

# 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<sup>1</sup>, Yuval Atzmon<sup>2</sup>, Gal Chechik<sup>1,2</sup>

<sup>1</sup>Bar-Ilan University, <sup>2</sup>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.

**FLUX-dev**

A dog to the right of a teddy bear



**SD 2.1**

A dining table above a suitcase



# Problem

It gets even worse when prompts include multiple relations.

A frog above a sneakers to the right of a umbrella

**FLUX-dev**



**+ Ours**

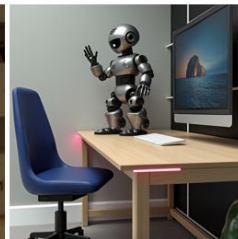


A robot above a table to the right of a chair

**FLUX-dev**



**+ Ours**



# 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.

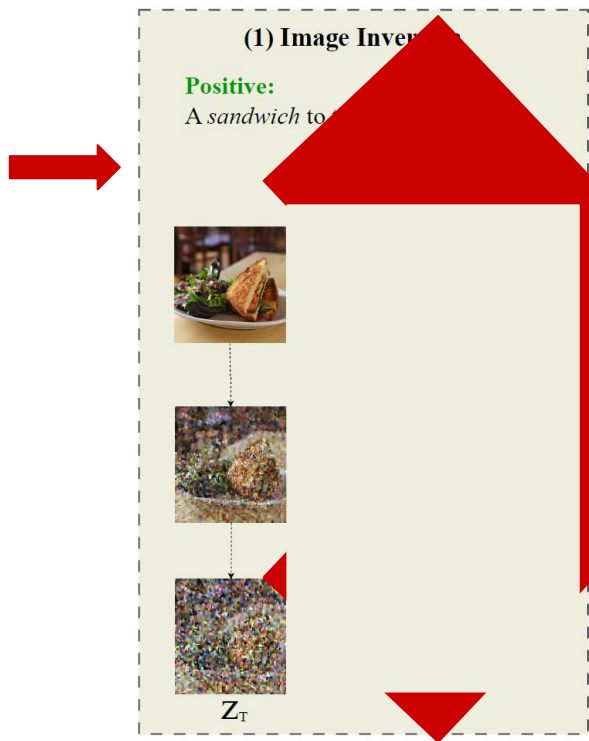
**Step 1: Learn spatial patterns from the cross-attention maps of images with correct relations (by training a relation classifier).**

**Step 2: Steer the generation process during inference using this learned classifier as the loss function.**

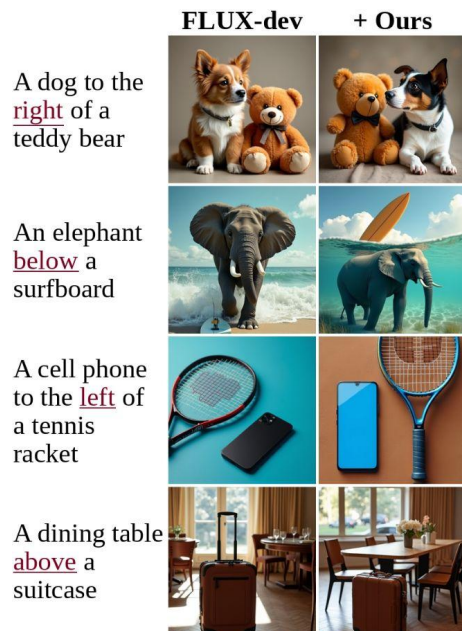
# Inference-time Optimization



# Training a Relation Classifier



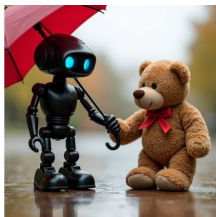
# Learn-to-Steer can be added to various models:



# Multiple Relations + Diagonal Relations



a frog **above** a sneakers; and the same sneakers **below** a teapot



a robot to the **left** of a teddy bear; and the same teddy to the **right** of an umbrella



a teapot **above** an umbrella; and the same umbrella to the **right** of a furby



a table to the **left** of a teddy bear; and the same teddy bear to the **right** of a mug; and a duck **above** a chair



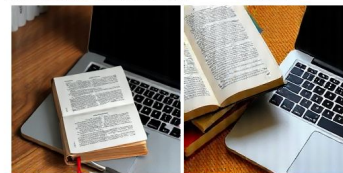
a table **below** a teapot; and the same teapot to the **left** of a backpack; and the same backpack to the **right** of a vase



a duck to the **right** of a sneakers; and the same sneakers **below** a vase; and the same vase to the **left** of a teddy bear

FLUX-schnell      Ours

A book **top-left** of a laptop



A dog **top-right** of a teddy bear



A toothbrush **bottom-left** of a pizza



A tv remote **bottom-right** of a cow



Thank You!