# T-SAM: Text Self-Attention Maps

For Semantic Alignment in Diffusion Models

# What is Text-to-Image Generation?

- Text → Image using diffusion models.
- Used in art, design, prototyping, automation.
- But they still struggle with semantic alignment.

> "Text-to-image diffusion models... can generate beautiful images... But even today, they often misunderstand what we are trying to describe."

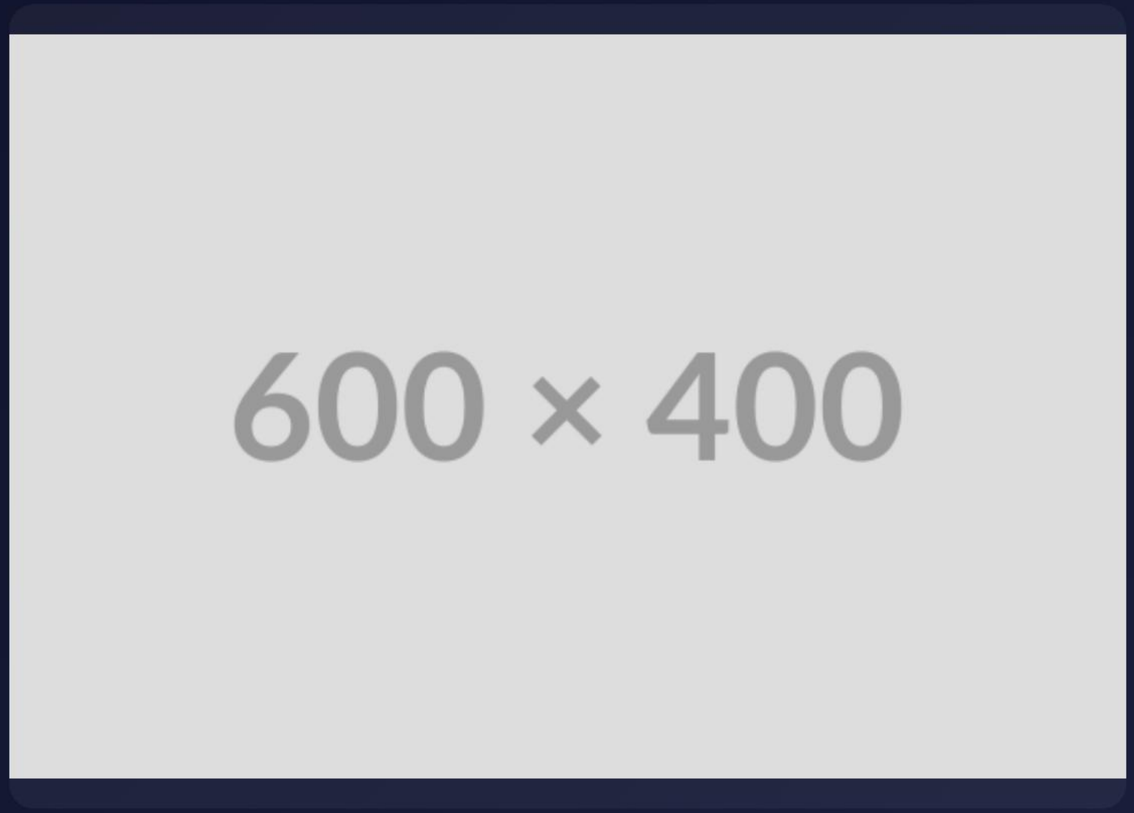
# Motivation: Typical Failures

**Prompt: “A black car and a white clock”**

Common Errors:

- Missing objects
- Wrong color binding
- Object–attribute confusion
- Blending of multiple objects

> “...the model may draw only one object, or swap the colors. This violation of meaning is called semantic misalignment.”



# Background: Cross-Attention

- Cross-attention links image pixels ↔ text tokens.
- Controls what part of the image listens to which word.
- Crucial for placing objects and attributes correctly.

> “If cross-attention for ‘black’ overlaps with ‘clock’, the clock might incorrectly become black.”

# Background: Self-Attention (Text Encoder)

- Self-attention captures word-to-word relationships.
- E.g., “black → car”, “white → clock”.
- This structure exists in the text encoder.

