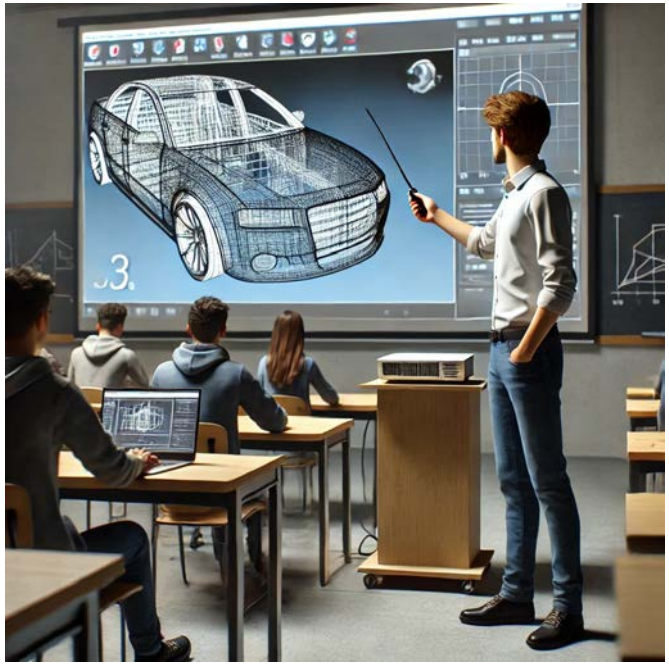
Text-to-3D Shape generation

TDT03: Advanced Topics in Visual Computing

H. Lee M. Savva A. X. Chang

Generative models are making incredible images



What about 3D?

The paper divides based on training data

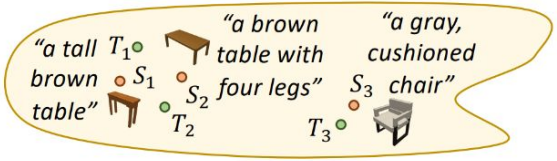
3D with Paired Text (**3DPT**)

"a tall brown table"

"a brown table with four legs"

"a gray, cushioned chair"

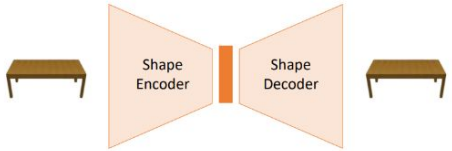
Aligned text-3D embedding



3D with Unpaired Text (**3DUT**)

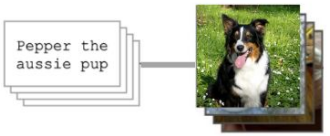
3D shape corpus

Learned 3D shape priors



No 3D data (**No3D**)

