ACM Multimedia 2022
Lisbon, Portugal | 10-14 October

STRUCT @PKU
Spatial and Temporal Restoration, Understanding and Compression

Microsoft

AI Illustrator: Translating Raw Descriptions into Images by Prompt-based Cross-Modal Generation

Yiyang Ma[1], Huan Yang[2], Bei Liu[2], Jianlong Fu[2], Jiaying Liu[1]

1 *Wangxuan Institute of Computer Technology, Peking University*

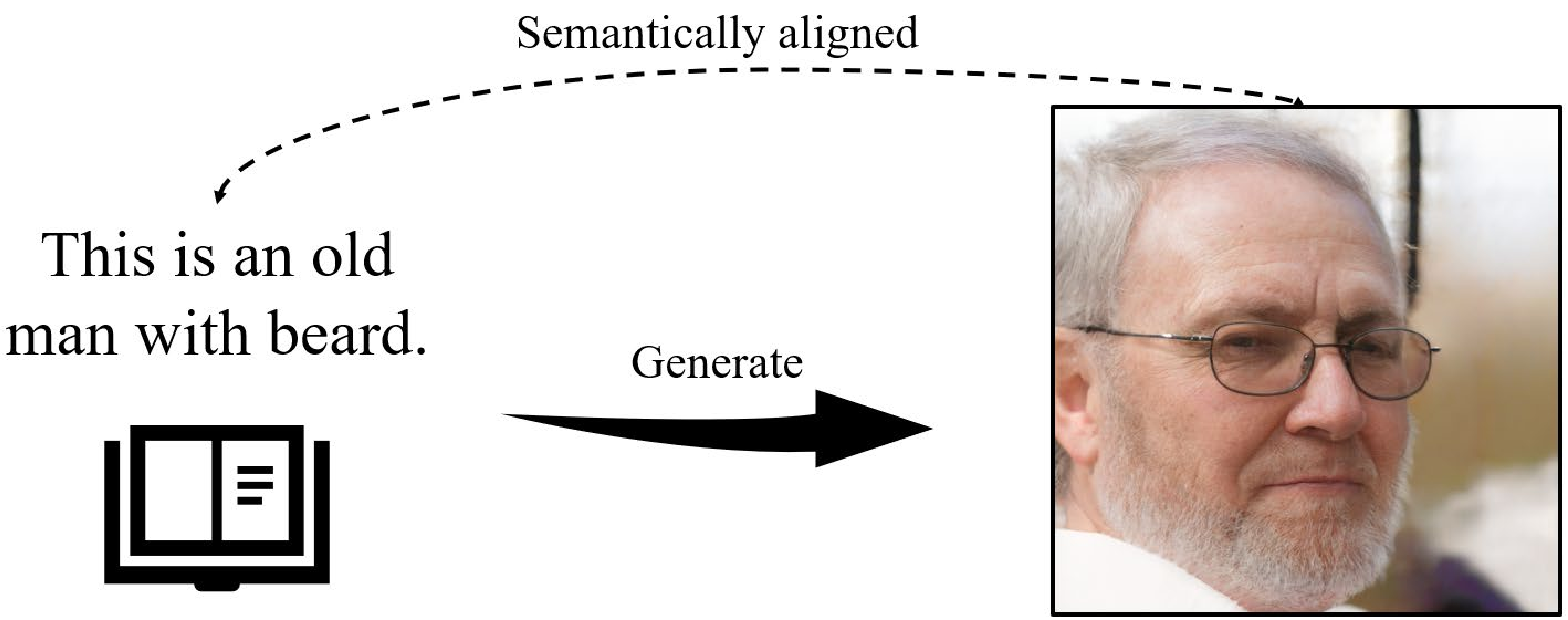
2 *Microsoft Research*

Aim and Challenge

- **Problem:** Translating raw descriptions to corresponding images

Descriptions can be complex and challenging

- descriptions may be abstract.
- descriptions may have multiple meanings which are hard to be semantically aligned.
- translated images should be impressive.



Aim and Challenge

- **Existing works:**

  There's a trilemma among

  - semantically alignment

  - open-world words

  - image quality

Our work aims at dealing with this trilemma.

Aim and Challenge

- **How to deal with these challenges?**

  Pretrained large scale models!

  - challenge of semantics:

    Contrastive Language-Image Pretraining (CLIP)

  - challenge of image quality:

    StyleGAN

Method

- **Main Idea:** transmit semantics through the pretrained models:

  Input Texts

  (1) $\rightarrow$ CLIP Text Embeddings (*CTE*s)

  (2) $\rightarrow$ CLIP Image Embeddings (*CIE*s)

  (3) $\rightarrow$ StyleGAN Z Space Embeddings (*SE*s)

  (4) $\rightarrow$ Translated Images

Method

- **Main Idea:** transmit semantics through the pretrained models:

  Input Texts

  (1) → CLIP Text Embeddings (*CTE*s)

  (2) → CLIP Image Embeddings (*CIE*s)

  (3) → StyleGAN Z Space Embeddings (*SE*s)

  (4) → Translated Images

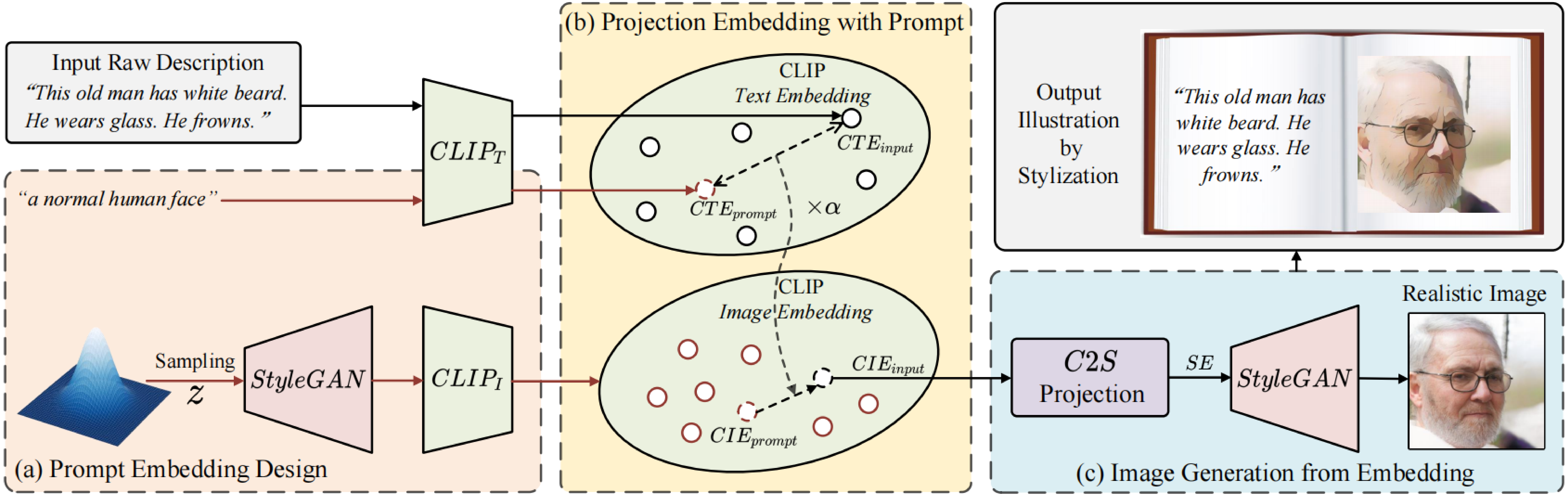
  Projection (1) and (4) can be done with existing models.

