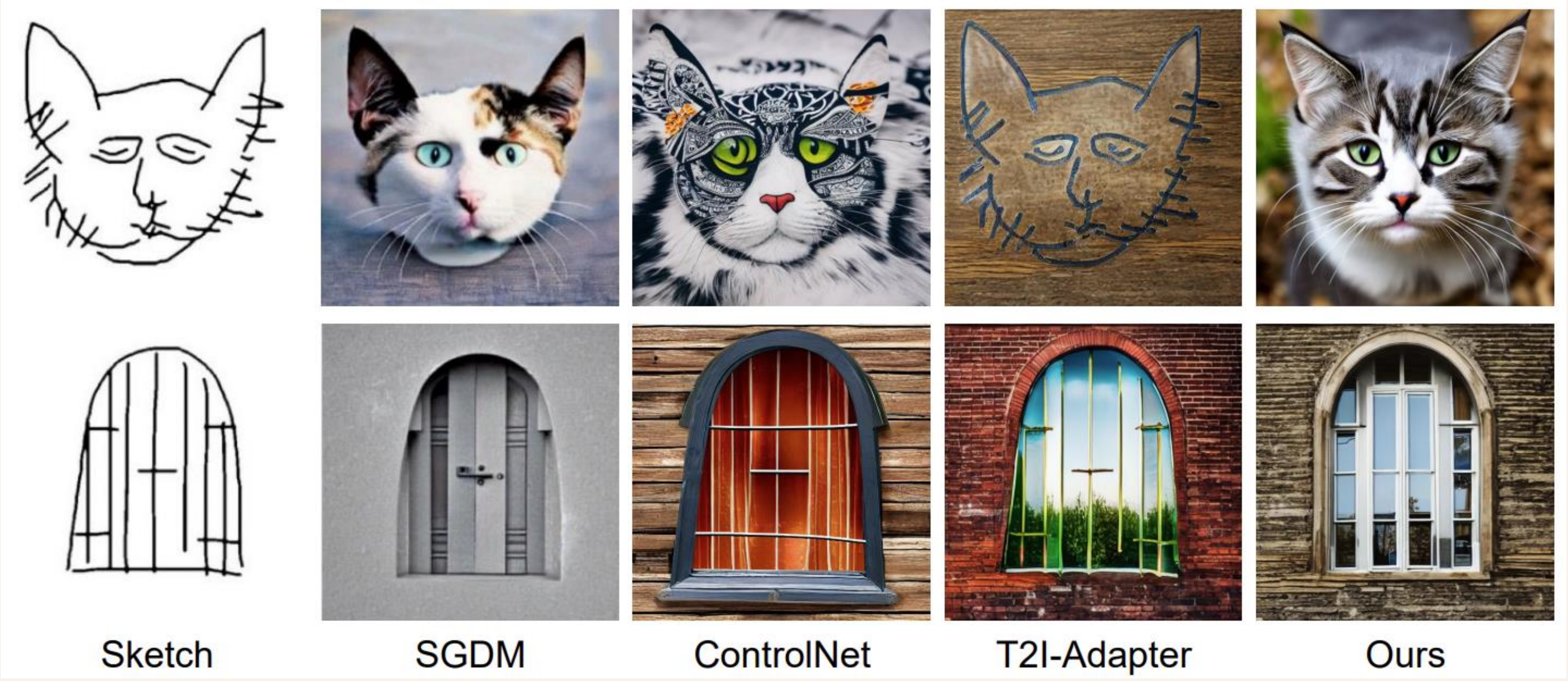
It's All About Your Sketch: Democratising Sketch Control in Diffusion Models

Subhadeep Koley[1,2], Ayan Kumar Bhunia[1], Deeptanshu Sekhri[1], Aneeshan Sain[1], Pinaki Nath Chowdhury[1], Tao Xiang[1,2], Yi-Zhe Song[1,2]
[1]*SketchX, CVSSP, University of Surrey*
[2]*iFlyTek-Surrey Joint Research Centre on Artificial Intelligence*

