Exploring Enhancements to Text-to-drawing Methods Based on vision-language models

Erin Ching-Hsuan Ho
x715106@stanford.edu, hchings@gmail.com

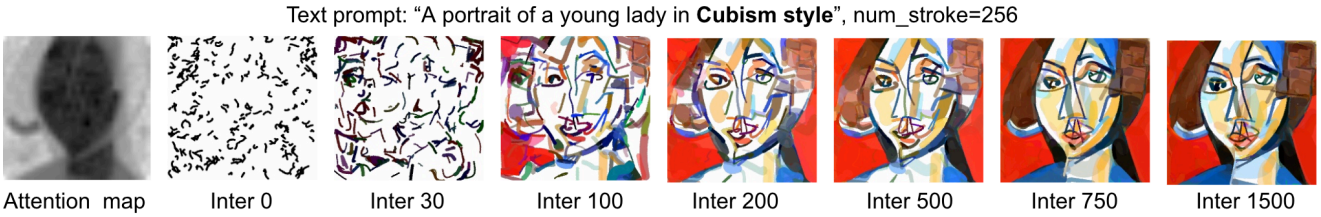

Figure 1. Visualization of the drawing process of the proposed method. Leftmost: Attention map used for guiding the optimization process. Rightmost: Final synthesized drawing.

Abstract

In the image generation domain, text-to-image has had significant progress over the past few years with the publications of deep learning models such as Diffusers[1] and DALL-E[2]. However, text-to-drawing synthesis, where outputs are composed of Bézier curves rather than pixels, remains comparatively underexplored. This paper examines the limitations of existing optimization-based text-to-drawing methods and introduces enhancements in initialization strategy and loss function, leading to improved visual outcomes.

1. Introduction

Text-to-drawing is a domain that focuses on generating sketches like human drawings, presenting a different style than high-fidelity images generated from diffusions or DALLA-E. While sketching might seem easier to generate given its lower fidelity nature compared to realistic images based on pixels, creating abstractions is difficult for machines to achieve.

The latest advancements in text-to-drawing leverage inference time optimization (e.g., CLIPDraw[3] and CLIPasso[4]), where the control points of Bézier curves are optimized based on the loss function during inference. To bridge the semantic gap between the input text prompt and the synthesized drawing, these methods commonly use CLIP (Contrastive Language-Image Pre-training)[5], a neural network that excels at encoding the semantic meaning of visual depictions, in its loss function. Despite CLIP's strong cross-modal understanding and retrieval capabilities, current CLIP-based text-to-drawing methods are prone to produce drawings that are visually unappealing and cluttered with messy sketches (A few examples in Figure 3 bottom row). This raises the question of whether comparing the semantics of the input text prompt and the synthesized drawing using CLIP is effective within this optimization framework.

To close out this gap, this paper investigates the challenges of CLIP-based text-to-drawing methods, using CLIPDraw as the baseline, and explores improvements. Section 2 provides a brief overview of related works, while Section 3 outlines our proposed method. In Section 4, we present our result analysis. The key findings are summarized as follows:

- **Impact of Initialization**: The initialization strategy can significantly affect the quality of the final drawing, depending on the input text prompt. Random initialization, as used in CLIPDraw, often results in suboptimal and messy drawing. In our proposed methods, we utilize either an attention map from the UNET of a pretrained Stable Diffusion model or Canny edge detection of the sample image from the diffusion model for initialization. Experiments show that by improving the placement of initial strokes, we can produce cleaner and more semantically meaningful drawings.

- **CLIP Loss**: The CLIP-based loss function, which compares the cosine similarity between the CLIP embeddings of synthesized drawing and the input text prompt, is helpful. However, it alone is insufficient to produce aesthetically pleasing and recognizable drawings, especially when the semantic complexity of the input text prompt increases.

- **Perceptual Loss**: Incorporating a perceptual loss such as LPIPS[6] of the synthesized drawing and a guided image, alongside the CLIP loss, significantly improves the quality of the generated drawings. This combination helps in capturing both the semantic meaning and the aesthetic appeal, effectively guiding the drawing synthesis process.