Large text-image corpus → Large vision-language model



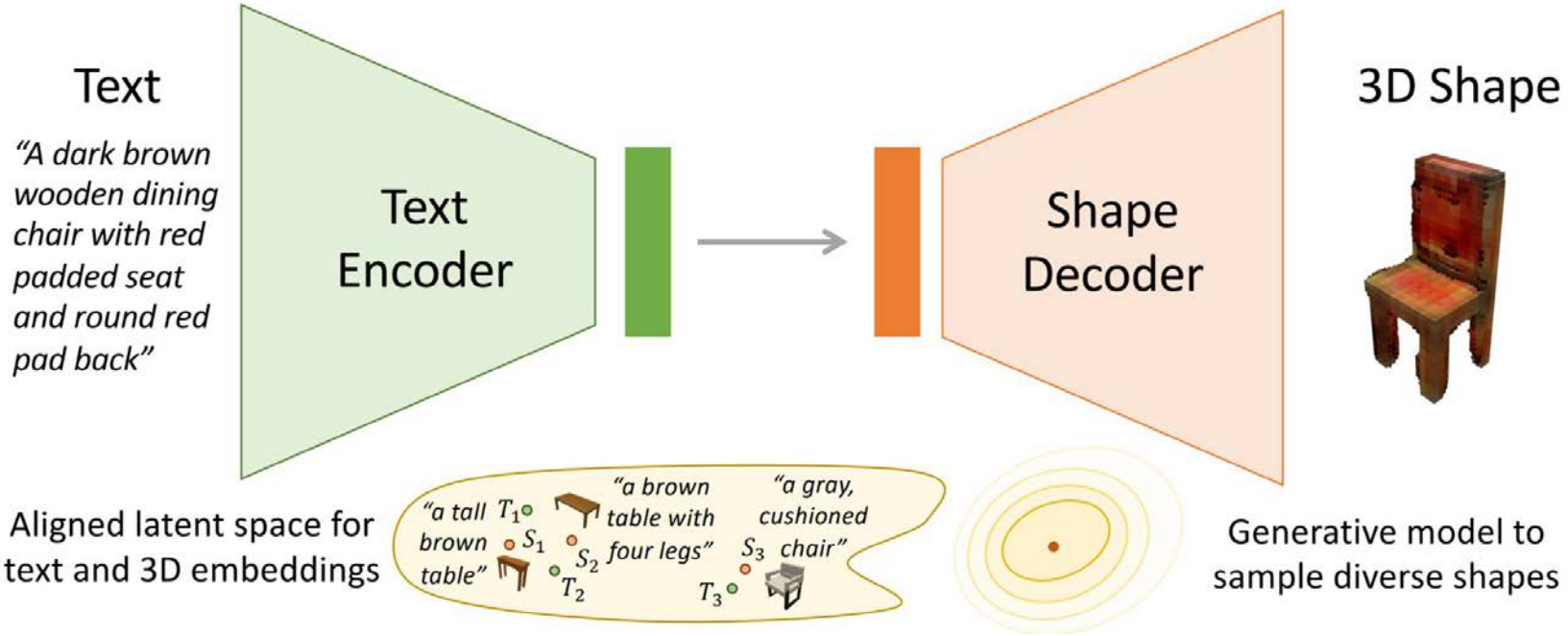
CLIP | DALLE-2 | ALIGN | IMAGEN

Prompt-based optimization of **differentiable** 3D representation

NeRF

DMTet

General structure

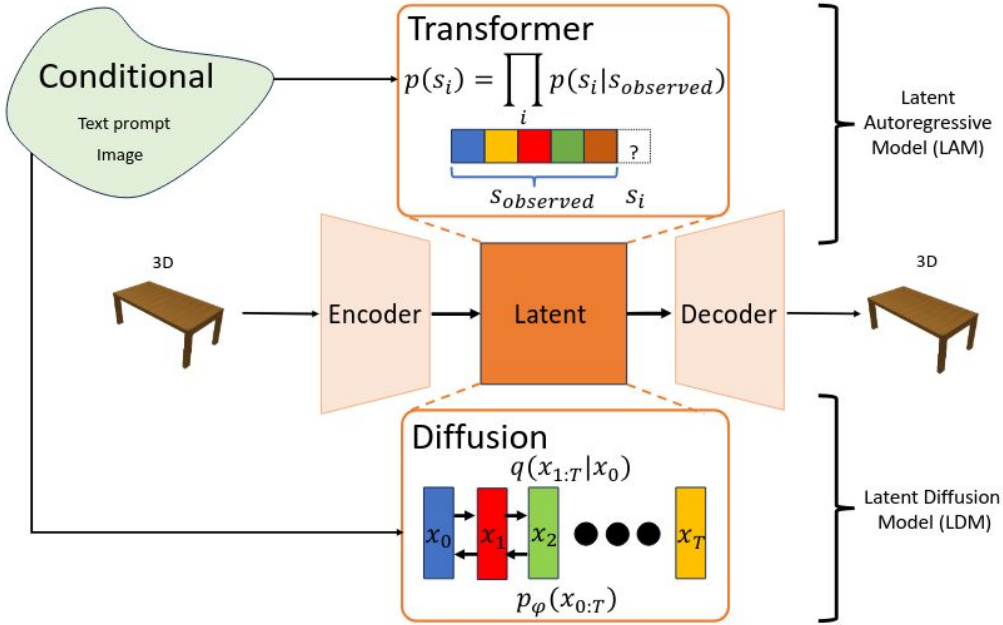


3D Paired Text

- Good results
- Can’t generate outside of dataset
- Few datasets

Two stages

1. Train VAE to learn latent space
