Learn-to-Steer:

Data-Driven Loss Functions for Inference-Time Optimization in Text-to-Image Generation

3D animation of a bearded man wearing glasses holding a banana
(to the left of | above | to the right of | below) his face in the kitchen

Sapir Esther Yiflach[1], Yuval Atzmon[2], Gal Chechik[1,2]
[1]Bar-Ilan University, [2]NVIDIA

Problem

Text-to-Image models often fail to portray the correct spatial arrangement of objects.

They struggle to generate even a single spatial relationship between two subjects accurately.

A dog to the right of a teddy bear

FLUX-dev

A dining table above a suitcase

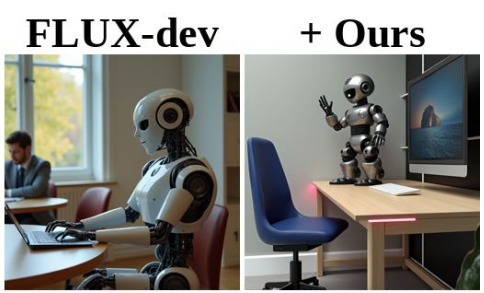
SD 2.1

Problem

It gets even worse when prompts include multiple relations.

A frog above a sneakers to the right of a umbrella

A robot above a table to the right of a chair



Current Approaches

1. Model Fine-Tuning → **Catastrophic Forgetting**
2. Test-time Optimization → **Rely on handcrafted heuristics**

Our Idea

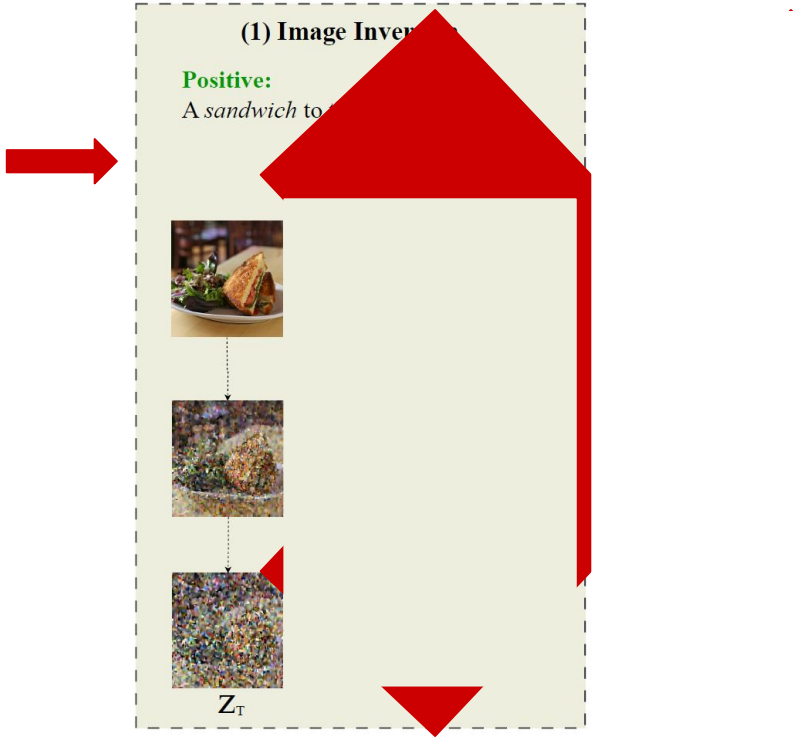
Instead of manually designing a loss function, we learn it from data.

<u>Step 1:</u> Learn spatial patterns from the cross-attention maps of images with correct relations (by training a relation classifier).

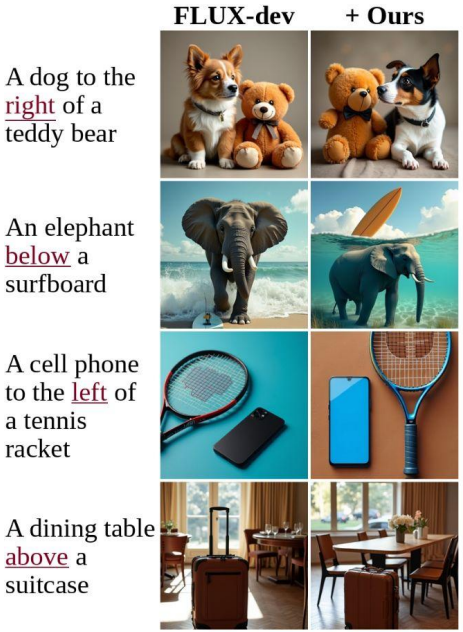
<u>Step 2:</u> Steer the generation process during inference using this learned classifier as the loss function.

Inference-time Optimization

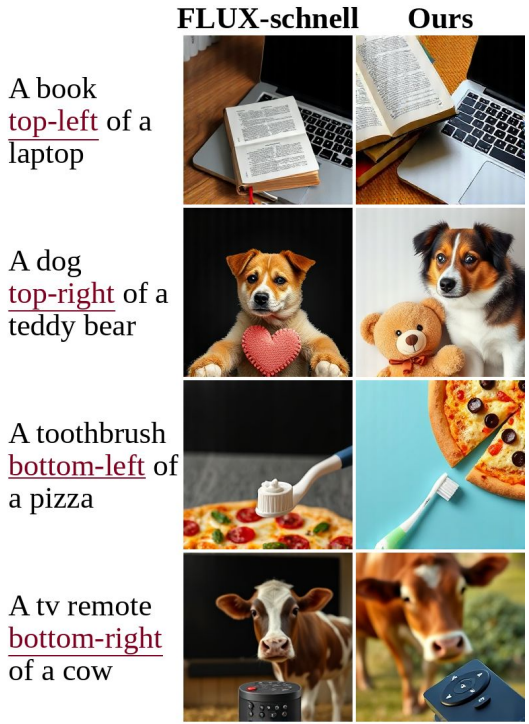
Training a Relation Classifier



Learn-to-Steer can be added to various models:



Multiple Relations + Diagonal Relations



Thank You!