- **Complex Scene Handling**: Our method demonstrates improved capability in handling complex scenes compared to CLIPDraw. By integrating perceptual loss and attention map-guided initialization, it can generate more detailed and contextually accurate drawings. While style transfer is not our goal, the proposed method can capture styles via the input text prompt (Figure 7).

2. Related Work

In recent years, the field of drawing synthesis have shifted from a supervised training approach using text-to-image generative models, to the synthesis-through-optimization paradigm, which eliminates the need for training data. Notable contributions using this paradigm include CLIPDraw[3], CLIPasso[4], CLIPascene[7], StyleCLIPDraw[8], and DiffSketcher[9], each of which leverages pre-trained models like CLIP or Stable Diffusion with slightly different optimization loop architecture to synthesize drawings from text or images based on Bézier curves, achieving varying degrees of abstraction and artistic goals.

CLIP (Contrastive Language-Image Pretraining) aligns images and textual descriptions within a shared latent space using contrastive learning. This powerful alignment allows CLIP to understand and generate images based on textual input, making it a robust tool for image synthesis and manipulation. By optimizing parameters to maximize the similarity between generated images and text prompts, CLIP can produce visually coherent and contextually appropriate images without extensive dataset-specific training.

Stable Diffusion utilizes denoising diffusion probabilistic models (DDPMs) to generate images through a series of refinement steps, transforming noisy data into coherent images. This method excels in text-to-image synthesis, producing high-quality, photorealistic images from textual descriptions. Stable Diffusion models optimize a set of parameters guided by text prompts, integrating diffusion model guidance into the image generation process.

CLIPDraw presents an innovative approach to text-to-image synthesis by utilizing the

pretrained CLIP (Contrastive Language-Image Pretraining) model to guide the creation of vector-based drawings from natural language descriptions. Unlike traditional methods that require extensive training on specific datasets, CLIPDraw operates by optimizing a set of Bézier curves directly, ensuring the generated sketches are aligned with the given textual prompt. This method stands out for its ability to generate drawings with out-of-the-shelf pretrained models, opening a stream of related research that uses CLIP to compare the semantic distance of the synthesized drawing versus the input text prompt. **StyleCLIPDraw** further adds VGG net for enabling style transfer.

CLIPasso tackles a slightly different problem domain by focusing on representing a pixel image with as minimal strokes as possible. It takes an image instead of an input text prompt as input. This method also uses the CLIP model but emphasizes geometric and semantic simplification to produce sketches at varying levels of abstraction. Like CLIPDraw, CLIPasso defines sketches through Bézier curves and employs a differentiable rasterizer to optimize these curves with respect to a CLIP-based perceptual loss. CLIPasso aims to create recognizable and structurally sound sketches without relying on specific sketch datasets. However, the application of this method is rather limited. It only works when the input image only contains a single and very simple object without any background.

CLIPascene aims to address CLIPasso's limitation that the input image cannot be beyond a single object. It introduces a methodology for converting scene images into sketches with by separating the image into foreground and background regions and then synthesizing sketches each independently. Specifically, it trains two MLP networks to learn stroke locations and remove select strokes, making SceneSketch capable of sketching a more detailed and contextually rich image compared to previous works. However, this method has an inherent constraint where the input images must contain easily separable foreground and background.

DiffSketcher focuses on producing pencil-drawing-like free-hand sketches from a natural language input text prompt by leveraging the power of pre-trained text-to-image diffusion models. This method optimizes Bézier curves using an extended version of the score distillation sampling (SDS) loss, allowing the diffusion model to guide the sketch synthesis process. The resulting sketches maintain high recognizability and structural integrity.

3. Method

We formulate the text-to-drawing problem as below: Given a short text prompt P and a set of hyperparameters including number of strokes, max stroke width, num of control points per stroke, and canvas size, output a synthesized drawing D of the canvas size that is composed of a set of RGB Bézier curves compliant with the stroke hyperparameters.