> “The CLIP text encoder internally knows grammatical relations... But this information never reaches the image generator.”

# Prior Work: Attend-and-Excite

## What it does

Forces the model to attend to all subjects during generation.

This solves "catastrophic neglect" (missing objects).

## The Limitation

Good for “object presence,” but **not fine attribute control.**

*“...it does not solve attribute binding — it does not know which color belongs to which object.”*

# Prior Work: Linguistic Binding (SynGen)

- Uses an external syntax parser (like SpaCy).
- Encourages correct attribute–entity pairs.
- Separates incorrect pairs.
- Good, but depends heavily on parser quality.

> "But if the parser fails or prompt is unusual, the method breaks. Also, [it] requires extra computation..."

# Prior Work: CONFORM / Attention Regulation

- **Concept:** Regulates attention using contrastive constraints.
- **Limitation 1:** Needs hand-designed token grouping.
- **Limitation 2:** Sometimes creates hard, unnatural boundaries in the generated image.
- **Quote:** "...it still needs hand-designed token groups and can make images look unnatural."

# This Paper: Key Observation 1

"

"This paper shows that cross-attention blindly follows text embeddings. If embeddings for 'black' and 'white' look similar, then their attention maps overlap, causing color confusion."

"

Observation: Words with similar embeddings → similar cross-attention maps.
Problem: Embeddings do not encode syntax properly.

# This Paper: Key Observation 2

> “The text encoder’s self-attention perfectly captures grammar. But due to the attention-sink effect, the final token embeddings lose these relationships, so the UNet never sees them.”

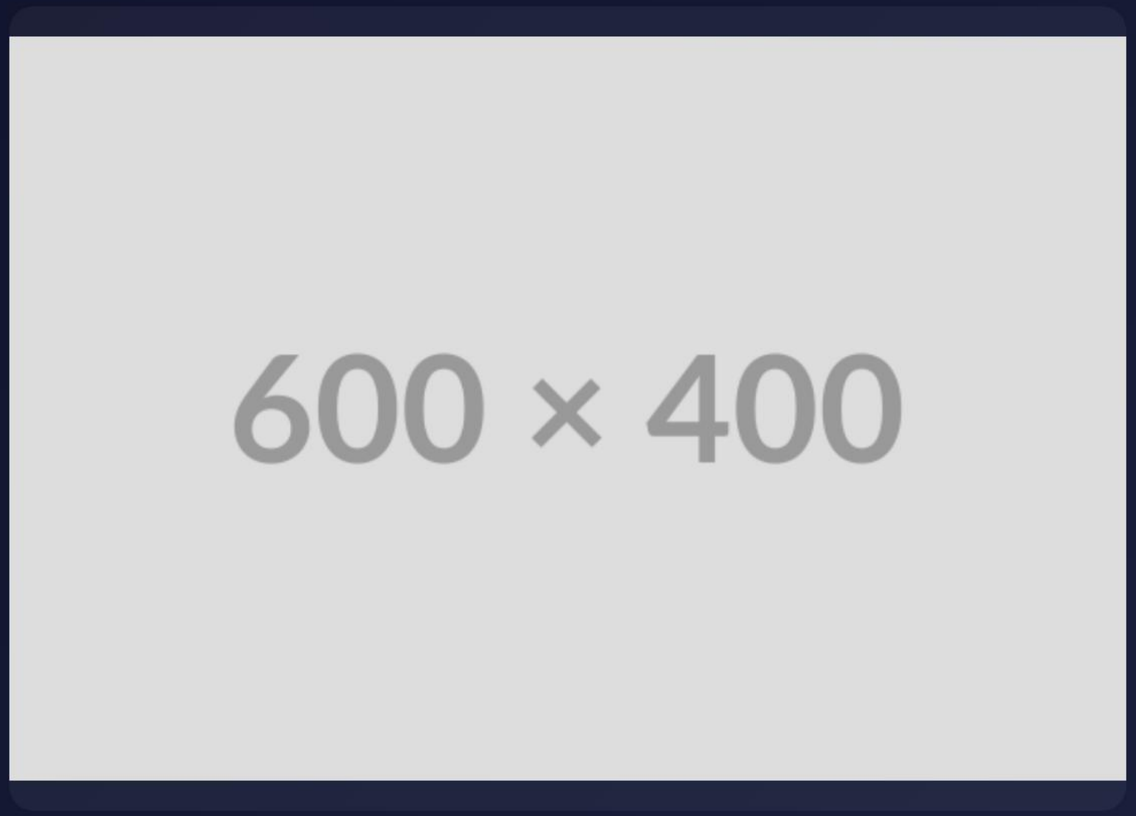
Observation: Self-attention **does** encode syntax correctly.
Problem: Embedding computation destroys this due to attention sink at .