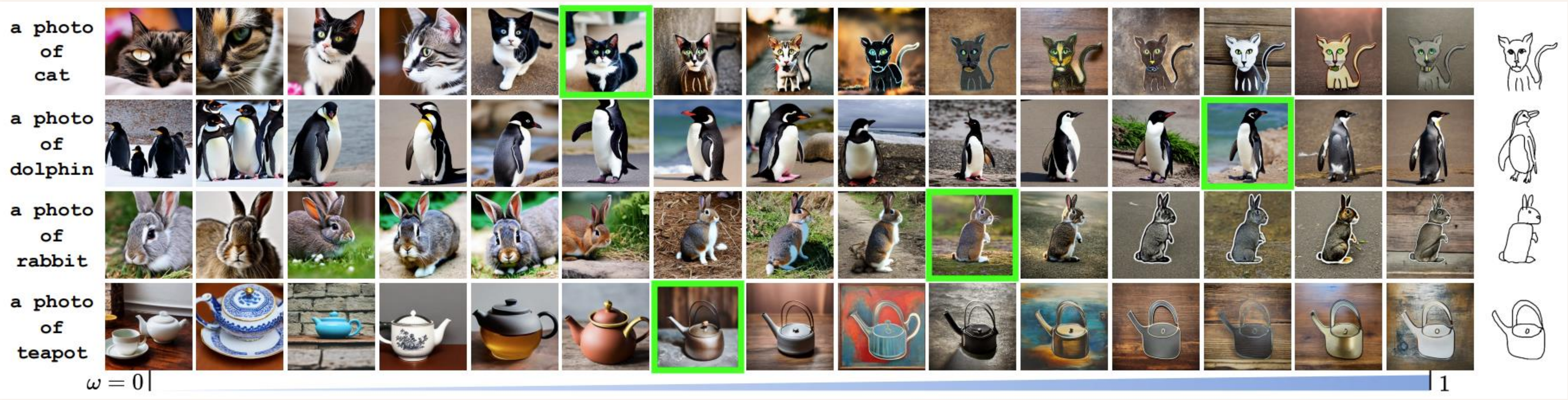
CVPR JUNE 17-21, 2024 SEATTLE, WA

Got 2 Minutes? Start here ⬇



Summary

- **Democratises sketch control**, enabling real amateur sketches to generate accurate images.
- Identifies the root-cause behind **deformed and non-photorealistic** outputs of existing diffusion-based Sketch-to-Image frameworks.



What's wrong with Sketch-to-Image DM?

- Sketches depict **significant shape-deformity** and hold less contextual information than other pixel-perfect conditioning signals (*e.g.*, masks).
- Lack of suitable prompts negatively impacts result. Ensuring a **balance between sketch and text-conditioning** requires manual intervention.
- Exiting methods (*e.g.*, SGDM) employs **spatial sketch-conditioning**.