Algorithm 1 The proposed text-to-drawing algorithm

Input: Input text prompt P; Number of strokes N; Pre-trained CLIP and Stable Diffusion model Φ; Perceptual loss coefficient; Hyperparameters such as *iterationNum*, *canvasSize*, and *maxStrokeWidth*.

Begin:

$EmbedPrompt = \text{CLIP}(P)$
$\{Curves_0, \ldots, Curves_N\} = \text{InitializeByAttentionMap}(\Phi.\text{attentionMap})$
for $i = 0$ to $iterationNum$ **do**
  $Image = \text{DiffRender}(Curves)$
  $EmbedAugmentedImg = \text{CLIP}(\text{Augment}(Image))$
  $CLIPLoss = -\text{CosineSim}(EmbedPrompt, EmbedAugmentedImg)$
  $PerceptualLoss = \text{LPIPS}(\Phi.\text{sampleImage}, Image)$
  $Loss = CLIPLoss + PerceptualLoss * perceptualLossCoeff$
  Backpropagate: $Curves \leftarrow \text{Minimize}(Loss)$
end for

To get the final drawing D, we followed the synthesize-through-optimization paradigm with improvements on the initialization method and the loss function. Figure 2 provides an overview of the optimization loop architecture and Algorithm 1 showcases the pseudo-code of the proposed method. High-level steps are as follows:

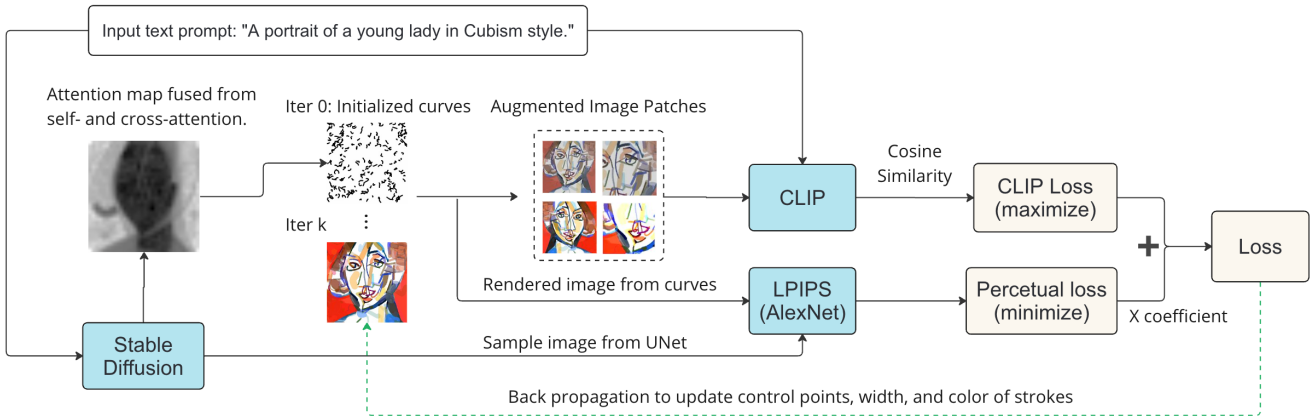

Figure 2. A high-level view of the gradient decent loop structure of the proposed method. Blue boxes are pre-trained models that are frozen. Once loss is calculated, it runs back propagation to update curves to form a new synthesized drawing.

a. Given the text prompt *P*, we first sample an image from a pre-trained Stable Diffusion model Φ. We offer two initialization strategies: The first and recommended method places initial strokes based on the attention map of the U-Net in Stable Diffusion. The second method places initial strokes around the Canny edges of the sample image instead. All strokes are initialized as black.

b. With the initialized drawing, we begin the optimization loop. At each iteration *i*, we calculate the loss based on the CLIP loss between the augmented synthesized drawing *Di* and the input text prompt *P*, as well as the perceptual loss between the synthesized drawing *Di* and the sample image from the diffusion model. The image augmentation pipeline contains operations such as color jitter, random crop, and random resize to improve the outcome.

c. We perform backpropagation to update the control points, stroke width, color, and opacity of the Bézier curves to synthesize a new drawing. This process is repeated until the loss curve converges.