2. Train generator to predict latent vector from text

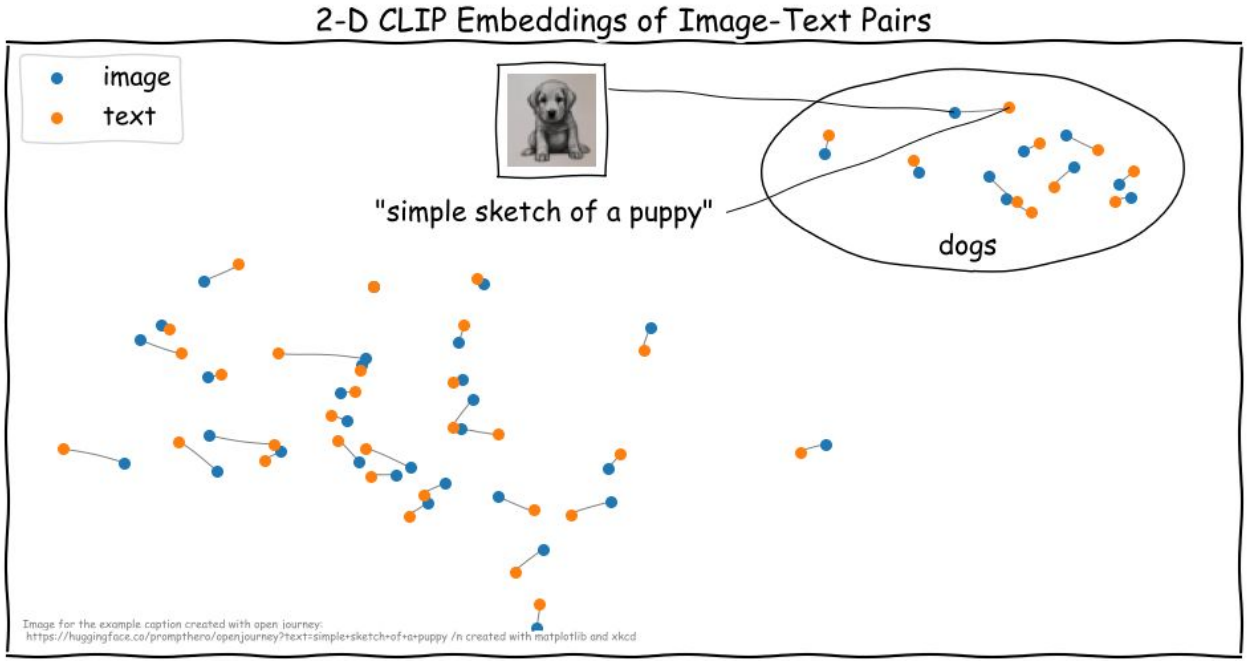


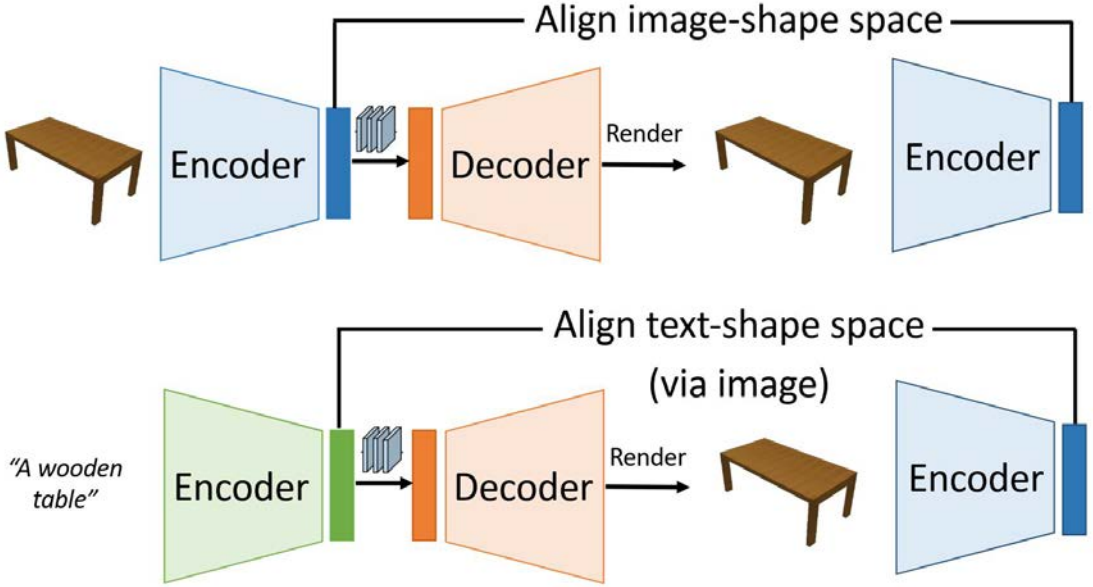
Unpaired 3D Data

-

Secret Weapon: image-and-text embeddings

- Other models like CLIP can be used




Align image-shape space
Encoder
Decoder
Render
Encoder
Align text-shape space
(via image)
"A wooden table"
Encoder
Decoder
Render
Encoder

