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

Yiyang Ma†
Wangxuan Institute of Computer Technology, Peking University
Beijing, China
myy12769@pku.edu.cn

Huan Yang
Microsoft Research
Beijing, China
huayan@microsoft.com

Bei Liu
Microsoft Research
Beijing, China
bei.liu@microsoft.com

Jianlong Fu
Microsoft Research
Beijing, China
jianf@microsoft.com

Jiaying Liu†
Wangxuan Institute of Computer Technology, Peking University
Beijing, China
liujiaying@pku.edu.cn

ABSTRACT
AI illustrator aims to automatically design visually appealing images for books to provoke rich thoughts and emotions. To achieve this goal, we propose a framework for translating raw descriptions with complex semantics into semantically corresponding images. The main challenge lies in the complexity of the semantics of raw descriptions, which may be hard to be visualized (e.g., “gloomy” or “Asian”). It usually poses challenges for existing methods to handle such descriptions. To address this issue, we propose a Prompt-based Cross-Modal Generation Framework (PCM-Frame) to leverage two powerful pre-trained models, including CLIP and StyleGAN. Our framework consists of two components: a projection module from Text Embeddings to Image Embeddings based on prompts, and an adapted image generation module built on StyleGAN which takes Image Embeddings as inputs and is trained by combined semantic consistency losses. To bridge the gap between realistic images and illustration designs, we further adopt a stylization model as post-processing in our framework for better visual effects. Benefiting from the pre-trained models, our method can handle complex descriptions and does not require external paired data for training. Furthermore, we have built a benchmark that consists of 200 descriptions from literature books or online resources. We conduct a user study to demonstrate our superiority over the competing methods of text-to-image translation with complicated semantics.

CCS CONCEPTS
• Computing methodologies → Computer vision.

KEYWORDS
Text-to-image translation, Text-to-image semantic alignment, Embedding prompt

ACM Reference Format:
Yiyang Ma, Huan Yang, Bei Liu, Jianlong Fu, and Jiaying Liu. 2022. AI Illustrator: Translating Raw Descriptions into Images by Prompt-based Cross-Modal Generation. In Proceedings of the 30th ACM International Conference on Multimedia (MM ’22), October 10–14, 2022, Lisboa, Portugal. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3503161.3547790

Figure 1: Illustrations generation examples of the proposed AI Illustrator framework. The realistic images are further style transferred as the final illustration design. The descriptions are raw descriptions obtained from the Internet or excerpted from books. The major attributes are underlined. An interesting fact is that even Cho Chang is not explicitly expressed as an Asian girl, our method can infer this from her name and generate an ethnic Chinese girl.

1 INTRODUCTION
Characters or scenarios in books often attract readers to image what they really look like. People expect that there could be a method to translate raw descriptions into images in order to help
people imagine. The result images of translation must match the
descriptions in the aspect of semantics. In the past, this work can
only be done by humans, since there are two challenges. The first
challenge is the raw descriptions can be long and complicated, and
are hard to be visualized precisely. The descriptions not only can be
specific, e.g., skin color or hair length, but also can be abstract, e.g.,
the emotions or personalities. And there may be multiple semantics
in one description. The second challenge is that the images should
have fine-grained attributes which should be impressive. In this
work, we pay attention to dealing with such descriptions with long
sentences and complex semantics, and illustrate them with semantic
consistency and high quality.

Existing methods [9, 17, 18, 26, 29–31, 33, 43] of text-to-image
translation have achieved great results due to the recent cross-
modal researches. However, such complex descriptions are difficult
for them to handle. Xu et al. [31] and Tao et al. [26], use paired data
to train, so their performances heavily rely on the quality of datasets
and they cannot process those words out of the vocabulary. That
means it is nontrivial for them to translate raw descriptions in our
task. Xia et al. [30] and Patashnik et al. [17], have the ability to deal
with complex texts, while they fail in some cases (e.g., text contains
rich semantics, as shown in Figure 5) because their optimization
methods cannot take full advantage of pre-trained vision-language
models. Besides, optimization methods are also time-consuming
so it is unpractical to widely utilize these methods. In conclusion,
previous methods cannot resolve the two challenges we mention.