In the following sub-sections. We discuss the key improvements of the proposed method from CLIPDraw.

3.1 Initialization Strategy

The randomized initialization used by ClipDraw is prone to undesired results because the objective function of text-to-drawing is highly non-convex, and therefore strokes can easily converge into local minimums. Moreover, since the algorithm synthesizes drawing at inference time, once the strokes are stuck at local minimums, there is very little the model can do to correct the drawing synthesis process (Figure 3 bottom row shows examples of undesired results).

To address this, we experimented with two alternative initialization strategies. The first approach is to use the attention map of a vision model. While we can use ViT (Vision-Transformer), here we choose to use Stable Diffusion, as it will be used to generate the prior for the drawing synthesis later. The attention-map initialization process places initial curves based on the fused product of the cross-attention map and the self-attention.

The second initialization approach leverages Canny Edge detection [10] of the sample image from the diffusion model. This approach is less ideal, as the Canny Edges of an image do not indicate the semantic importance. For example, in the right image of Figure 4, many detected Canny Edges are depicting the background,

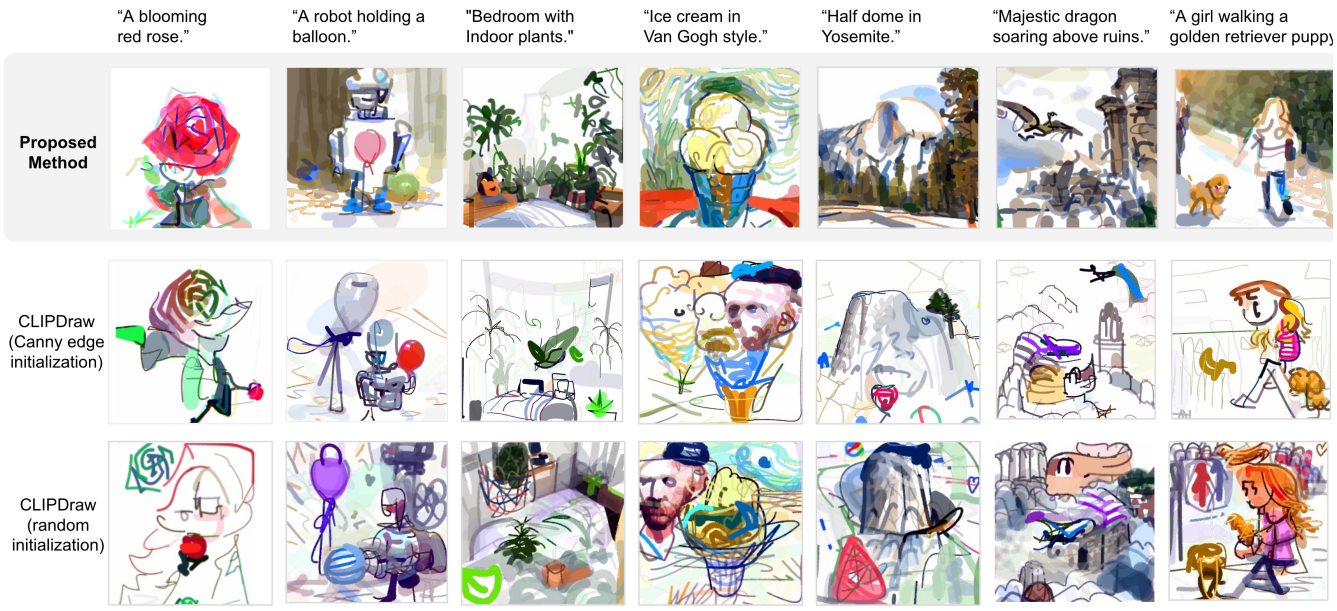

Figure 3. Various drawings synthesized by the proposed method compared to those generated by CLIPDraw with Canny Edge initialization and the original CLIPDraw (random initialization) using the same input prompt. All results in this comparison used the same number of strokes of 128, no negative text prompts, and the same image augmentation pipeline.

contributing little to the semantic meaning of the input text prompt.

