You Text, We Sketch: Text Guided Diffusion For Vectorized Sketch Generation

Yash Shah, Samy Cherfaoui**

Department of Computer Science, Stanford University

*Equal contribution

Introduction

Vectorized sketches that make use of strokes and point-slopes are a more natural way of thinking about how humans draw (generate) images or sketches.

Advantages

- ✓ They can be scaled up or down without affecting image quality, as opposed to raster images.

- ✓ Lightweight in file size

Applications

- Icon and logo generation
- Complex design generation that fits on a business card and a billboard
- Sketch completion and “healing”
- Domain adaptation to OOV word prompts

Text Conditioning

A more natural way of communicating the kind of sketch to be generated using complexities of human language, following popularization of DALL-E 2, Glide, Imagen, and more!

Problem Setup

Inputs:

$\forall i \in [\mathcal{N}], \boldsymbol{s}_i = \{\forall j \in [\mathcal{K}], (\Delta x^{(j)}, \Delta y^{(j)}, g^{(j)})\}$

$\forall i \in [\mathcal{N}], c_i \leftarrow$ Text prompt for $\boldsymbol{s}_i$

Output:

Learn to generate a sketch from noise given a text prompt

Background

We extend SketchKnitter [1] by conditioning on text prompts. To do so, we incorporate ideas from Glide [2].

SketchKnitter

- Generates vectorized sketches unconditionally from noise using DDIMs
- Learns to predict binary pen state for each stroke point
- Conditions on part of the sketch for completion and “healing” tasks

Glide

- Guided diffusion for text-conditioned raster image synthesis
- Uses a text encoder to condition on natural language
- Trains a 3.5B parameter diffusion model
- Compares CLIP guidance against Classifier-free guidance

Dataset

Google’s *Quick, Draw!* dataset [4]

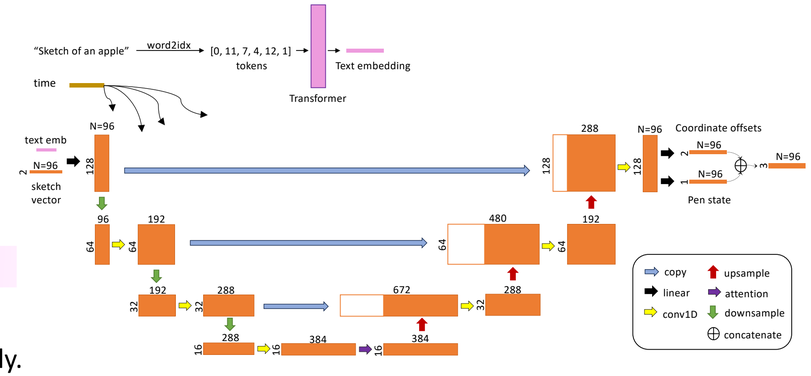
Classes

Text prompts

<start>
{ this | here | image | sketch }
{ is | of }
{ a | an | the }
{ apple | umbrella | moon | shoe | lion | fish }
<end>

Methods

- U-Net conditions on text embeddings
- Embeddings also projected to dimensionality of attention layers and concatenated to attention context at each layer



Classifier-free guidance [3]

During training, randomly discard conditioning to train unconditionally.

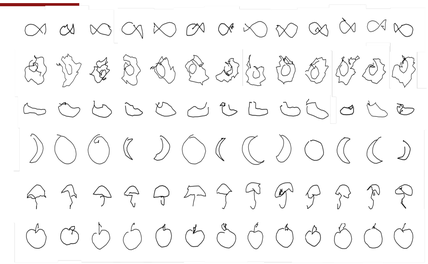
During sampling, $\hat{\epsilon}_\theta(\boldsymbol{x}_t|c) = \epsilon_\theta(\boldsymbol{x}_t|\emptyset) + w \cdot (\epsilon_\theta(\boldsymbol{x}_t|c) - \epsilon_\theta(\boldsymbol{x}_t|\emptyset))$
where w is the guidance strength

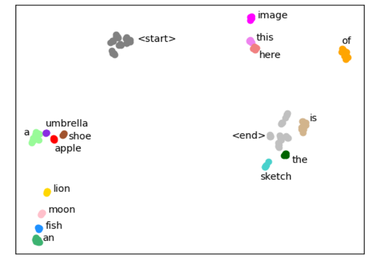
Results

- Evaluated our samples via Fréchet inception distance (FID), geometry score (GS), and Kynkäänniemi precision and recall.
- Conducted experiments by generating 512 samples for every possible grouping of (class, w) where w ∈ {1, 2, 3, 4, 5}, representing the sampling guidance strength.

Table 1: Model evaluation on FID, GS, Precision, and Recall (n=512 samples) using the best w for a given class. This is w=5 for all classes except `moon` which achieves maximum performance at w=1.

Model	FID ↓	GS ↓	Prec ↑	Rec ↑
SketchKnitter (Unconditional)	6.9	3.4	0.52	**0.88**
Ours (on text prompts for class `apple`)	6.4	3.1	**0.60**	0.75
Ours (on text prompts for class `umbrella`)	7.0	4.5	0.52	0.54
Ours (on text prompts for class `moon`)	6.6	**3.0**	0.58	0.71
Ours (on text prompts for class `shoe`)	**6.0**	3.2	0.57	0.77
Ours (on text prompts for class `lion`)	18.4	5.6	0.13	0.52
Ours (on text prompts for class `fish`)	6.3	3.1	0.59	0.74
Average across classes (w=5)	8.5	3.8	0.50	0.66
Average across classes (w/o `lion`) (w=5)	6.5	3.4	0.57	0.68





References

[1] Wang, et al. SketchKnitter: Vectorized sketch generation with diffusion models. In *ICLR*, 2023.
[2] Nichol, et al. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *ICML*, 2022.
[3] Ho and Salimans. Classifier-free diffusion guidance. In *NeurIPS*, 2021.
[4] https://github.com/googlecreativelab/quickdraw-dataset