In this work, we propose a simple but effective framework named
**Prompt-based Cross-Modal Projection Framework (PCM-Frame)**
to translate raw descriptions into illustrations. Our method can
resolve the two challenges above because of the appropriate design
with two pre-trained models including StyleGAN and Contrastive
Language-Image Pre-training (CLIP). However, how to jointly lever-
age the two pre-trained models is nontrivial and challenging. That
is because there are enormous gaps between the different latent
spaces of StyleGAN and CLIP and it is hard to bridge these gaps.

To achieve this goal, we introduce two novel modules in our
method. First, a prompt-based projection module which projects
**Text Embeddings** to **Image Embeddings** is proposed. Prompt, as usu-
ally a method of adapting pre-trained models to downstream tasks,
has been extensively utilized in natural language processing (NLP)
problems [14]. In our work, they are used to connect the two lat-
tent spaces of CLIP, representing “a normal description” and “a
normal image” in order to be on behalf of the whole latent space.
CLIP encodes texts to **Text Latent Space** and images to **Image La-
tent Space**. Semantically aligned image-text pairs will be encoded
to embedding pairs which have high cosine similarity, and vice
versa. In this module, we take a specific pair of **Text Embedding** and
**Image Embedding** as “prompts” and migrate the input **Text Embed-
ning** which is extracted from the input text description to input
**Image Embedding** by the connection of prompt embeddings. Second,
we propose a module which projects **Image Embeddings** to Style-
GAN **Embeddings** and generates images from them by StyleGAN.
StyleGAN [7, 8] as one of the most notable Generative Adversarial
Network (GAN) [5] frameworks, helps us to generate images with
high quality. This module contains a network trained on paired
data which are randomly sampled. We randomly sample StyleGAN
**Embeddings**, generate images from them by pre-trained StyleGAN
and extract **Image Embeddings** from them by CLIP. The training
of this projection module does not require any external data. In
order to train the network with semantics maintained, we further
design combined semantic consistency losses for the training of
this network to ensure we can generate semantically accurate im-
ages by StyleGAN. At last, we use a style transfer method [28] to
cartoonize them so that these images can be used as illustrations
for books. Two results of our method are provided in Figure 1.

Besides, we build a benchmark consisting of 200 raw descriptions
of characters which are challenging to visualize. We evaluate the
performance of our method on some of the descriptions and conduct
a user study on these translation results. In summary, this work
has the following contributions:

- We propose a framework that can translate complicated raw
text descriptions into illustrations with high semantic consis-
tency, quality, and fidelity.
- We propose a novel prompt-based projection from Text Em-
beddings to Image Embeddings. We also propose a loss func-
tion which helps our framework keep the semantic consistency and the training process doesn’t require any external
paired data.
- Experiments demonstrate the superiority of our method. The
user study shows that more than 89% of 2200 votes from 22
subjects prefer our method to other state-of-the-art methods.

## 2 RELATED WORK

### 2.1 Text-to-Image Semantic Alignment

In order to translate a text into an image, checking whether the
text and the image are semantically aligned is necessary.

In the early years, works like [2, 12, 21, 31, 38, 39, 41] train a
pair of text encoder and image encoder separately to semantically
align them. Lee et al. [21] train two simple networks to extract
embeddings from texts and images respectively. For those pairs
of text and image which relatively semantically aligned, the em-
bedding pairs will have a higher dot product and vice versa. Xu
et al. [31] propose a method with an attention mechanism called
Deep Attentional Multimodal Similarity Model (DAMSM). Li et al.
[10, 11], Tao et al. [26], Zhu et al. [43] also use this model. All the
methods above need to be trained for a certain translation task. So,
their performance heavily relies on the quality of the datasets and
they cannot encode the words out of the vocabulary of the datasets.

The methods of extracting fine grained level features of images
[3, 42] are also inspiring. Following the success of BERT [1] in
language tasks, recent works typically use transformer [27] as
baseline models. A recent model, based on Contrastive Language-
Image Pre-training (CLIP) [19], builds two encoders of transformer
for a wide range of images and texts. CLIP is trained on 400 million
text-image pairs which are collected from a variety of publicly
available sources on the Internet. There are two latent spaces, one
for texts and another for images. In our work, we take CLIP as our
text-to-image alignment checking module.

### 2.2 Text-to-Image Translation

Due to the great development of GANs in many fields [32, 34–37, 40],
most existing text-to-image translation methods are GAN-
based. They can be roughly divided into two categories.