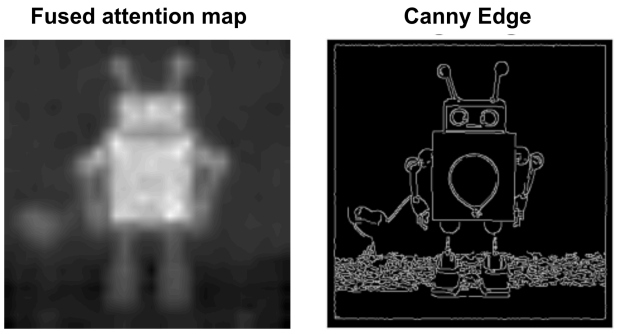

Figure 4. Example of fused attention map and Canny Edge of the same input text prompt.

3.2 Loss function

While drawing-through-synthesis methods are sensitive to the initial placement of Bézier curves, improving the initialization method alone has a limited impact on the quality of the final drawing. Through experiments, we identified the CLIP-based loss function adopted by CLIPDraw and its subsequent research as the root cause for undesired results.

To recap, the CLIP-based loss proposed by CLIPDraw first derives the embeddings of the input text and the augmented drawing from the CLIP model at an iteration. It then updates the loss function by subtracting the cosine similarity of those two CLIP embeddings. The augmentation step typically involves random perspective, color jitter, and random resize cropped to improve the result. While CLIP is highly effective at capturing semantic relationships between images and text, the CLIP-based loss which purely compares the semantic similarity of the synthesized drawing with that of the input text prompt is far from sufficient. As an example, CLIP will still assign a high similarity score to a messy, ugly drawing of roses (bottom row and the first column of Figure 3), despite that the drawing is not appealing to human perception.

With this observation, we complement semantic loss from CLIP with perceptual loss to capture aesthetics in the optimization process. The perceptual loss is calculated based using LPIPS (Learned Perceptual Image Patch Similarity) [6],

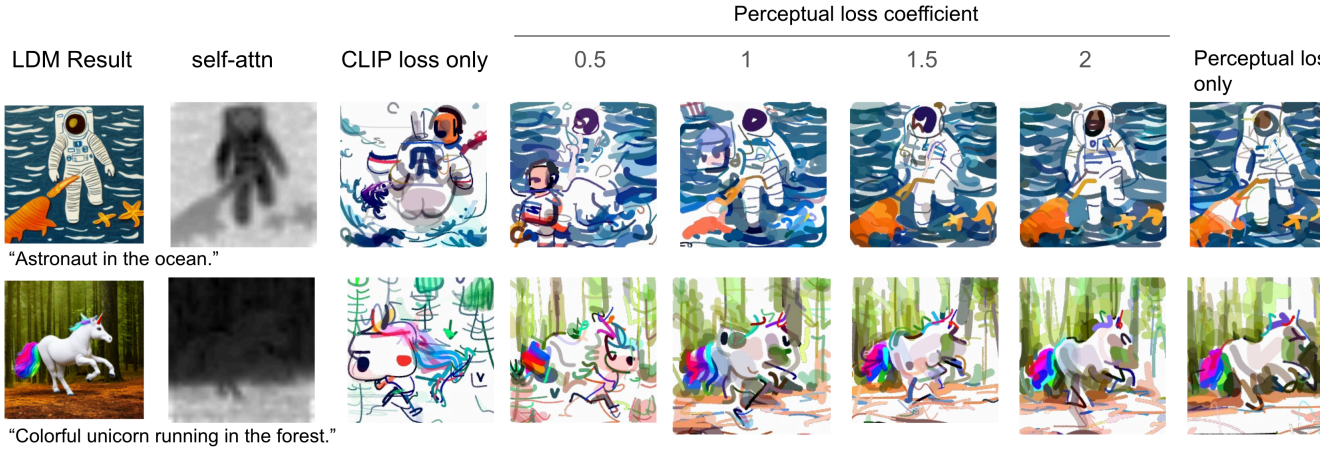

Figure 5. Ablation study showing the impact of different components of the proposed loss function (e.g., CLIP loss and perceptual loss) on the results. Starting from the left, we have the image generated by LDM from the text prompt, which guides the drawing, and its self-attention map used for placing initial strokes. Moving to the right, we gradually increase the weight of the perceptual loss from 0 (CLIP loss only) to exclusively using perceptual loss.