  (1): CLIP Text Encoder

  (4): Pretrained StyleGAN

Method

- **Pipeline:** two projections within the latent spaces of the pretrained models.

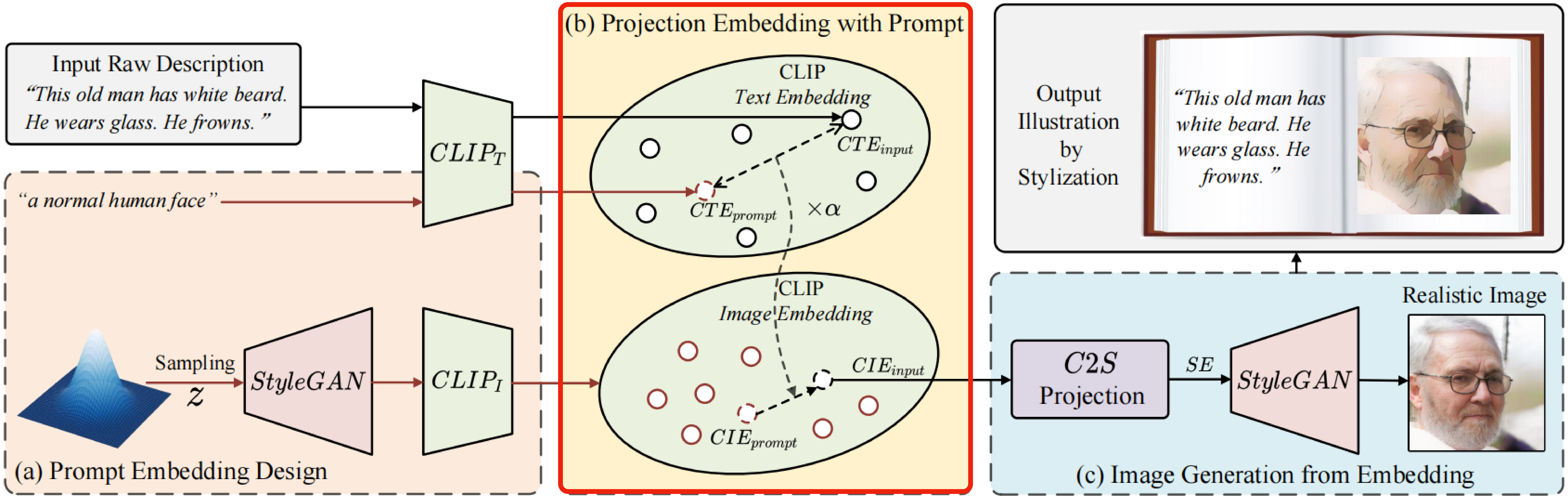


- text embeddings to image embeddings

- CLIP to StyleGAN

Method

- **Pipeline:** two projections within the latent spaces of the pretrained models.

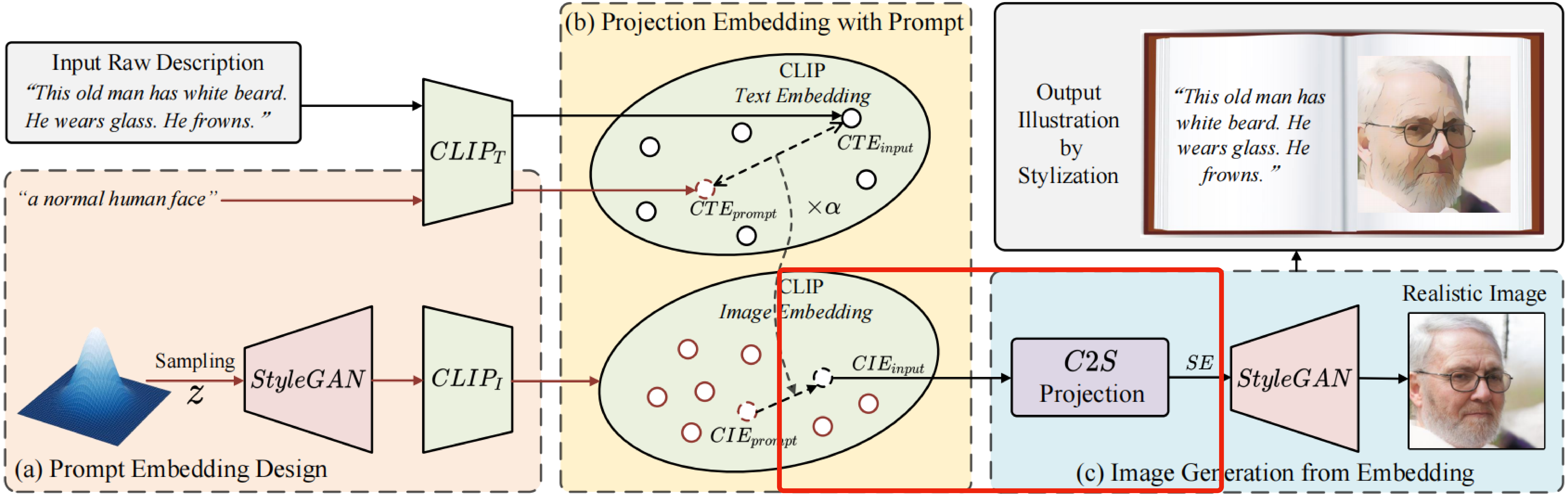


- text embeddings to image embeddings

- CLIP to StyleGAN

Method

- **Pipeline:** two projections within the latent spaces of the pretrained models.



- text embeddings to image embeddings

- CLIP to StyleGAN

Method

- **The First Projection: Text Embeddings to Image Embeddings**

CLIP has two latent spaces:

- Text latent space
- Image latent space

Semantically aligned text-image pairs will have embedding pairs which have small cosine distances.