Figure 2: The entire framework of PCM-Frame. \(CLIP_f\) and \(CLIP_l\) denote the text encoder and image encoder of CLIP. Beforehand, in (a), we design \(CT_{\text{prompt}}\) and \(CE_{\text{prompt}}\) which are expressed as red dotted circles and used in (b). The \(CT_{\text{prompt}}\) is extracted by CLIP from a certain sentence while the \(CE_{\text{prompt}}\) is obtained from a big set of \(CLIP_s\) (expressed as red circles). At inference time, the input raw text description is encoded to \(CT_{\text{input}}\) by CLIP and projected to the corresponding \(CE_{\text{input}}\) in (b). Then, the \(CE_{\text{input}}\) is projected to the corresponding \(SE\) from which we can generate a semantically aligned realistic image by StyleGAN in (c). At last, the generated realistic image can be further style transferred as the final illustration design. The specific architecture of \(C2S\) projection network is shown in Figure 3.

The first category doesn’t use pre-trained generation models, e.g., StyleGAN \([7, 8]\). These methods train an image generator themselves. The pioneering work of Reed et al. [21] approaches text-to-image translation by training a conditional GAN \([16]\) with text embeddings extracted from a pre-trained text encoder. Xu et al. [31] introduce an attention module between images and texts. Following the attention module which is proposed by Xu et al. [31], Zhu et al. [43] introduce a memory writing gate and Tao et al. [26] propose a backbone that generates images directly by Wasserstein distance. Ramesh et al. [20] build a large model with over 12-billion parameters and show a great diversity of text-to-image translation.

The second category uses existing generation models so that the quality of their generated images is better and the training process can be shorter. But because of the domain limitation of the existing generation models, the images they generate are limited in certain domains. Xia et al. [29] map input text to StyleGAN latent space, while Xia et al. [30] use cosine similarity of text and image embeddings encoded by CLIP as a loss function to optimize an embedding in StyleGAN latent space. Due to the usage of CLIP, Xia et al. [30] can process texts with more complex semantics. But its performance is random and visually unpleasant. Patashnik et al. [17] propose three methods to manipulate an existing image. The first method of latent optimization they propose can be used as an image generation method by giving an initial image. Gal et al. [4] transfer images to new text-guided domains by fine-tuning StyleGAN. We use StyleGAN2 [8] in our work.

2.3 Prompt Method in NLP Domain

Prompt is a recently proposed method to “re-formulate” downstream tasks so that these tasks can be resolved by pre-trained models [14]. It can be regarded as an “intermediate”. It is first introduced to solve NLP problems. Liu et al. [14] introduce such an example: when recognizing the emotion of a social media post, “I missed the bus today”, we may continue with a prompt “I felt so ___”, and ask the language model (LM) to fill the blank with an emotion-bearing word instead of giving the LM only the sentence “I missed the bus today,” which do not have a precise task. In this way, the downstream task is re-formulated so that a pre-trained LM can handle. Without such “prompt”s, if we want to leverage pre-trained models on a downstream task, the models have to be finetuned on the corresponding data, while finetuning is time-consuming and the performances heavily rely on the quality of the dataset.

In NLP domain, there have been many automatic methods \([13, 22, 25]\) of designing prompts for certain tasks instead of manually specifying \([15, 24]\). We refer the reader to the survey \([14]\) for extensive exposition and discussion on prompt. The methods of prompt have been widely used in NLP domain. No attempt of applying the prompt method in Cross-Modal problems or Computer Vision problems so far. In our work, we leverage the idea of prompt and propose a specific pair of embeddings as “prompt”s to help us bridge the two latent spaces of CLIP. Our method is a novel understanding of prompts which is different from the methods proposed in NLP.

3 PROMPT-BASED CROSS-MODAL GENERATION FRAMEWORK

In this section, we will introduce our AI Illustrator framework that can translate raw text descriptions into vivid illustrations. Our framework consists of two major modules which is shown in Figure 2. The first module in (b) projects Text Embeddings (abbreviated as \(CTE\) and \(C\) denotes CLIP) to Image Embeddings (abbreviated as \(CIE\) and \(C\) denotes CLIP) with prompt embeddings and the second module in (c) generates images from the projected \(CIE_s\). In particular, we encode the input text to \(CTE_{\text{input}}\) using CLIP and project it to \(CIE_{\text{input}}\) via the first module with a pair of prompt embeddings.
Then, via the second module, the CIEinput is projected to the corresponding StyleGAN Z Embedding (abbreviated as SE, and we use StyleGAN2 in our work) and we generate an image from it which is semantically aligned to the original text by pre-trained StyleGAN. To bridge the gap between realistic images and illustrations, we style transfer the generated images at last. We do not build a direct projection from input text to SE because there are few paired data. So, we take CIE as a transition to connect the texts and SEs.