a metric used to evaluate the perceptual similarity between images. Unlike traditional pixel-wise metrics such as Mean Squared Error (MSE) or Peak Signal-to-Noise Ratio (PSNR), LPIPS aims to align more closely with human perception. Using LPIPS, we're able to guide the drawing synthesis with the sample image from the diffusion model. The final loss function is:

$$\mathcal{L}_{\text{total}} = -\frac{1}{n}\sum_{j=1}^{n} \text{CLIP}(\text{imageAug}_j(D_i), P) + \lambda \mathcal{L}_{\text{LPIPS}}\left(\text{Decoder}(S^{\phi}(z_t|y;t)), D_i\right) \quad (1)$$

The first term is the CLIP loss, where n is the number of augmentations, imageAug*j* is the *jth* image augmentation operation, *Di* is the rendered image from the generated curves at iteration *i*, and *P* is the input text prompt. We want to maximize the semantic similarity between generated drawing and input prompt, hence the negative sign for cosine distance. The second term is the perceptual loss using LPIPS, where the first argument is the decoded image sampled from the diffusion model, which is compared against *Di* on perceptual similarity. The goal is to maximize perceptual similarity.

4. Results

In this section, we evaluate the result of the proposed method using both qualitative and quantitative approaches. We chose CLIPDraw as our baseline for drawing quality because it is the most cited optimization-based synthesis method using CLIP in recent research. Subsequent papers like CLIPasso, StyleCLIPDraw, and CLIPascene have made minor adjustments to the problem formulation, such as limiting strokes to black and white or using an image instead of a free-form text prompt. However, they did not propose notable improvements on the optimization loop structure, making direct comparisons to those methods unnecessary.

4.1 Qualitative Comparison with Existing Synthesis-through-optimization Methods

In Figure 3, we demonstrate that our proposed method produces significantly better drawings than CLIPDraw in terms of aesthetics, capability to capture semantic meaning, and being more human-recognizable. This improvement is consistent across a wide range of text prompts, from concrete objects and abstract concepts to more complex scenes.

Furthermore, unlike CLIPascene, which requires separating the image into foreground and background for drawing synthesis, our method generates recognizable backgrounds and objects in a unified process.

We also show incremental improvements in the proposed methods regarding initialization strategy and loss function. The naive random initialization strategy used by CLIPDraw (bottom row of Figure 3) often results in suboptimal drawings that get stuck in local optima. For instance, in the first text prompt of Figure 3, CLIPDraw attempts to draw multiple incomplete roses spread across the canvas, resulting in a non-aesthetic outcome with messy color chunks. Similarly, in the rightmost column, while our proposed method successfully synthesizes a puppy in a sensible position, CLIPDraw is optimized for placing multiple golden color chunks on the canvas, creating unappealing outcomes. The second row of Figure 3 illustrates that simply replacing random initialization with Canny Edge initialization does result in cleaner drawings, but does not have significant improvement when the loss function is only composed of CLIP loss.

Regarding the loss function, the first row of Figure 3 shows that introducing perceptual loss guided by images generated from Latent Diffusion Models, rather than relying solely on CLIP, produces drawings that better capture the intended semantics compared to the baseline method. For example, despite Bézier curves intrinsically has lower fidelity than pixel-level images, the proposed method is able to accurately capture the shape of Half Dome (column 5) and a dragon (column 6) where the baseline method failed to.