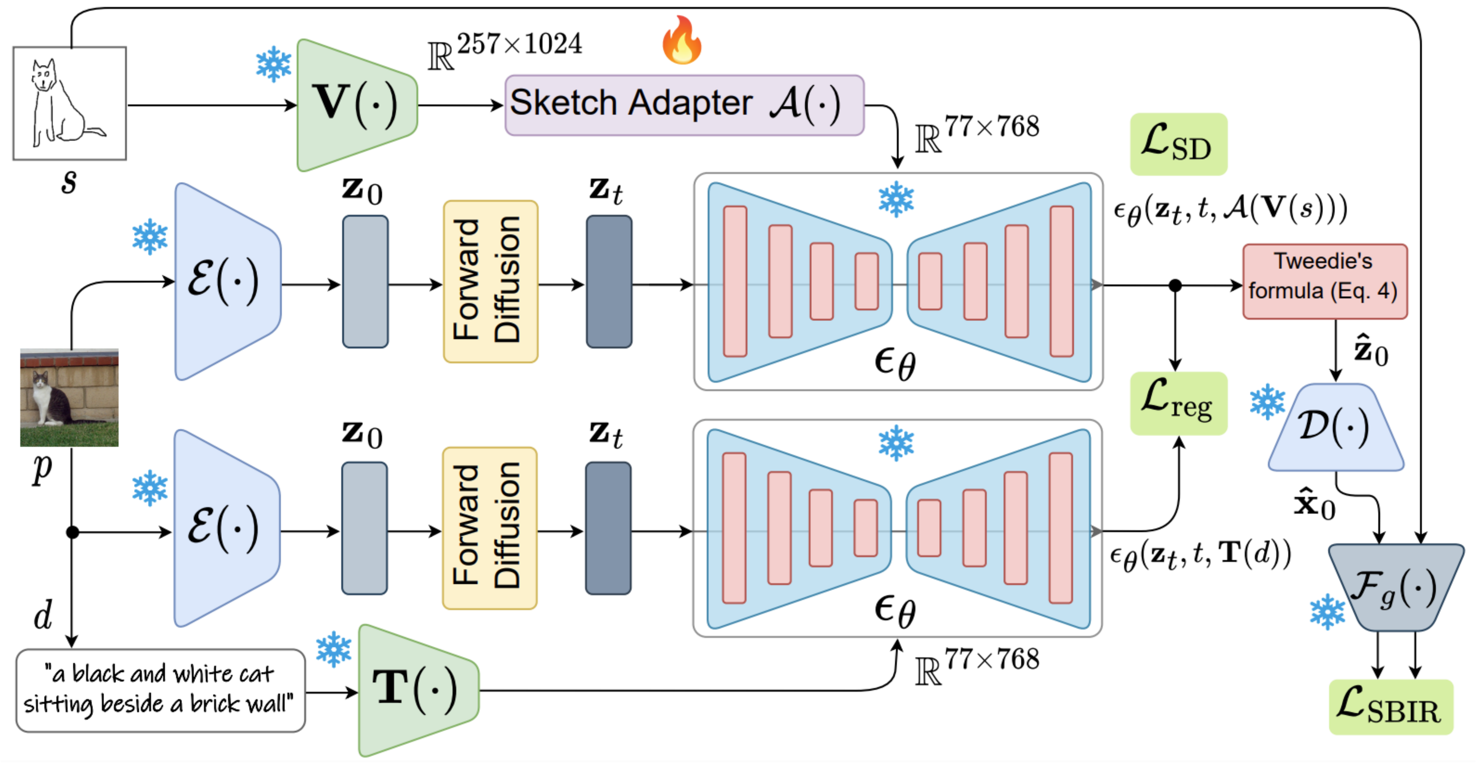
Solutions

- **Eliminate spatial sketch-conditioning** by converting the input sketch into an equivalent fine-grained textual embedding, thereby preserving users' semantic-intent without pixel-level spatial alignment.
- **Fine-grained discriminative loss** for maintaining the fine-grained sketch-photo correspondence.
- Introduces **sketch-abstraction-aware t-sampling**. For highly abstract sketch, a higher probability is assigned to larger t and vice-versa.

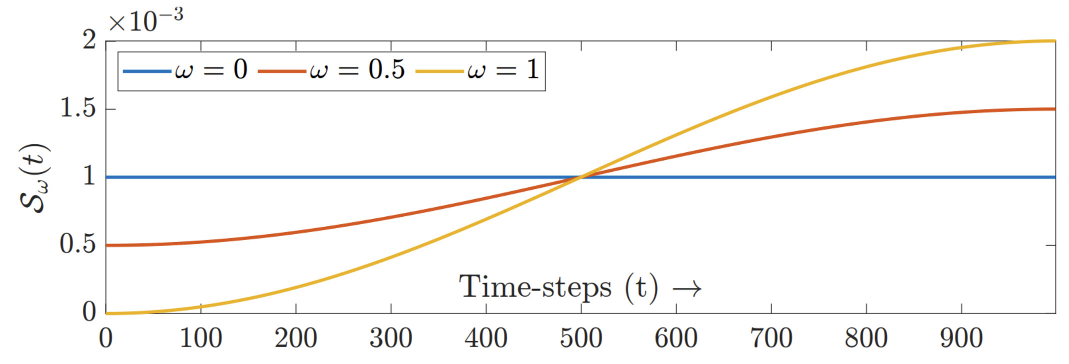
Proposed Model

- **Salient Components**
  1. Fine-grained discriminative guidance via pre-trained FG-SBIR model.
  2. Super-concept preservation loss via synthetically generated textual prompts.
  3. Adaptive t-sampling based on input sketch-abstraction.

- We encode sketches as sequence of feature vectors as an **equivalent fine-grained textual embedding**.
- To ensure a **fine-grained matching** between sparse freehand sketches and pixel-perfect photos, we use a pre-trained FG-SBIR model $\mathcal{F}(\cdot)$.



- For learning $\mathcal{A}(\cdot)$, we use a discriminative SBIR loss that calculates cosine similarity $\delta(\cdot,\cdot)$ between input sketch and output photo features from $\mathcal{F}(\cdot)$.
- We posit that textual captions being less fine-grained than a sketch, acts as a **super-concept** of the corresponding sketch.
- We use a pre-trained SoTA captioner to synthetically generate caption d for every ground truth photo p. Then, at each t, the noise predicted through text-conditioning ($T(d)$) acts as a reference to calculate a **regularisation loss** to learn $\mathcal{A}(\cdot)$.
- **High and low-level semantic structures** of the output image tend to manifest in different stages of the denoising process.
- We thus adjust the **time-step sampling procedure** based on the input sketch's abstraction level.



- To assess sketch-abstraction, we design a **CLIP-based sketch classifier**, that provides a score where scores from $0 \rightarrow 1$ denotes more to less abstract sketches.

Experiments & Results