3.1 Prompt Embedding Design

As we point out in Section 1, CTEprompt and CIEprompt are used to bridge the text and image latent spaces of CLIP and these prompts should represent “a normal description” and “a normal image” respectively. Otherwise, there may be a big distance between the prompt pair and the target pair which may lead to failure. So, it is important to assign an appropriate pair as prompts. Our prompt design is shown in the Figure 2 (a). The specific method of using these prompts will be further explained in Section 3.2.

To make sure that the prompt can represent all data, the prompt embedding should be extracted from a big enough set of data. We assume that the prompt embedding should have the largest average cosine similarity to all the embeddings in the set. Their lengths are all normalized because only their orientations contain semantics. Taking $\mathbf{y}$ to denote the prompt embedding and $\mathbf{x}_i$ to denote the $i$-th the embeddings in the set, the problem can be formulated as

$$\max \frac{z}{y} \sum_{i=1}^{n} \frac{\mathbf{y} \cdot \mathbf{x}_i}{\|\mathbf{y}\| \cdot \|\mathbf{x}_i\|} \tag{1}$$

$$s.t.\|\mathbf{y}\| = 1, \tag{2}$$

where $\cdot$ denotes the dot product of vectors, $n$ denotes the number of data in the set and $z$ denotes the average cosine similarity between the prompt embedding and all other embeddings. Note that this is a non-linear programming problem which is hard to resolve directly. To address the above issue and obtain prompt embeddings, we propose to find the physical meanings of the equations which will help us to solve this problem.

Because the lengths of all of the embeddings are all normalized, Equation 1 can be simplified to

$$\max z = \frac{1}{n} \sum_{i=1}^{n} \frac{\mathbf{y} \cdot \mathbf{x}_i}{\|\mathbf{y}\| \cdot \|\mathbf{x}_i\|} \tag{3}$$

From Equation 3, by using associative law and commutative law of addition and multiplication, we can get

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

The above equation represents a hyperplane and $z$ is the constant parameter. The farther the hyperplane is from the origin point, the larger the absolute value of $z$ is. And as the feasible region of this problem is a hypersphere with symmetry which is shown in Equation 2, we can move this hyperplane as far as possible if we want to maximize $z$. It is easy to see that $z$ will be the largest when the hyperplane and the hypersphere are tangent and $\mathbf{y}$ is the unit normal vector of the hyperplane at this moment. Through analytic geometry, the unit normal vector of the hyperplane is a vector as shown below:

$$\mathbf{y}' = \frac{1}{\sum_{i=1}^{n} \mathbf{x}_i} \mathbf{y} = \frac{\mathbf{y}}{\|\mathbf{y}\|} \tag{5}$$

It can be seen that the vector $\mathbf{y}'$ is the arithmetic average of all embeddings with unit length.

Specifically, for the image prompt, we can build a big set of images by randomly sampling in Z space of StyleGAN and generating images from such latent embeddings. For the text prompt, because it’s hard to get a big set of descriptive texts with a large semantic range and the text itself is a carrier of semantics, we can manually specify a certain sentence like the methods mentioned in [14]. For example, for text-to-human face translation, we can simply set sentence “A normal human face.” and extract its CTE as CTEprompt. This choice is further discussed in our ablation Section 4.4. But, if there is a text set which has high enough quality, our method can be used to obtain a better CTEprompt.

3.2 Embedding Projection with Prompt

In this section, we encode the input text to CIEinput and manage to obtain the corresponding CIEinput via the first module shown Figure 2 (b) with the prompts we get in last section. The prompts can be regarded as a pair of “intermediate’s”. To be more specific, the CTEprompt is subtracted from the input CIEinput, then this difference is added to the CIEprompt. The result is the semantically corresponding CIEinput of the input text. This module is shown in Figure 2 (b) and the projection can be formulated as

$$CIEinput = CIEprompt + (CTEinput - CTEprompt) \tag{6}$$

From the perspective of prompt, the summation can be regarded as “re-formulation” as Liu et al. [14] propose. And in contrast, we propose that the subtraction can be regarded as “de-formulation”. This projection makes our next module of projection between CIEs and SESs can process CIEinput without fine-tuning on CTE-SE pairs which are not easy to obtain. The validity of the linear operations in Equation 6 is supported by the character of CLIP. Further discussions are provided in supplementary materials.

