GlueGen: Plug and Play Multi-modal Encoders for X-to-image Generation

Can Qin\textsuperscript{1,2}, Ning Yu\textsuperscript{1}, Chen Xing\textsuperscript{1}, Shu Zhang\textsuperscript{1}, Zeyuan Chen\textsuperscript{1}, Stefano Ermon\textsuperscript{3}, Yun Fu\textsuperscript{2}, Caiming Xiong\textsuperscript{1}, Ran Xu\textsuperscript{1}

\textsuperscript{1}Salesforce Research \textsuperscript{2}Northwestern University \textsuperscript{3}Stanford University

Background

- Text-to-image (T2I) synthesis, generating photorealistic images from text prompts, has witnessed a tremendous surge in capabilities recently.

"an astronaut riding a horse"

Motivation

- How to enhance the current text encoder of T2I model with more powerful language models?
- How to plug and play multi-modal encoders to enable X-to-image generation without time-consuming retraining?

Proposed Model

Framework of GlueGen

- GlueGen can plug in off-the-shelf pre-trained components, including:
  1. More powerful language model: T5-3B
  2. Multi-lingual Language Models: XLM-Roberta
  3. Audio Encoders: AudioCLIP

Details of GlueNet

- To achieve a desired GlueNet, we propose:
  1. Encoder-decoder structure as shown by (b)
  2. Alignment and reconstruction as illustrated by (a)
  3. Dense regression module with TokenMLP and SeqMLP as (c)
  4. Without the need of re-training Unet