No3D

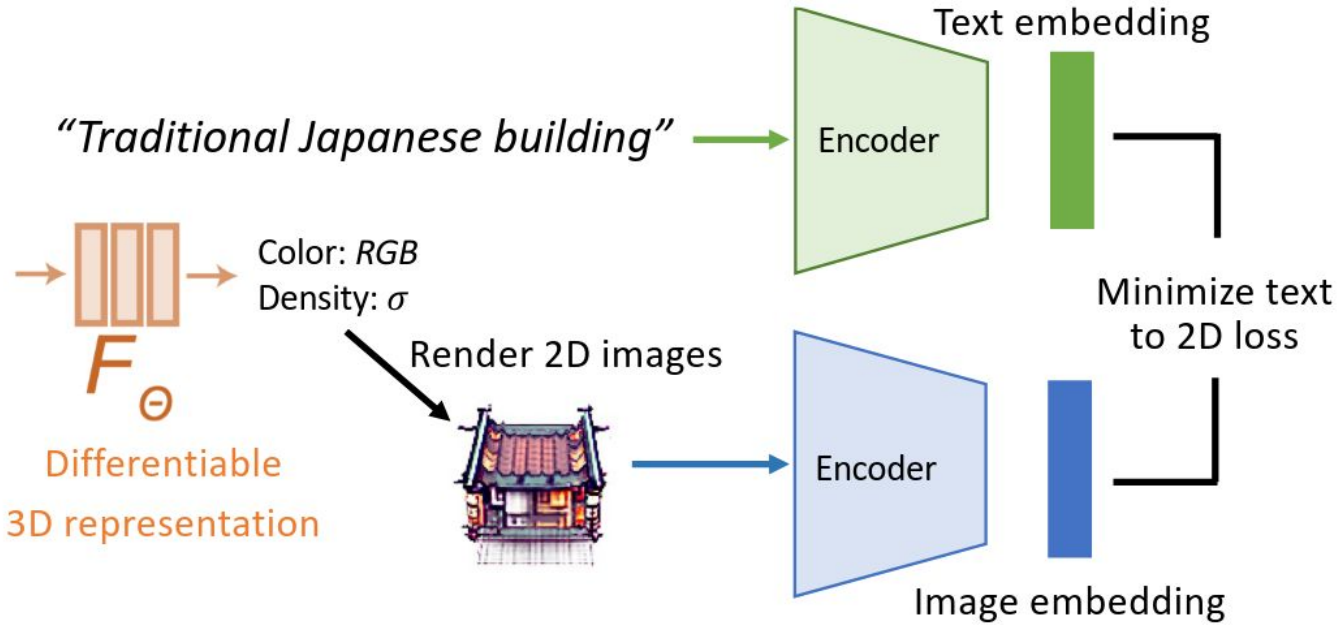
Can you teach a model to generate 3D output without any 3D training data?