The inference time of this module only uses linear operations among CTEinput, CTEprompt and CIEprompt to get CIEinput. That means this module can run quite efficient and stably.

In our work, the difference between CTEinput and CTEprompt will be multiplied by a constant to control the distinctiveness. So, the actually used method is

$$CIEinput = CIEprompt + \alpha \cdot (CTEinput - CTEprompt) \tag{7}$$

where $\alpha$ is the constant. It can vary in the range of 1 to 2 and is set to 1.75 which is capable for most descriptions empirically.

3.3 Image Generation from Embedding

After getting the CIEinput, we manage to generate the corresponding image from it. In the module shown in Figure 2 (c), we map the CIEinput we get above to the semantically aligned SE from which we can generate the image we expect. We take a combination of fully connected layers with dense connection as this projection and this network is named as CLIP-to-StyleGAN (C2S) projection network. The architecture of this network is shown in Figure 3. To
train this network, we propose to generate CIE-SE pairs. As the distribution of \( Z \) space of StyleGAN is standard normal distribution determinately, we can randomly sample SEs in \( Z \) space and generate images from these SEs by pre-trained StyleGAN. Thus, we can extract CIEs from these images by CLIP. In this way, we can get theoretically infinite CIE-SE pairs to train our network.

The key to this projection is to keep the semantics of CIEinput, so the loss of this network should be designed purposefully. Besides, we should also guarantee that the SEs are within \( Z \) space so that we can generate images by StyleGAN. In order to optimize the above network, we propose the following combined loss functions.

In order to keep the semantics of CIEinput, we can use CLIP to check the semantic alignment of the generated image and CIEinput. In particular, we extract the CIE from the generated image as \( CIE_{rebuild} \) and design a loss function called \( L_{sem\_cons} \) to minimize the cosine distance between \( CIE_{rebuild} \) and \( CIE_{input} \). The subscript of the loss function is short for Reconstructing Semantics Consistency. Taking \( G \) to denote pre-trained StyleGAN and \( CLIP_{I} \) to denote the image encoder of CLIP, this loss is calculated by

\[
L_{sem\_cons} = \text{CosDis}(CIE_{input}, CLIP_{I}(G(SE_{pred}))).
\]

(8)

For the basic constraint of the network, we use a \( l1 \) loss function called \( L_{l1} \) between \( SE_{pred} \) and \( SE_{true} \). It is calculated by

\[
L_{l1} = ||SE_{pred} - SE_{true}||_1.
\]

(9)

Besides, the network should make sure that the predicted \( SE_{pred} \) is within the \( Z \) space of StyleGAN. Otherwise, StyleGAN cannot generate images from the \( SEs \) out of the latent space. Because the distribution of \( Z \) space is clearly a standard normal distribution, all the \( SEs \) should have mean values of 0 and standard deviations of 1. This character helps us to design a regularization loss called \( L_{reg} \) to ensure the generated \( SEs \) are all within \( Z \) space easily and this is the reason why we use \( Z \) space. This loss is calculated by

\[
L_{reg} = ||\text{mean}(SE_{pred})||_1 + ||\text{std}(SE_{pred}) - 1||_1.
\]

(10)

To sum up, the total loss of our network is shown below.

\[
L = \lambda_{sem\_cons} \cdot L_{sem\_cons} + \lambda_{l1} \cdot L_{l1} + \lambda_{reg} \cdot L_{reg}.
\]

(11)

where \( \lambda_{sem\_cons}, \lambda_{l1} \) and \( \lambda_{reg} \) are the corresponding weights of all the losses. These values are 1.0, 0.3, 0.3 respectively.

After getting the \( SE \), we can simply generate the corresponding image from it by StyleGAN. But there is still a gap between the generated images and illustrations because most illustrations are more abstract than realistic images. In order to bridge this gap, we adopt an existing stylization method [28] to cartoonize the images at last and get the final design of illustration.

### 4 EXPERIMENTAL RESULTS

In this section, we show a lot of experimental results to demonstrate the superiority of our framework and evaluate the effectiveness of the modules we propose. The implementation details to ensure the reproducibility are provided in the supplementary materials.