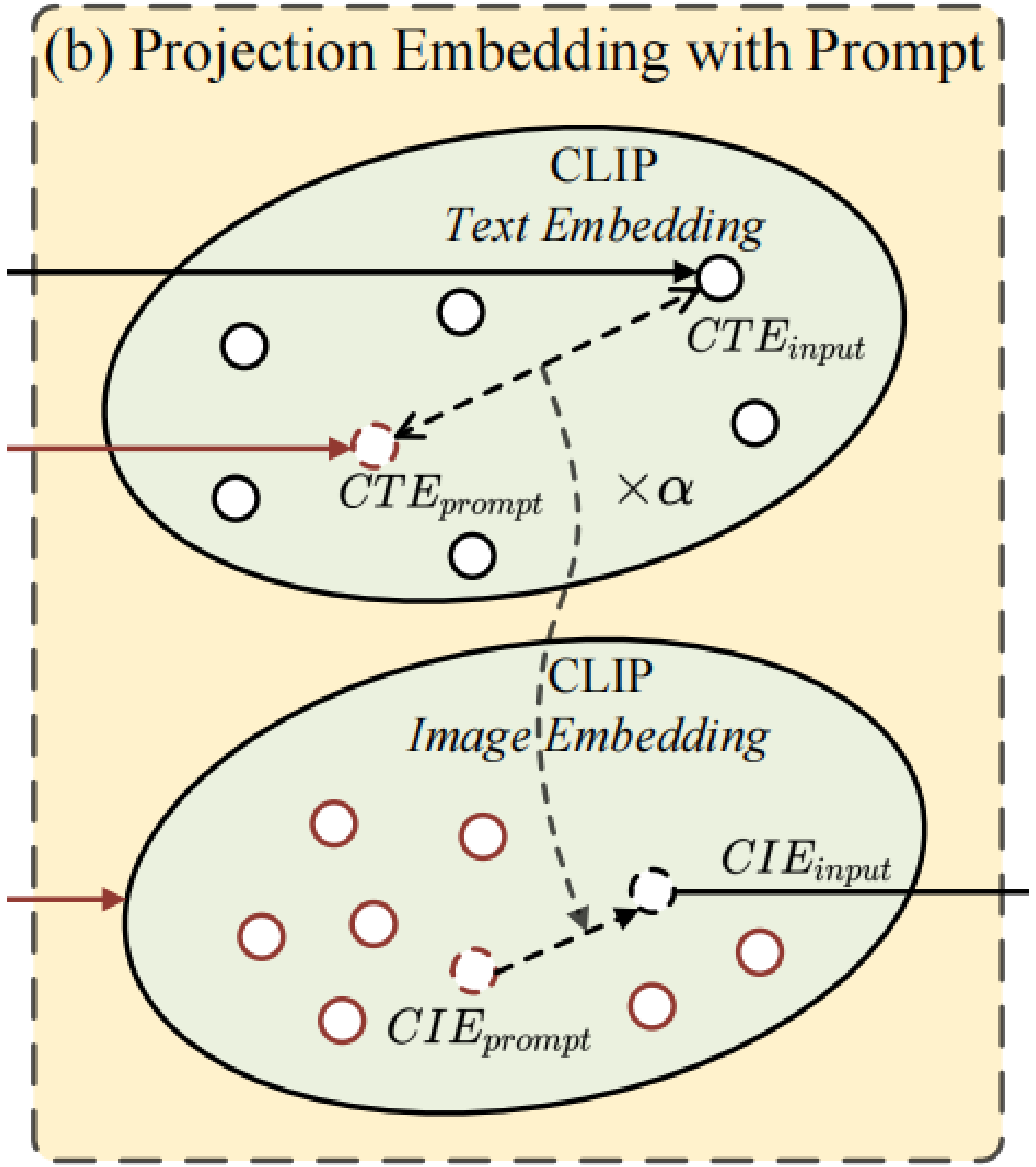
Method

- **The First Projection: Text Embeddings to Image Embeddings**

Due to the character of CLIP, for two pairs of matched texts and images, we have:

$$CTE_1 - CTE_2 = CIE_1 - CIE_2 \qquad (1)$$

If we can find a semantically aligned pair of representative embeddings, we can project input *CTE*s to corresponding *CIE*S.



Method

▸ **The First Projection: Text Embeddings to Image Embeddings**

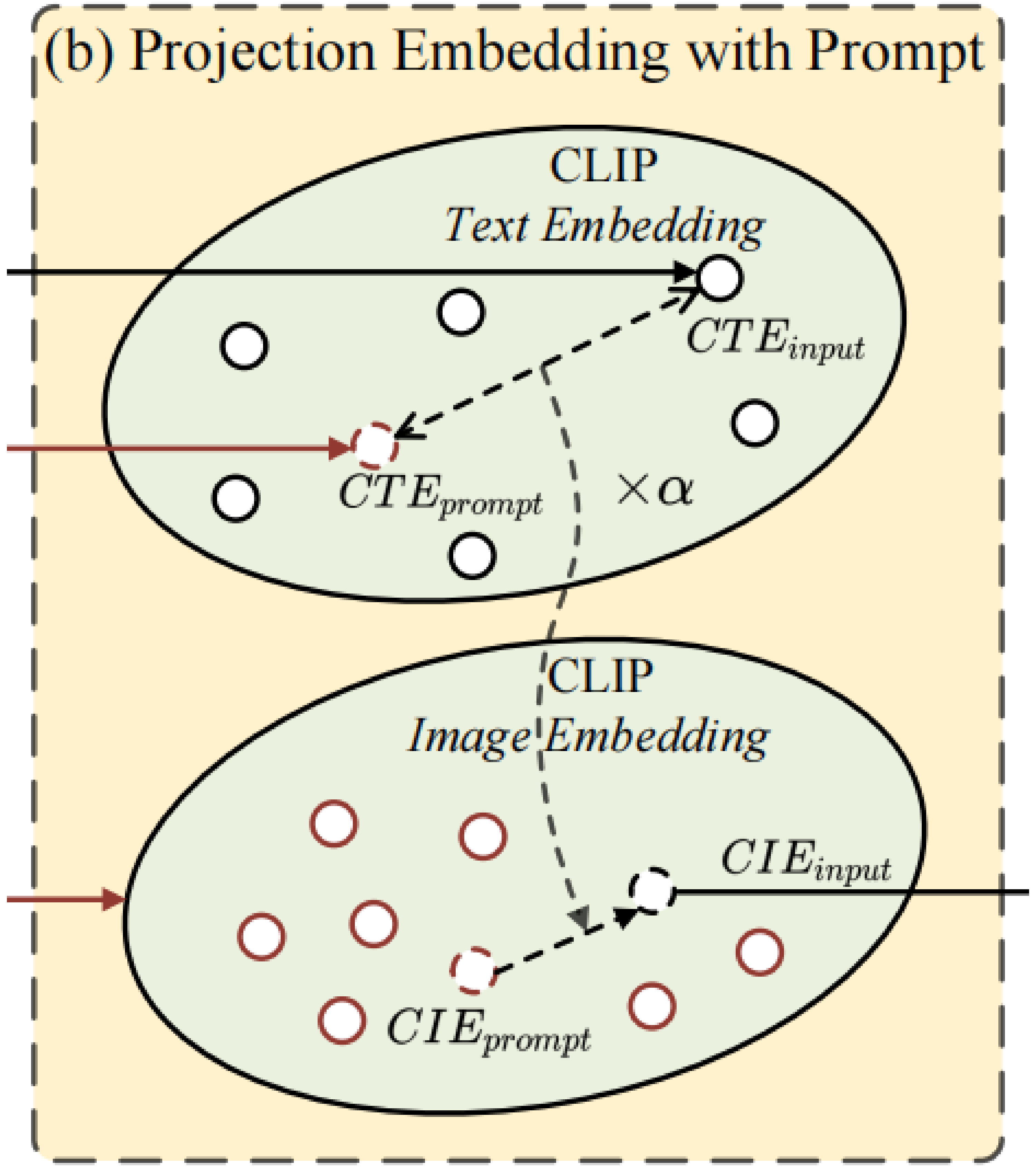
The "representative" pair is a prompt pair to latent projection. We have:

$$CIE_{input} = CIE_{prompt} + (CTE_{input} - CTE_{prompt}) \quad (2)$$

In practice, we use:

$$CIE_{input} = CIE_{prompt} + \alpha \cdot (CTE_{input} - CTE_{prompt}) \quad (3)$$

To control the distinctiveness of the projection.