# This Paper: T-SAM (Text Self-Attention Maps)

The idea is extremely elegant:

- Extract text self-attention matrix **T** (the correct syntax).
- Compute cross-attention similarity matrix **S** (image meaning).
- **Aim: Make S match T.**

> "Use the self-attention matrix T... and guide cross-attention S... to match it."



# This Paper: Optimization (Simple Loss)

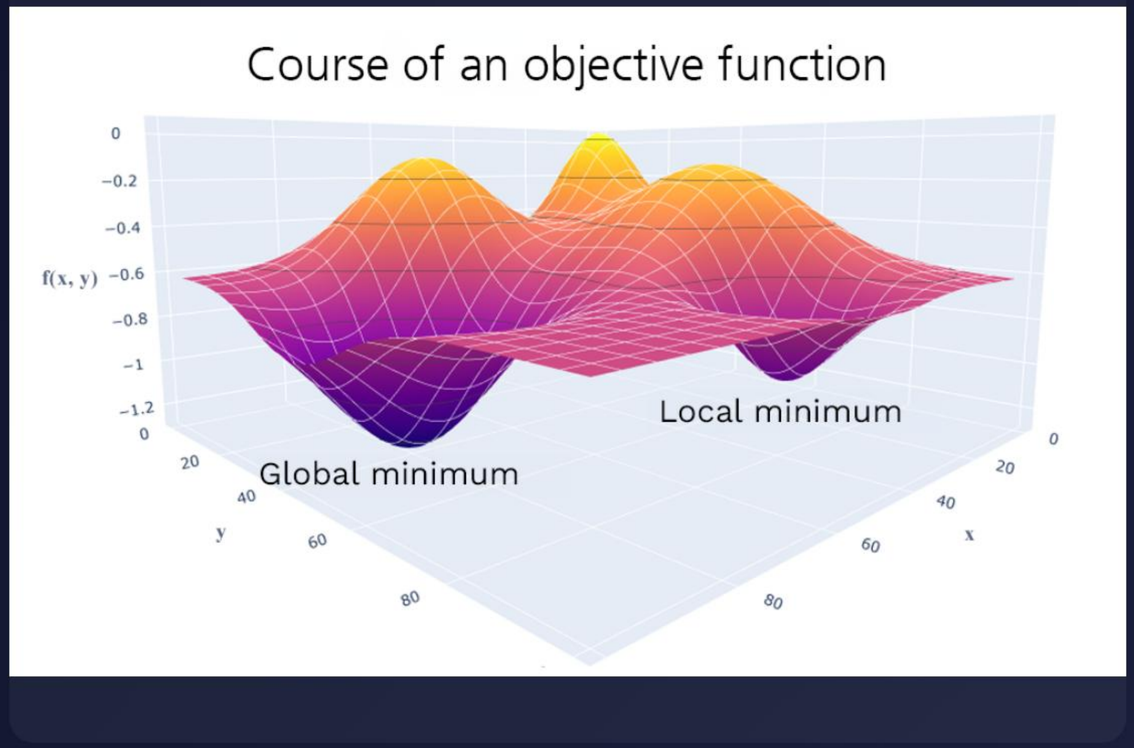
A simple loss is applied:

$$\text{Loss} = |T - S|$$

A small gradient update nudges the latent z:

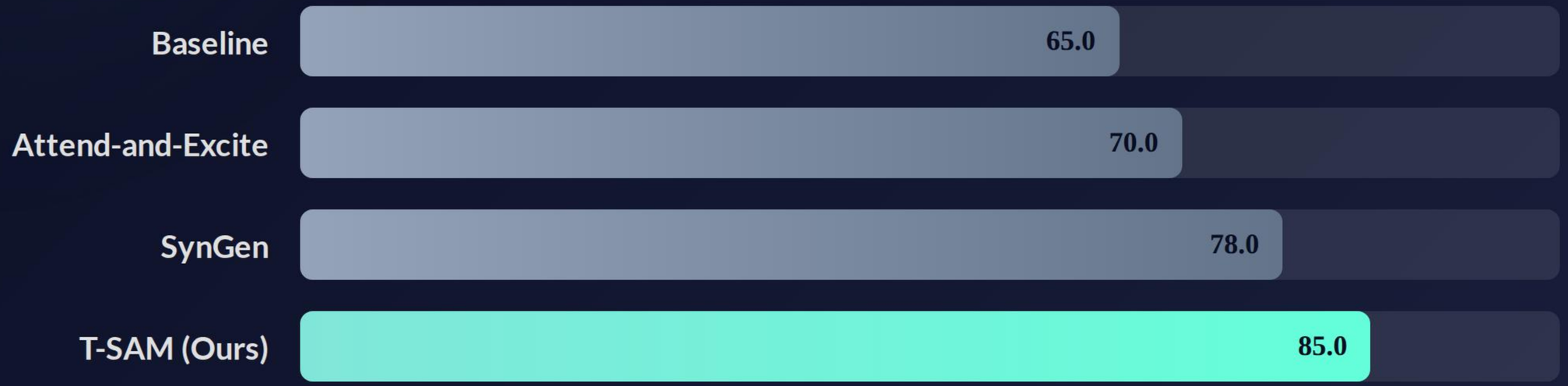
$$z' = z - \alpha \nabla (|T - S|)$$

*“They do not retrain the model. They simply nudge the latent at a few early denoising steps...”*



# Experiments (TIFA Benchmark)

- Tested on 4000+ natural language prompts.
- T-SAM improves TIFA score significantly.
- Handles colors, shapes, relations better than baselines.