#### 4.1 Baseline Methods

To demonstrate the superiority of the method we propose, we compare our translation results with state-of-the-art methods including DF-GAN (CVPR 2022) [26], TediGAN-B (arXiv 2021) [30] and StyleCLIP (ICCV 2021) [17]. StyleCLIP, as a work aiming at image manipulation, can also be used for text-to-image translation with the proposed first technique of latent optimization by assigning an origin latent. TediGAN-B and StyleCLIP can process complicated texts due to the usage of CLIP, while DF-GAN cannot. For the task on human faces, we retrained DF-GAN on the Multi-Modal CelebA-HQ dataset Xia et al. [29]. Other previous methods which need to be retrained e.g., DM-GAN (CVPR 2019) [43], SD-GAN (CVPR 2019) [33] and AttnGAN (CVPR 2018) [31] get approximate results to DF-GAN and they also cannot process the words out of the vocabulary of the dataset, so we only compare with DF-GAN.
4.2 Comparisons to State-of-the-Art Methods

Qualitative Comparison. First, we qualitatively compare our results with the results of other methods. To demonstrate the ability to translate the input text description to fine-grained images without the interference of style transfer, all the results of this section are not style transferred. Because DF-GAN can only process the words within the vocabulary of training, we first translate sentences with simple words which appear in the Multi-Modal CelebA-HQ dataset in order to compare with DF-GAN fairly. The results are shown in Figure 4. It can be seen that our method performs much better than other works. We translate all the semantics of the input text and the images are visually pleasing, while the results of other methods cannot guarantee, e.g. we translate the description of “black woman” and the description of “eyeglasses” successfully while other methods fail. Then we show translation results on relatively complex texts in Figure 5. There are many descriptions which are hard to be visualized or do not have direct relations to the human face, e.g., “monk” and “16-18 years old”. We translate those indirect descriptions into details in the result images and exclude irrelevant semantics while other methods like StyleCLIP fail.

We show the diversity of our results in Figure 6 because one description may match multiple images. We translate one text into diverse images by style mixing a random SE in the first few layers of StyleGAN. This is a special mechanism supported by StyleGAN. In order to make every result shown in this paper reproducible, all other results are generated without any random style mixing.

We also conduct a user study on the results for human faces of both simple and complicated texts. The users are asked to judge which one is the most photo-realistic (abbreviated as Real. Prefer.) and the most semantically aligned (abbreviated as Acc. Prefer.) to the given text. The cases with simple words contain 20 text-image pairs and cases with complicated words contain 80 text-image pairs. A total of 22 subjects participates in this user study and 2200 votes are collected. The results are shown in Table 1. Quantitative Comparison. We evaluate inception score (IS) [23] to compare the diversity and quality of the generated images in Table 2. All the results are calculated from 100 samples except DF-GAN. The result of DF-GAN is calculated from 20 samples.

4.3 Illustration Results of Raw Descriptions

The final goal of our work is to automatically generate illustrations from raw descriptions. In order to demonstrate the ability of our framework on such tasks, we show more translation results of book characters in Figure 7 and Figure 1. At the same time, we offer the style transferred images of them. The style transfer method is provided by Wang and Yu [28]. These transferred images can be directly used as illustrations for books. Some of the descriptions are excerpted from books and some of them are obtained from the Internet. These descriptions contain very complex semantics.

|                  | Ours | StyleCLIP | TediGAN-B | DF-GAN |
|------------------|------|-----------|-----------|--------|
| Com. Acc. Prefer. (%) | 90.6 | 5.1       | 4.3       | -      |
| Real. Prefer. (%)    | 78.3 | 20.8      | 0.9       | -      |
| Sim. Acc. Prefer. (%)| 83.8 | 9.5       | 1.9       | 4.8    |
| Real. Prefer. (%)    | 75.5 | 23.8      | 0.2       | 0.5    |

Table 1: The user study for translation results on text descriptions. “Com.” denotes “with Complicated Words” and “Sim.” denotes “with Simple Words”. The best numbers are bold. For descriptions with complicated words, there are 1760 votes. For descriptions with simple words, there are 440 votes.

|                  | Ours | StyleCLIP | TediGAN-B | DF-GAN |
|------------------|------|-----------|-----------|--------|
| IS ↑             | 3.229| 1.323     | 3.191     | 2.503  |

Table 2: Inception score comparison of generated results from different methods. ↑ means the higher the better.
“Nami is a Swedish girl. She is a slim young woman of average height with orange hair and brown eyes. Most people consider her to be very attractive or even beautiful.”