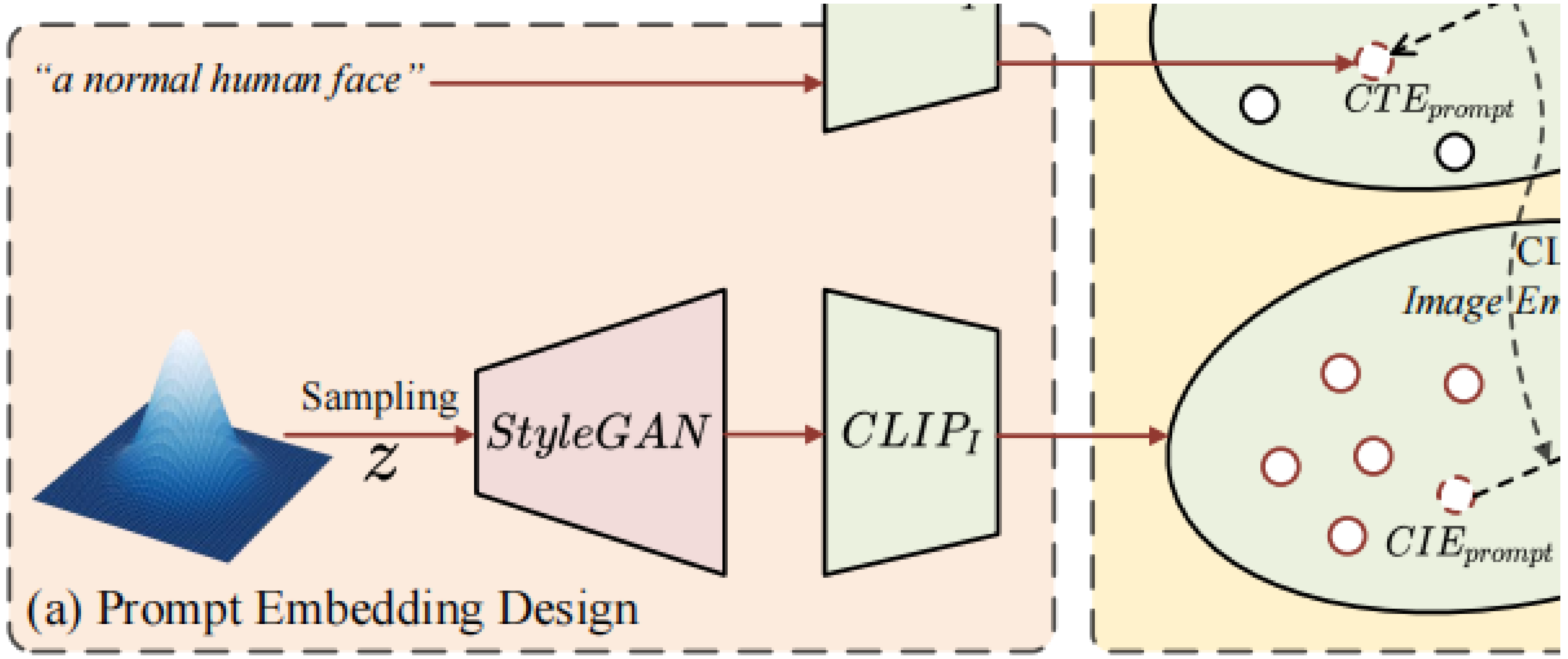
▸ The First Projection: Text Embeddings to Image Embeddings

How to find the prompt embeddings?

Because they are “representative”, they should have the largest average cosine similarity to all the embeddings.

$$\max_{\boldsymbol{y}} z = \frac{1}{n}\sum_{i=1}^{n}\frac{\boldsymbol{y}\cdot\boldsymbol{x}_i}{|\boldsymbol{y}|\cdot|\boldsymbol{x}_i|} \tag{4}$$

$$s.t. |\boldsymbol{y}| = 1 \tag{5}$$



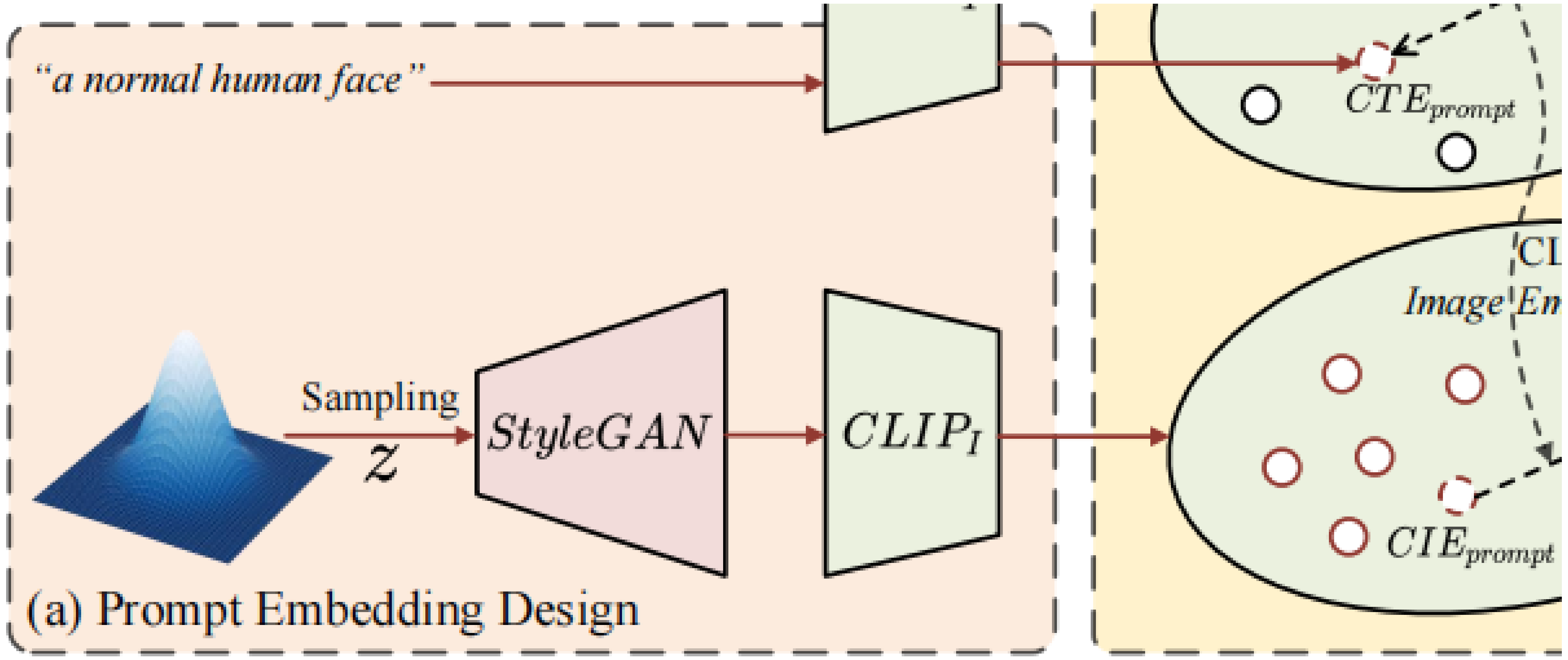
(a) Prompt Embedding Design

- **The First Projection: Text Embeddings to Image Embeddings**

We can simplify Eqn. 4 as:

$$\max_{\boldsymbol{y}} z = \boldsymbol{y} \cdot \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i \qquad (6)$$

which is the equation of a hyperplane.