4.2 Comparing initialization strategies

The training loss curves for different initialization strategies (Figure 6) reveal interesting insights into their effectiveness. Notably, the Canny edge initialization method, without the aid of perceptual loss, sometimes performs worse than random initialization. This is likely because the CLIP loss alone is insufficient to guide the optimization process out of local minima when starting from Canny edges, which may not align well with the semantic content of the text prompt.

In contrast, the random initialization and attention map initialization methods show different behaviors. While the random initialization can sometimes get stuck in suboptimal local minima, the attention map initialization generally converges slightly faster. This suggests that attention maps provide a better starting point by aligning initial strokes more closely with the important regions of the image.

However, convergence speed is not the sole metric of interest. The goal is the quality of the synthesized drawing, which has been discussed in the previous section. The attention map initialization not only converges faster but also leads to higher-quality outputs, as it better captures the semantic and structural details of the text prompt. This combined approach of optimizing initialization and integrating perceptual loss results in more aesthetically pleasing and semantically accurate drawings.

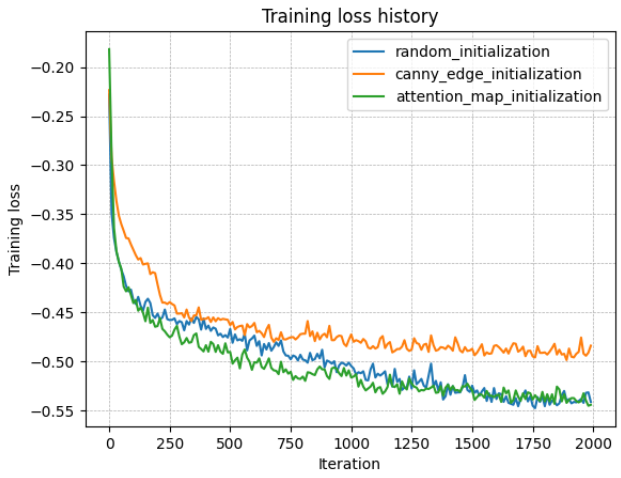

Figure 6. Training loss curves of three different initialization strategy of the input prompt "Ice cream in Van Gogh style".

4.3 Ablation Study on Loss Component Impact

Given that adding LDM-guided perceptual loss significantly improves the synthesized drawing, we conducted an ablation study to understand its impact compared to the other component of the loss function (i.e., CLIP cosine similarity loss). A few examples using different coefficients of the perceptual loss are shown in Figure 5.

Our study shows that while the image-text embedding in CLIP has solved a robust range of image-based recognition tasks, CLIP loss alone is far from adequate to optimize the control points of strokes to form aesthetic drawings. This is expected because a good drawing should not only capture semantic but also be aesthetical to human perception. In Figure 5, as we increase the weight for the perceptual loss, which means increasing the impact of the guided image sample from Stable Diffusion, the quality of the drawings increases. Interestingly, in the rightmost column of Figure 5, where we solely use perceptual loss and exclude CLIP loss, we found that the results are slightly worse than when both perceptual and CLIP losses are used. This indicates that semantic loss should still play an important role in text-to-drawing synthesis.

4.4 Qualitative Evaluation

As a proxy for qualitative evaluation, we followed CLIPDraw and CLIPasso to use a pretrained classifier network to evaluate the category-level recognizability of the drawing generated by the proposed method. While the proposed method can synthesize drawings with abstract concepts, we only test on input text prompts that at least include one to two concrete objects to avoid noises. Due to GPU resources and long inference time for each text prompt, only 50 input text prompts are tested.

Despite this being the most common qualitative evaluation for text-to-drawing synthesis, we do not find this metric indicative or practical. For example, CLIPDraw produced a messy drawing with the text prompt in Figure 5 column 2 ("A robot holding a balloon"), and the proposed method is superior when judged by human eyes. However, due to CLIPDraw's messy drawings of many ballons, the correct class score is slightly higher than that of the proposed method. This makes selecting meaningful text prompts to evaluate class scores improvement challenging, and easy to introduce bias. Within the text prompts we've tested that rule out such cases, we do see a slight improvement of 13% in the classification scores of the correct classes, compared to CLIPDraw as the baseline.

