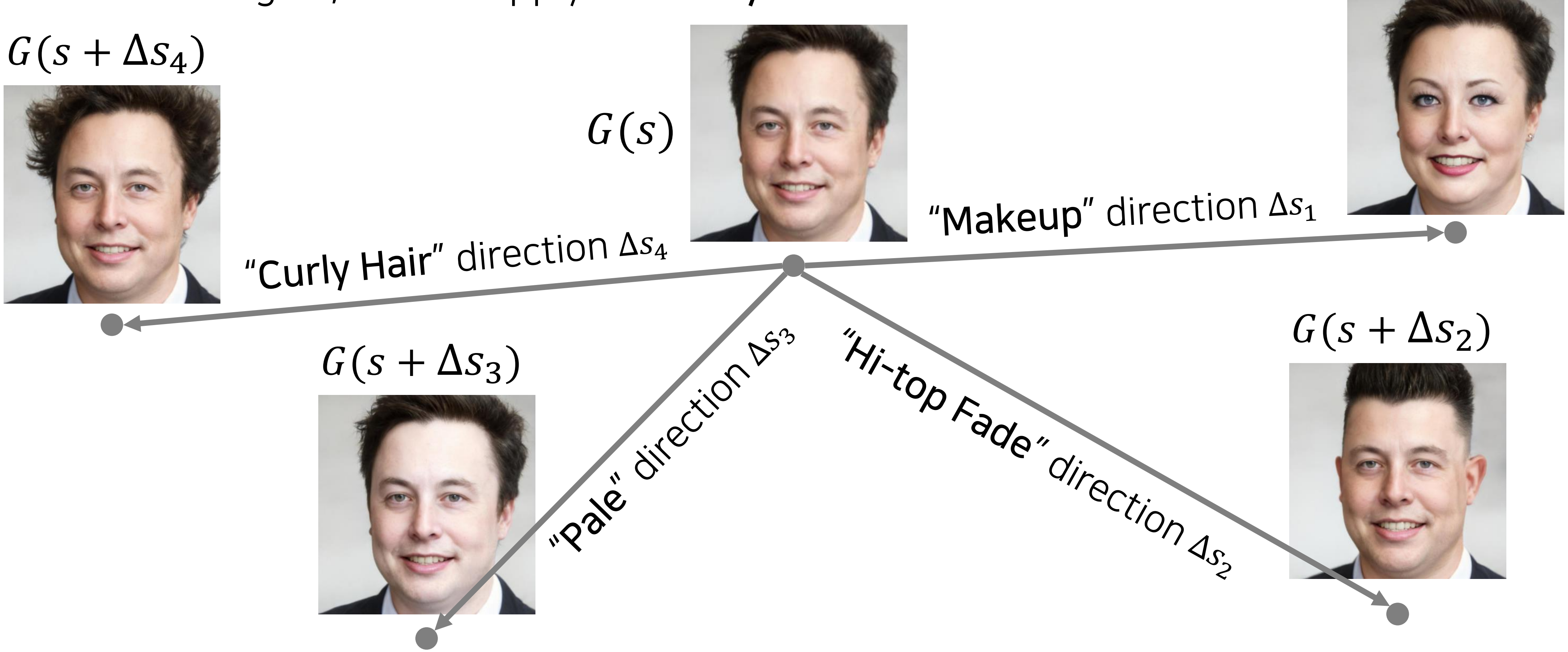
- **Latent mapper** is trained **to manipulate the desired attributes of the image** as indicated by the text prompt t, while preserving the other visual attributes of the input image.



After trained per text prompt (10 hours), the mapper manipulates attributes in one forward.

StyleCLIP Method ③ Global Directions

- Find a **global direction** $\Delta s \in S$ in a StyleGAN's style space S.
  - After finding Δs, we can apply Δs to **any latent** s.



Comparisons

- StyleCLIP is **not limited** to preset manipulation directions.
  - However, StyleCLIP shows **competitive results** on even common attributes.