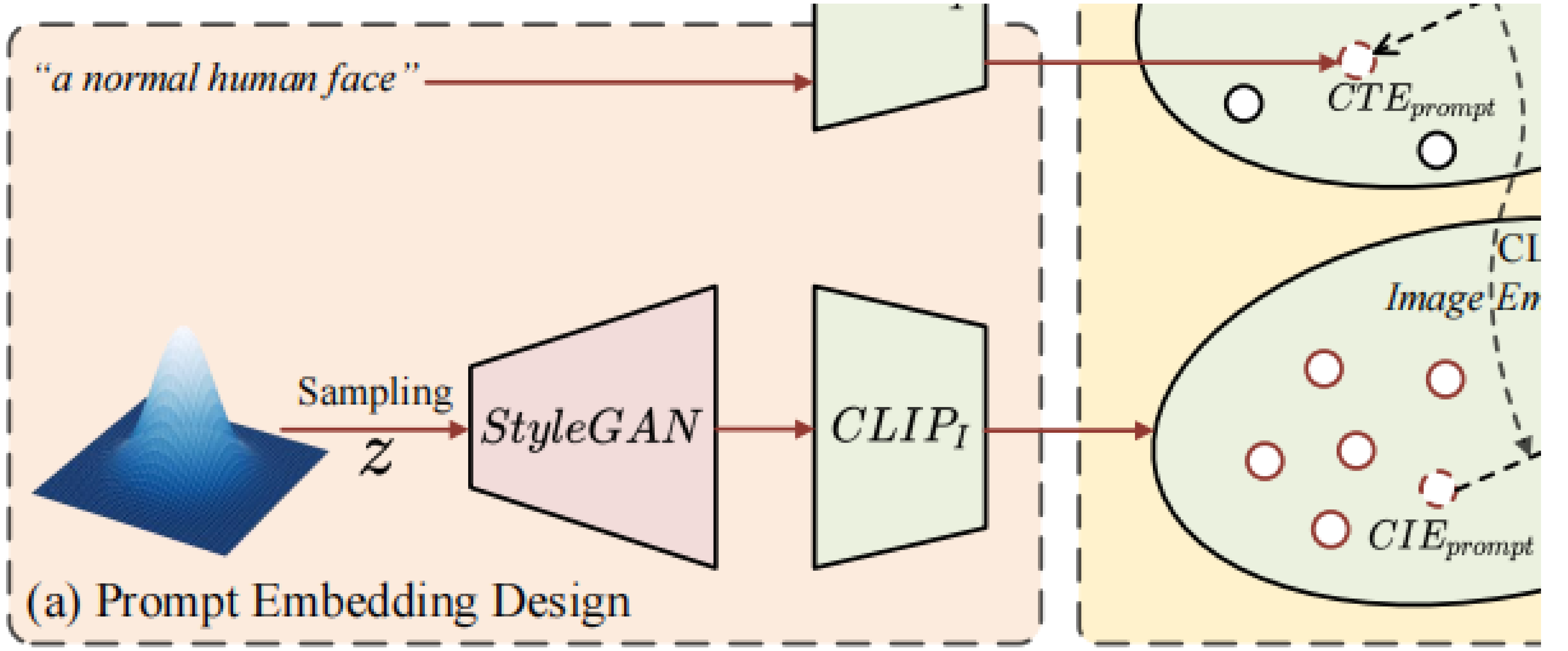
z will be biggest at the time of the hyperplane (Eqn. 6) and the hypersphere (Eqn. 5) are tangent. At this time,

$$\boldsymbol{y}' = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i, \ \boldsymbol{y} = \frac{\boldsymbol{y}'}{|\boldsymbol{y}'|} \qquad (7)$$

- **The First Projection: Text Embeddings to Image Embeddings**

For images, we can sample a large number of images by StyleGAN and calculate image prompt embedding through Eqn. 7.

For texts, we can simply specify a sentence which contains the meaning of “general” or “normal” like “A normal human face.”.

