



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



Yiyang Ma<sup>1</sup>, Huan Yang<sup>2</sup>, Bei Liu<sup>2</sup>, Jianlong Fu<sup>2</sup>, Jiaying Liu<sup>1</sup> 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.

semantically aligned.

- translated images should be impressive.



- descriptions may have multiple meanings which are hard to be

Semantically aligned



Generate





## Existing works:

There's a trilemma among

- semantically alignment
- open-world words
- image quality

Our work aims at dealing with this trilemma.



- How to deal with these challenges? Pretrained large scale models! - challenge of semantics:
  - Contrastive Language-Image Pretraining (CLIP)
  - challenge of image quality: StyleGAN





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

Input Texts (1)  $\rightarrow$  CLIP Text Embeddings (CTEs) (2)  $\rightarrow$  CLIP Image Embeddings (CIEs)  $(3) \rightarrow StyleGAN Z Space Embeddings (SEs)$  $(4) \rightarrow \text{Translated Images}$ 



**Main Idea:** transmit semantics through the pretrained models: Input Texts (1)  $\rightarrow$  CLIP Text Embeddings (CTEs) (2)  $\rightarrow$  CLIP Image Embeddings (CIEs)  $(3) \rightarrow StyleGAN Z Space Embeddings (SEs)$  $(4) \rightarrow \text{Translated Images}$ 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.



- CLIP to StyleGAN

text embeddings to image embeddings



- CLIP has two latent spaces:
  - Text latent space
  - Image latent space
- pairs which have small cosine distances.

Semantically aligned text-image pairs will have embedding



Due to the character of CLIP, for two pairs of matched texts and images, we have:  $CTE_1 - CTE_2 = CIE_1 - CIE_2$  (1) If we can find a semantically aligned pair of representative embeddings, we can project input *CTE*s to corresponding *CIE*S.



The "representative" pair is a prompt pair to latent projection. We have:  $CIE_{input} = CIE_{prompt} + (CTE_{input} - CTE_{prompt})$  (2) In practice, we use:  $CIE_{input} = CIE_{prompt} + \alpha \cdot (CTE_{input} - CTE_{prompt})$  (3) To control the distinctiveness of the projection.





How to find the prompt embeddings? Because they are "representative", they should have the largest average cosine similarity to all the embeddings.

$$\max_{\boldsymbol{y}} \boldsymbol{z} = \frac{1}{n} \sum_{i=1}^{n} \frac{\boldsymbol{y} \cdot \boldsymbol{x}_{i}}{|\boldsymbol{y}| \cdot |\boldsymbol{x}_{i}|}$$
(4)  
$$s.t.|\boldsymbol{y}| = 1$$
(5)







The First Projection: Text Embeddings to Image Embeddings We can simplify Eqn. 4 as:

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

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,

$$y' = \frac{1}{n} \sum_{i=1}^{n} x_i, \ y = \frac{y'}{|y'|}$$
 (7)



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.".





## The Second Projection: CLIP Embeddings to StyleGAN Embeddings

The network architecture is shown below.



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



### The Second Projection: CLIP Embeddings to StyleGAN Embeddings The training loss consists of 3 parts.

- Basic constraint of the network:

 $\mathcal{L}_{l1} = ||SE|$ 

Semantic consistency loss:

 $\mathcal{L}_{sem\_cons} = CosDis(CIE_{input}, CLIP_I(G(SE_{pred})))$ (9)The regularization loss which ensures the predicted SE is in the StyleGAN

Z space:

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

The total loss is the combination of the three losses.

$$S_{pred} - SE_{true}||_1$$
 (8)



### 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

- 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.





**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 openworld words, we first show the translation results containing only the words of Multi-Modal CelebA dataset.

PCM-Frame (Ours) StyleCLIP TediGAN-B DF-GAN

She is black

woman, and she

has black hair.

This man has

brown hair. He

wears eyeglasses.

He is smiling.

She has black hair. She wears lipstick. She has bushy eyebrows.

This woman has straight hair and chubby face.









# 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.



She is a cute girl with white skin and big eyes.







### 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.

This is a church <u>in</u> <u>the dusk</u>. <u>Yellow</u> <u>and dim light</u> falls on the church. There is <u>no cloud</u> <u>in the sky</u>.

### Realistic Result



Illustration



Here is a <u>gloomy</u> church. This is a <u>Gothic</u> church with <u>spires</u>. The <u>sky</u> is <u>gray</u>.











This cat has <u>long</u> <u>hair</u>. Its <u>paws are</u> <u>straight and in</u> <u>front of its body</u>. Its <u>hair is orange</u>.

Here is a <u>fat</u> cat with <u>white and</u> <u>grey hair</u>. It looks <u>vigilant</u>. Its <u>ears</u> <u>stand straight</u>.

This is a cat with <u>black and white</u> <u>hair</u>. It <u>stands</u> <u>before a yellow</u> <u>wall</u>.





### Experiments

### **Manipulation Results on Generated Images** Our method can also be used to manipulate the generated images via the equation below: $CIE_{target} = CIE_{origin} + \alpha \cdot (CTE_{target} - CTE_{origin})$ (11)



An old man with beard.

Generated Image



A woman with black hair.

+black hair



A face. A face with black hair.  $\alpha = 0.5$ 

+bright skin



A face. A face with bright skin.  $\alpha = 0.4$ 





A face. A smiling face.  $\alpha = 0.3$ 





A face with an open mouth. A face.  $\alpha = 0.6$ 



A face with glasses. A face.  $\alpha = 0.25$ 

-happy



A happy woman A woman.  $\alpha = 0.5$ 



A face. A face with big eyes.  $\alpha = 0.2$ 

+glasses



A woman. A woman with glasses.  $\alpha = 0.15$ 



## **Ablation Studies** The ablation consists of 2 parts. First, we demonstrate the efficiency of the proposed loss functions.

This is an old man with beard.





## Ablation Studies

# The ablation consists of 2 parts. Second, we demonstrate the efficiency of the proposed prompts.

Image promptText prompt fromwith zeros as SECelebA-MultiModal

Lucie has pretty figure, a quantity of <u>golden hair</u> and a pair of <u>blue eyes</u>.

She has <u>brown</u> <u>eyes</u>, <u>pale skin</u>. She is famous for her fashionably small waist.





Our 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.





### ACM Multimedia 2022

Lisbon, Portugal | 10-14 October





### STRUCT @PKU Spatial and Temporal Restoration, Understanding and Compression



## **Presenter: Yiyang Ma** myy12769@pku.edu.cn