“Celie is a young black girl living in Georgia in the early years of the twentieth century. She is largely uneducated; her letters to God are written in non-standard dialect.”

“Hamlet is angry, dejected, depressed, and brooding; he is manic and energetic, a man who loathes himself and his fate.”

“Santiago is an aging, experienced Cuban fisherman who has gone eighty-four days without catching a fish. He has thick beard.”

“Hamlet is angry, dejected, depressed, and brooding; he is manic and energetic, a man who loathes himself and his fate.”

“Nami is a Swedish girl. She is a slim young woman of average height with orange hair and brown eyes. Most people consider her to be very attractive or even beautiful.”

| Description | Realistic Result | Illustration |
|-------------|------------------|--------------|
| “Nami is a Swedish girl. She is a slim young woman of average height with orange hair and brown eyes. Most people consider her to be very attractive or even beautiful.” | | |
| “Celie is a young black girl living in Georgia in the early years of the twentieth century. She is largely uneducated; her letters to God are written in non-standard dialect.” | | |
| “Hamlet is angry, dejected, depressed, and brooding; he is manic and energetic, a man who loathes himself and his fate.” | | |
| “Santiago is an aging, experienced Cuban fisherman who has gone eighty-four days without catching a fish. He has thick beard.” | | |

| Description | Realistic Result | Illustration |
|-------------|------------------|--------------|
| “This is a church in the dusk. Yellow and dim light falls on the church. There is no cloud in the sky.” | | |
| “Here is a gloomy church. This is a Gothic church with spires. The sky is grey.” | | |
| “The church has white walls and black roof. It has one tall tower. The sky is as blue as the sea.” | | |
| “This is a cat with black and white hair. It stands before a yellow wall.” | | |
| “This cat has long hair. Its paws are straight and in front of its body. Its hair is orange.” | | |
| “Here is a fat cat with white and grey hair. It looks vigilant. Its ears stand straight.” | | |

Figure 7: Translation results for the face of famous characters on raw descriptions. The images of the left ones of each pair are our realistic image results while the right ones are style transferred illustrations as the final design. The descriptions are all raw descriptions which are obtained from the Internet or excerpted from books. The major attributes are underlined.

Figure 8: Translation results for non-face descriptions. The images of the left ones of each pair are our realistic image results while the right ones are style transferred illustrations as the final design. The major attributes are underlined.
It’s clear that our method successfully translates the underlined attributes, whether they’re specific or abstract. And those less relevant parts of texts without underlines do not have negative effects on our results. Besides, in Figure 1, there’s an interesting fact that our method translates the text description of Cho Chang into an Asian girl because our method infers this from her Chinese name, Cho. In order to demonstrate our capability of translating general descriptions, we also show similar illustration results on churches and cats which are non-face cases in Figure 8.

4.4 Ablation Study

There are two main factors that have effects on the quality of our work, the method of getting prompt and the loss functions of the C2S projection network. We demonstrate their effectiveness by giving an ablation study. In Figure 9, we show the results without the proposed losses. In Figure 10, we show ablation study on prompt design. As we discuss in Section 3.1, prompts should represent a normal semantic, we compare our design to several other designs which may contain the semantic of “normal”. For the image prompt, we use the average cosine distance between captions from Multi-Modal CelebA-HQ.

4.5 Failure Cases and Discussions

There are two kinds of failure cases of the proposed method. First, if the image we expect is out of the generation domain of StyleGAN, it is hard to be generated. Second, if there is more than one person described in one input, our framework may be confused and generate an image that contains attributes from not only one person. These failure cases are shown in Figure 11.

We consider that there may be two reasons for these failures. First, there’s a limitation of the generation domain of StyleGAN. Second, there’s still room for improvement of our framework to better leverage CLIP embeddings of texts with complicated texts.

5 CONCLUSIONS

We have proposed a framework for illustrating complicated semantics. Our framework is able to deal with various text inputs and generate impressive images with high quality, fidelity, and semantic alignment to the input texts due to our appropriate design with pre-trained models including CLIP and our StyleGAN. Our method is a general method that can be used to translate multiple kinds of objectives, e.g., human faces, churches, and cats. Extensive experiments on different kinds of input text descriptions demonstrate the superiority of our method. User study also shows that most people prefer our method to other state-of-the-art methods because of the visually pleasant results and semantic accurateness.