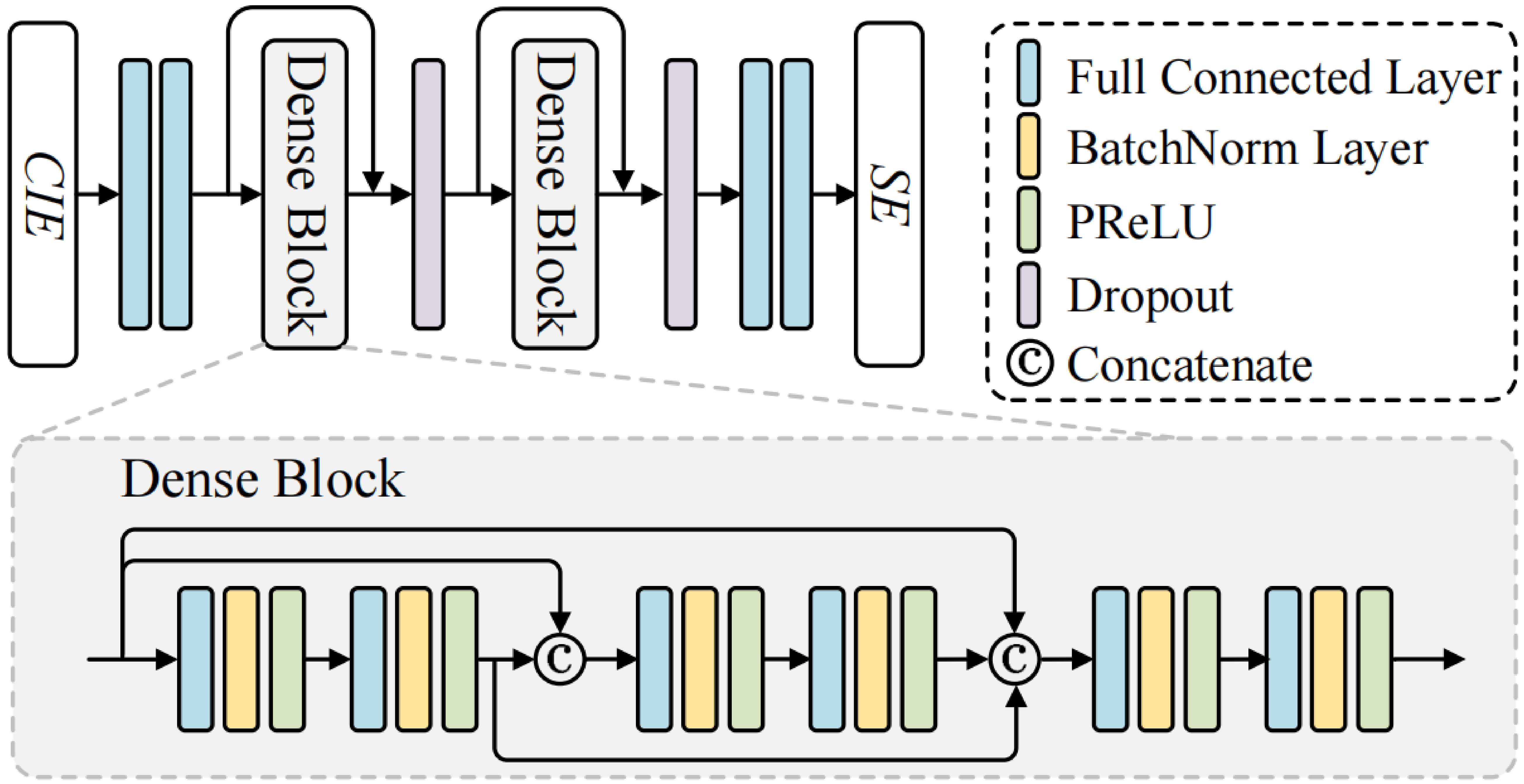


Method

- **The Second Projection: CLIP Embeddings to StyleGAN Embeddings**

  We build a NN to learn the projection. The training pairs are easy to get.

  The network architecture is shown below.



Method

- **The Second Projection: CLIP Embeddings to StyleGAN Embeddings**

The training loss consists of 3 parts.

Basic constraint of the network:

$$\mathcal{L}_{l1} = ||SE_{pred} - SE_{true}||_1 \tag{8}$$

Semantic consistency loss:

$$\mathcal{L}_{sem\_cons} = CosDis(CIE_{input}, CLIP_I(G(SE_{pred}))) \tag{9}$$

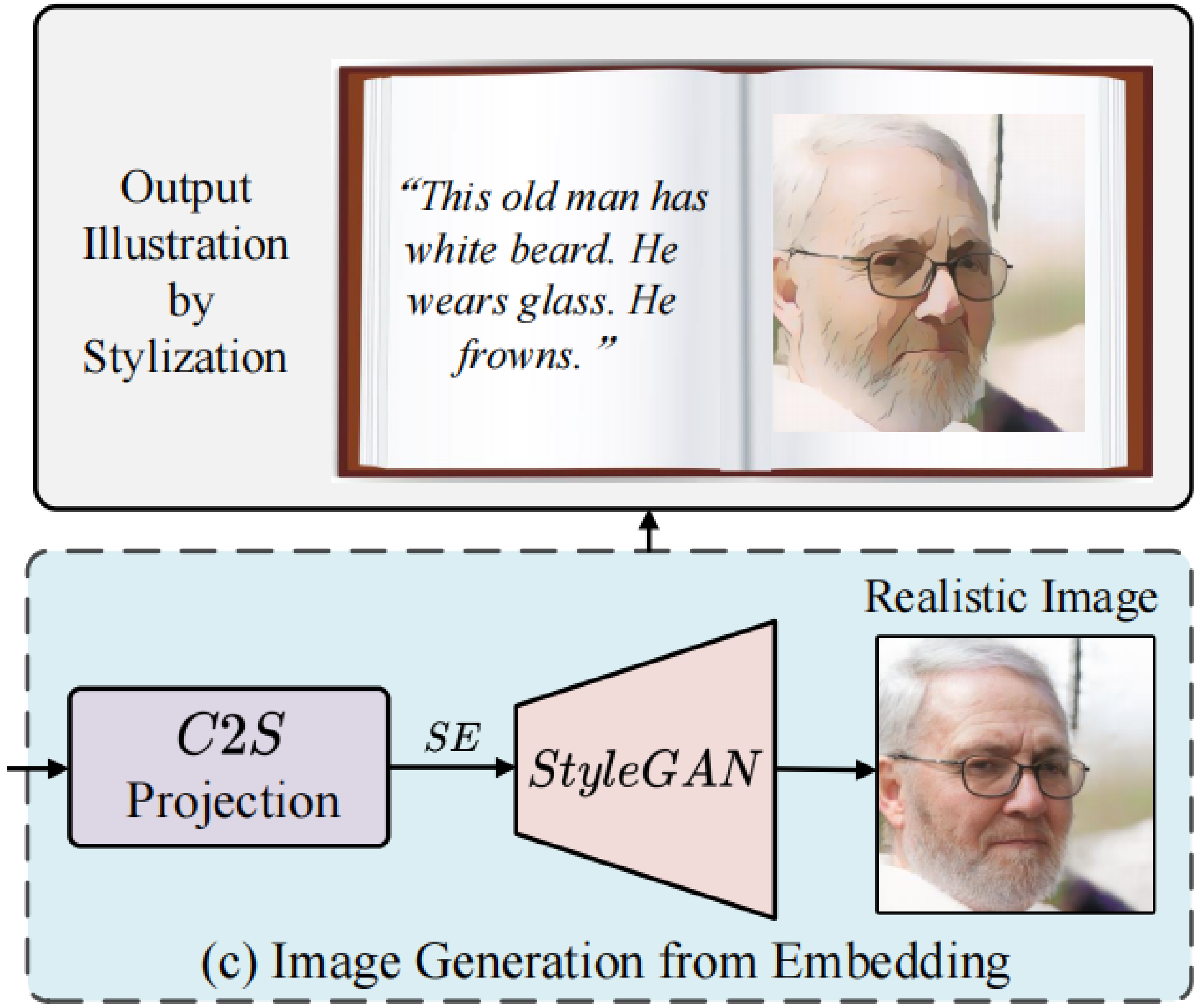
The regularization loss which ensures the predicted SE is in the StyleGAN Z space:

$$\mathcal{L}_{reg} = ||mean(SE_{pred})||_1 + ||std(SE_{pred}) - 1||_1 \tag{10}$$

The total loss is the combination of the three losses.

- **Cartoonlization at Last**

In order to use the translation results as illustrations, our pipeline can apply a stylization module to convert the realistic images to cartoon images.



Experiments

- **Experiments**

  - Texts containing only limited words.

  - Texts containing open-world words.

  - Diverse results on one same text input.

  - Non-face results and cartoon results.