2 solutions

Solution 1: Back to CLIP

Solution 2: Using diffusion models

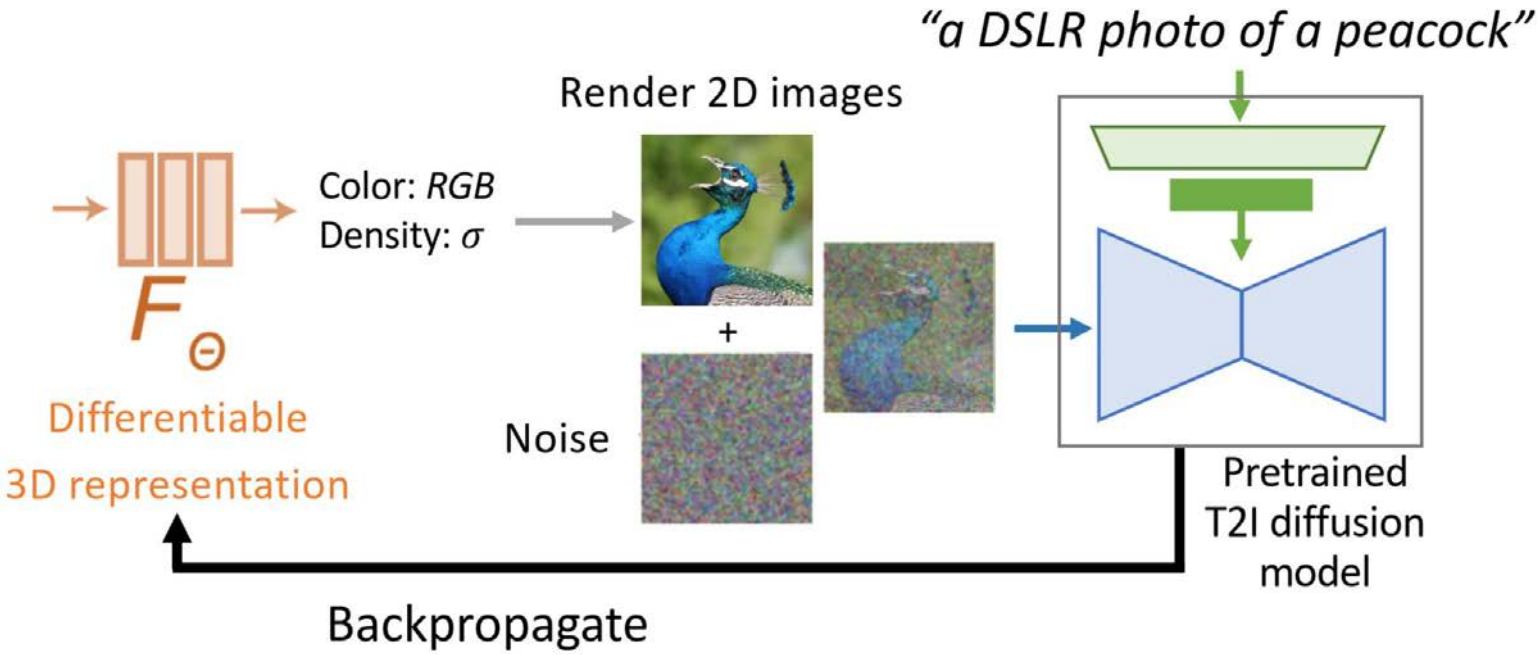
Solution 1: Back to CLIP



Works pretty good on textures, but the models are bad

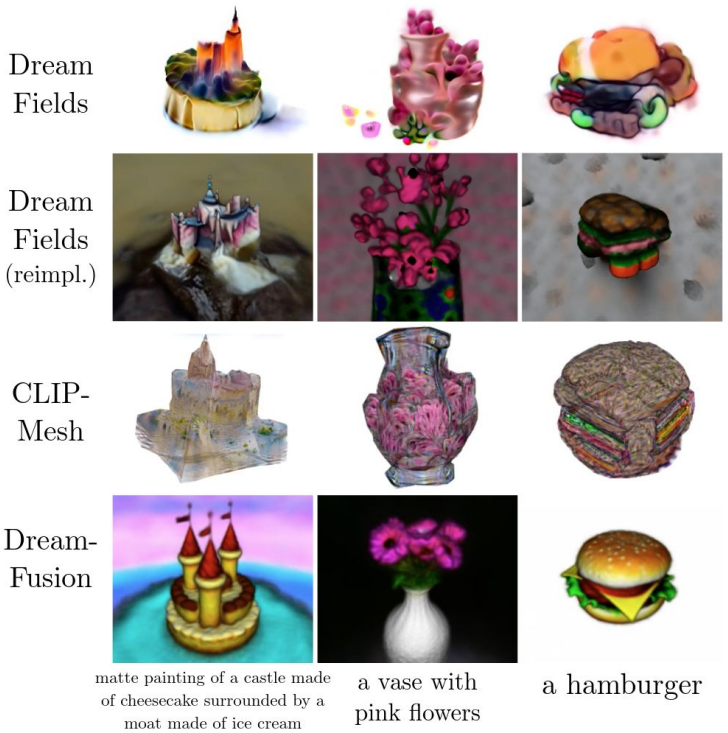
Dream Fields

Solution 2: Using diffusion models



Score Distillation Sampling (SDS) loss

CLIP < Diffusion



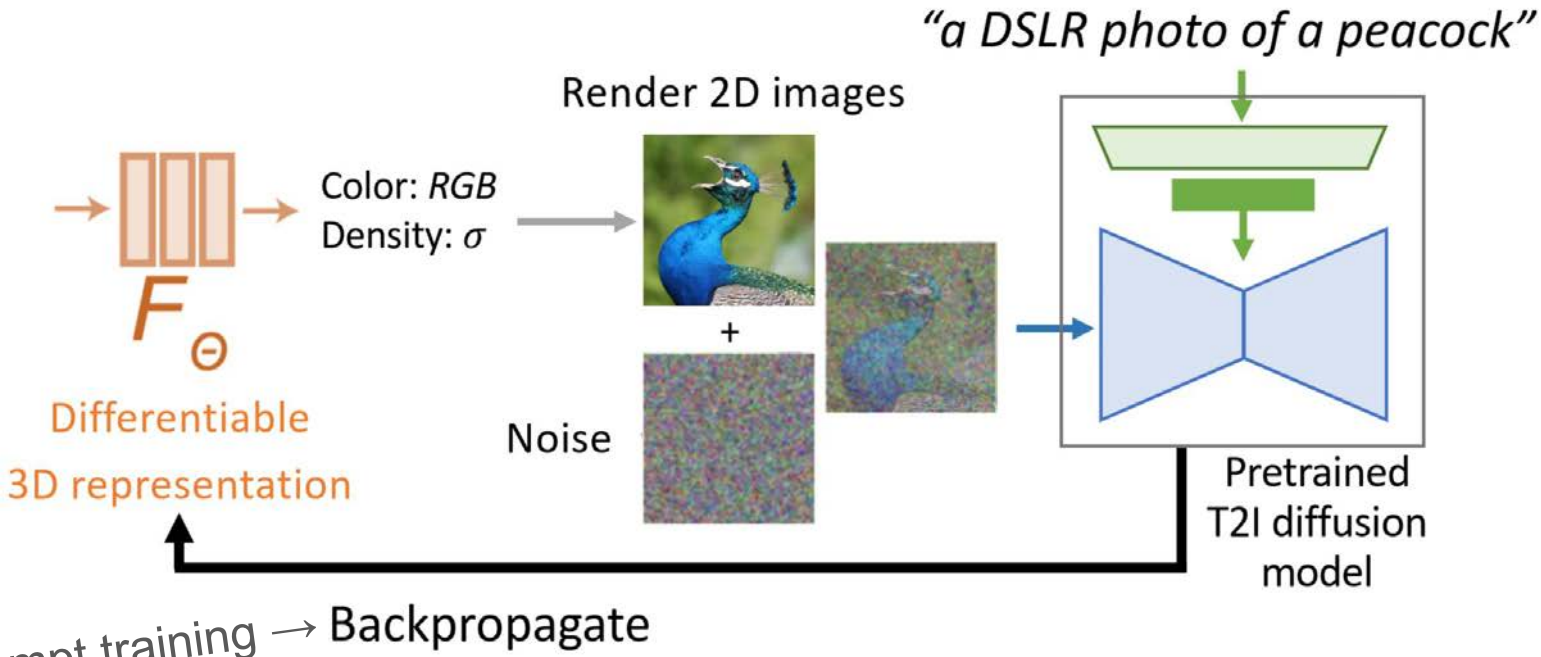
CLIP

CLIP

CLIP

Diffusion

Problems with using diffusion models



Problems with using diffusion models

Smooth shapes

Dream-Fusion

matte painting of a castle made of cheesecake surrounded by a moat made of ice cream

a vase with pink flowers

a hamburger

Problems with using diffusion models

The Janus Problem

Problems with using diffusion models

How to mitigate

- PerpNeg
  - Negative prompts
- MVDream
  - Fine tune guidance models to be 3D aware

Other interesting point the paper briefly mentions

- Image Conditioning - Image-to-Image given a new camera angle
- Image-to-3D - Generating 3D models from images
- Scene Generation - Generating scenes from either a set of pre-made models or from scratch

Set-the-scene