	CLIPDraw	Proposed method
Avg correct class(es) score(s) improvement	1	1.13x

Table 1. Percentage improvement of correct class scores using the proposed method compared to the CLIPDraw baseline, averaged over 50 selected text prompts.

4.5 Capturing Style

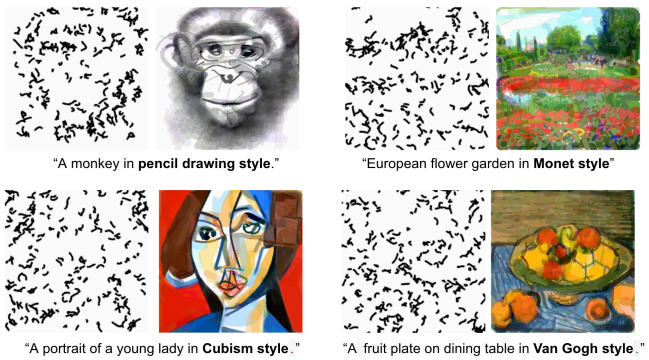

Figure 7. Examples of initialized strokes (left) and the final synthesized drawing (right) of four text prompts involving style. Num of strokes is 256. The perceptual loss coefficient is 1.8.

CLIP-based text-to-drawing methods such as StyleClipDraw added VGG-16 model into CLIPDraw's optimization structure to enable the drawings to fit the style of the input style image. However, we showcase in Figure 7 that the proposed optimization structure can capture style well through text-prompt without an additional VGG model, due to the incorporation of perceptual loss calculated from the sample image from the diffusion model.

This makes the proposed method have a wider use case than CLIP-only text-to-drawing methods such as CLIPDraw. The previous Figure 3 shows that CLIPDraw attempted to draw the face of Van Gogh instead of capturing the artist's style due to the limitation of its CLIP-only loss function. On the contrary, the proposed method can capture the nuances of style specified in the text prompt with perception loss. When the number of strokes is larger (e.g., 256), the synthesized drawings as shown in Figure 7 do have a comparable aesthetic to pixel images.

5. Discussion on limitations

While our proposed method significantly improves the quality of synthesized drawings compared to existing CLIP-based approaches, it has several limitations. One of the primary limitations is its difficulty in accurately rendering detailed subjects, such as human and animal faces. Despite the enhanced initialization strategies and loss functions, the method often struggles to capture the intricate features and nuances required for these types of drawings, resulting in less recognizable and less aesthetically pleasing outcomes.

Another limitation is the reliance on a pre-trained diffusion model for initialization and perceptual loss calculation. While this guidance helps improve overall drawing quality, it also introduces dependency on the quality and robustness of the diffusion model itself. If the diffusion model generates suboptimal images, the subsequent drawing synthesis may also be adversely affected.

6. Conclusions

In summary, we have presented an approach to text-to-drawing synthesis that addresses key limitations of existing CLIP-based methods through improved initialization strategies and the incorporation of perceptual loss. Our method demonstrates significant improvements in generating visually coherent and semantically accurate drawings. Future work will focus on refining these aspects to further enhance the applicability and robustness of text-to-drawing synthesis.

References

[1] High-Resolution Image Synthesis with Latent Diffusion Models, 2021
[2] Zero-Shot Text-to-Image Generation, 2021
[3] CLIPDraw: Exploring Text-to-Drawing Synthesis through Language-Image Encoders, 2021
[4] CLIPasso: Semantically-Aware Object Sketching, 2022
[5] Learning Transferable Visual Models From Natural Language Supervision, 2021
[6] The Unreasonable Effectiveness of Deep Features as a Perceptual Metric, 2018
[7] CLIPascene: Scene Sketching with Different Types and Levels of Abstraction, 2023
[8] StyleCLIPDraw: Coupling Content and Style in Text-to-Drawing Translation, 2022
[9] DiffSketcher: Text Guided Vector Sketch Synthesis through Latent Diffusion Models, 2023.
[10] Canny Edge Detection wiki