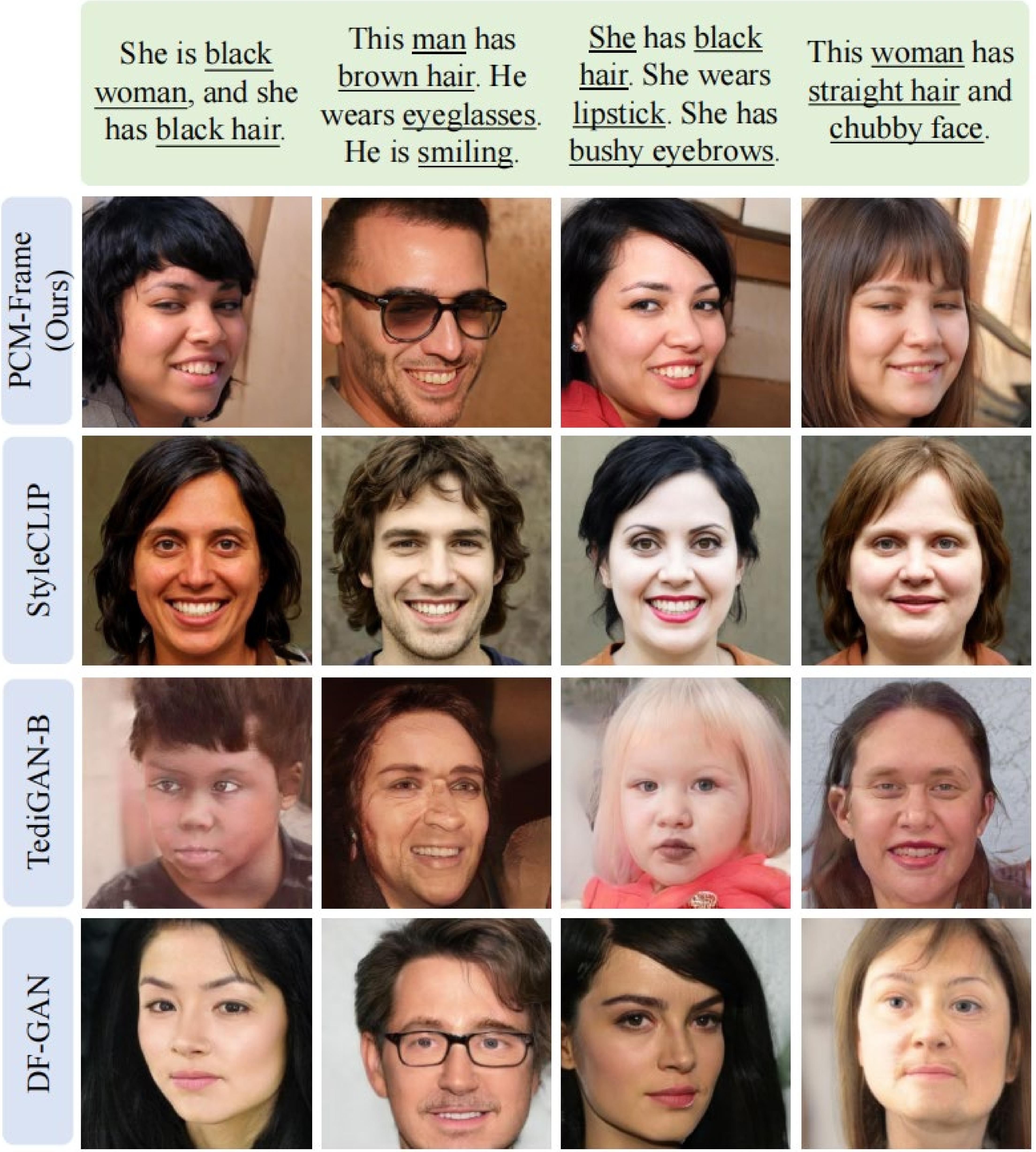
  - Manipulation results on generated images.

Experiments

- **Texts Containing Limited Words:**

  Our method is based on CLIP which can deal with open-world words.

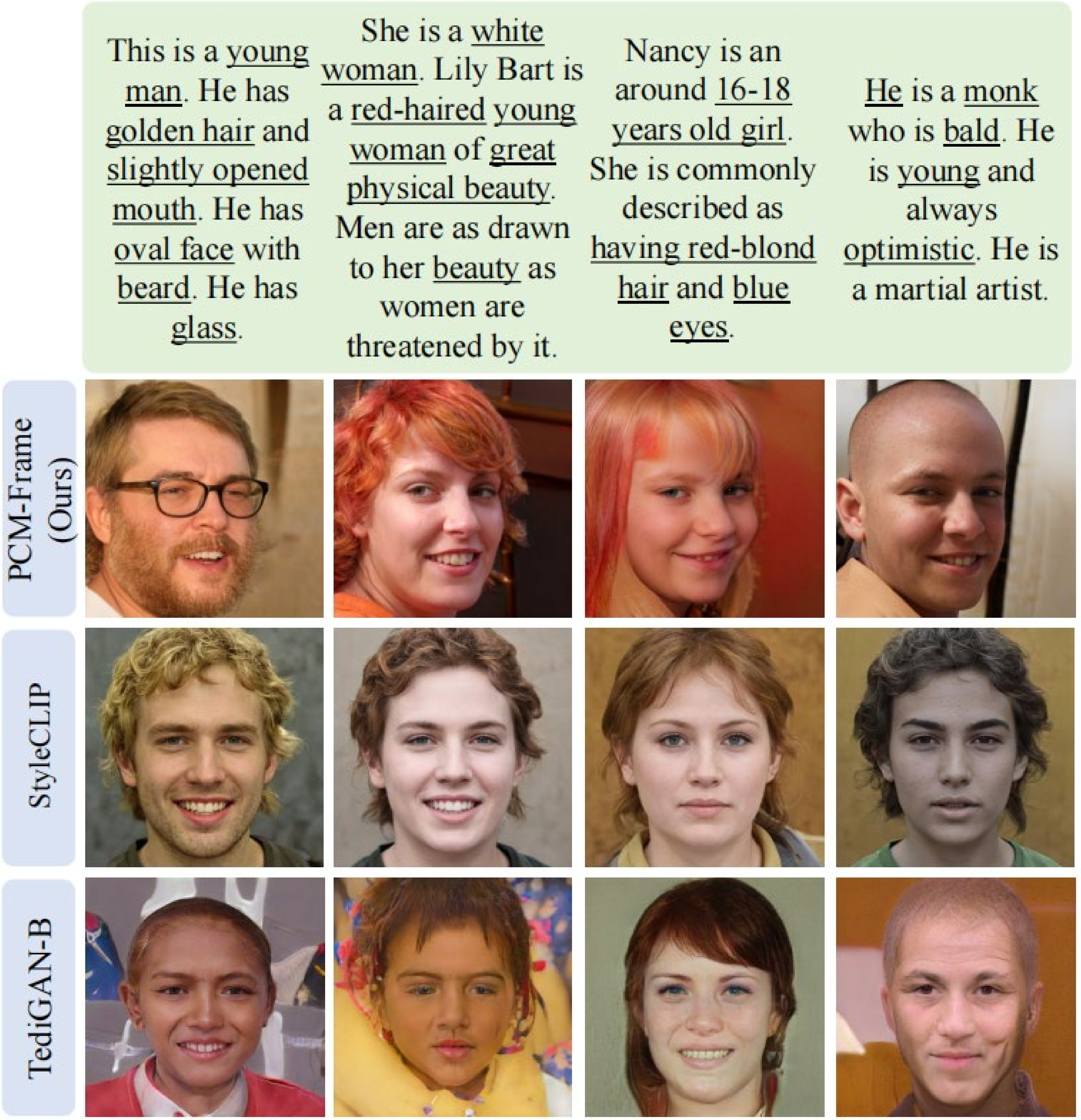
  But in order to compare with the methods which cannot process open-world words, we first show the translation results containing only the words of Multi-Modal CelebA dataset.



Experiments

- **Texts Containing Open-World Words:**

  Then, we show the translation results containing open-world words. This task is more challenging.



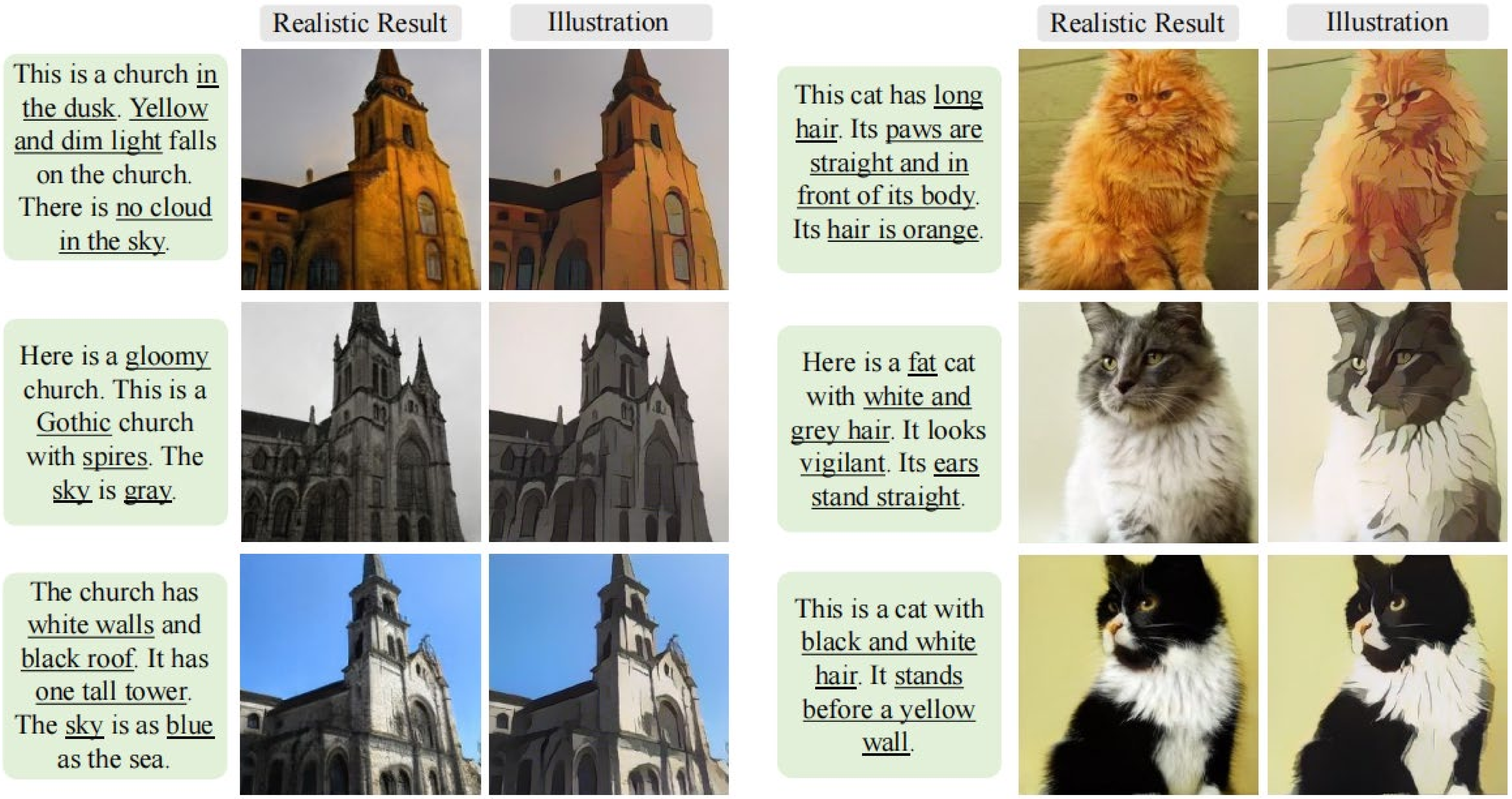
- **Diverse Results for One Single Text**

Our method can generate diverse results with one input by taking random SEs in certain layers of StyleGAN. The results are shown.

Experiments

- **Diverse Results for One Single Text**

Our method can also translating non-face images as long as we have the corresponding pretrained generative model.



- **Manipulation Results on Generated Images**

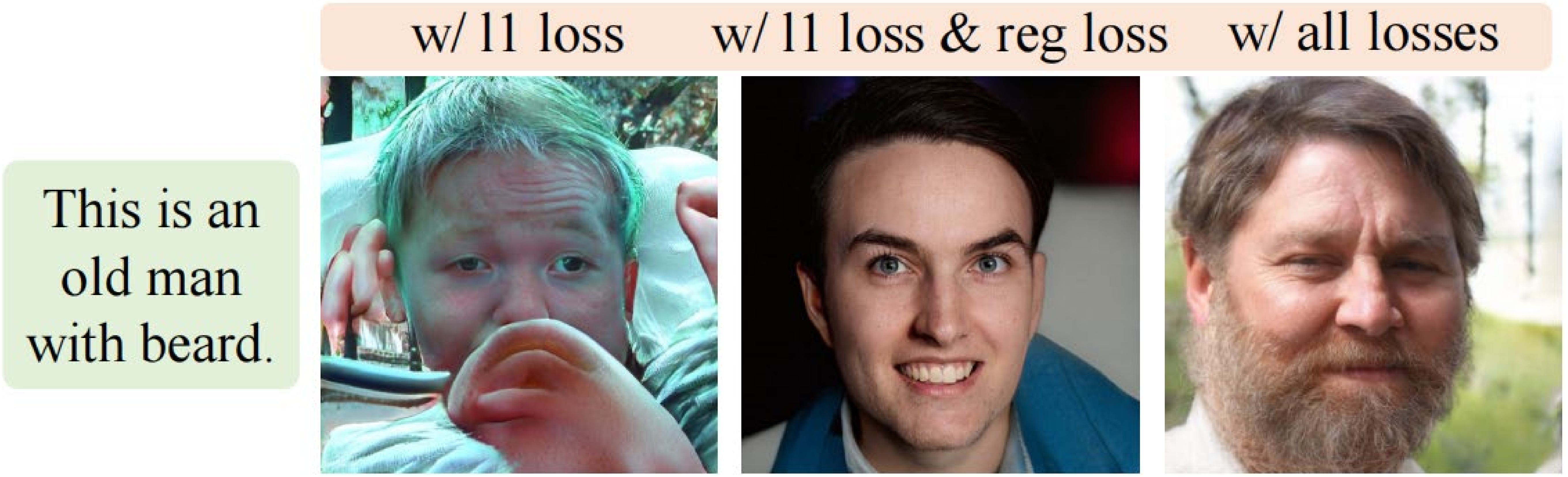
Our method can also be used to manipulate the generated images via the equation below:

$$CIE_{target} = CIE_{origin} + \alpha \cdot (CTE_{target} - CTE_{origin}) \quad (11)$$

- **Ablation Studies**

  The ablation consists of 2 parts.

  First, we demonstrate the efficiency of the proposed loss functions.



- **Ablation Studies**

The ablation consists of 2 parts.

Second, we demonstrate the efficiency of the proposed prompts.

Conclusion

- A framework to translate raw descriptions into images with high semantic consistency, quality and fidelity.
- The first to use prompt-based method to project text embeddings to image embeddings.
  - The method of using prompt embeddings.
  - The design of prompt embeddings.

Thank You

Presenter: Yiyang Ma
myy12769@pku.edu.cn