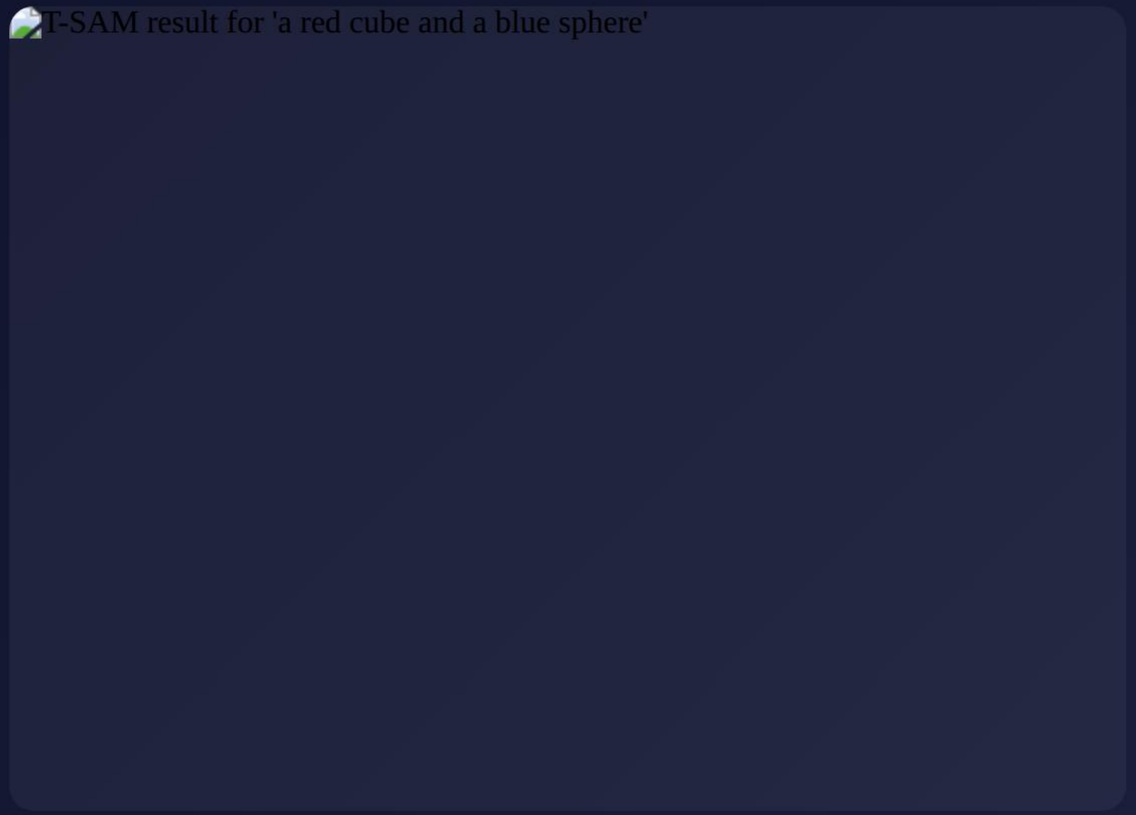

“On the TIFA benchmark, T-SAM outperforms prior approaches across categories like color, shape, and relation understanding.”

# Experiments (Structured Prompts)

Prompts like **“a red cube and a blue sphere.”**

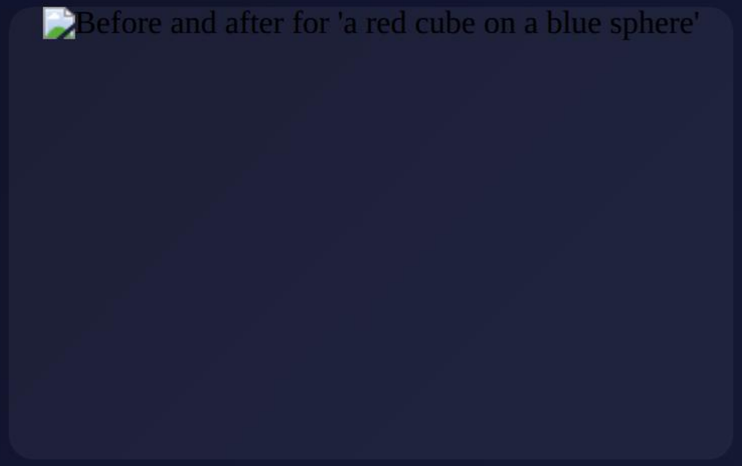
- Better attribute binding.
- No object missing.

> “For structured prompts, T-SAM consistently generates all objects with correct attributes.”



# Qualitative Results (Before vs. After)

"a black car and a white clock"




"a red cube on a blue sphere"

"a yellow bird and a green frog"

# Limitations

- **Computation:** Extra computation during inference (optimizes latent multiple times).
- **Tuning:** Hyperparameters (α, γ) matter and need tuning.
- **Scope:** Only text-side syntax is considered, not image-side structure.
- **Length:** Does not handle long, multi-sentence prompts yet.

# Future Direction

- **Speed:** Faster approximation of T → S alignment.
- **Integration:** Combining with image-side structure (e.g., segmentation, layout).
- **Scale:** Multi-sentence or story-level alignment.
- **Quote:** “The next step could be integrating visual constraints, improving efficiency, and extending syntax to long paragraph-level prompts.”

# Summary & Evolution of Alignment Techniques

## 1. Attend-and-Excite

Ensures **object presence**.
Solves missing objects.

## 2. Linguistic Binding

**Attribute–entity mapping** via an external parser.

## 3. T-SAM (This Paper)

**Parser-free**, self-contained syntax transfer.

## The Key Idea

*"The first to automatically extract syntax from the model's **own** self-attention..."*

# Questions?

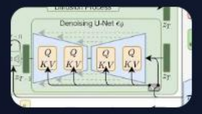
Thank you.

# Image Sources

https://imgcdn.stablediffusionweb.com/2024/5/22/47964e79-97f9-4612-bb3f-e569a345ffeb.jpg

Source: stablediffusionweb.com

---



https://eugeneyan.com/assets/stable-diffusion.webp

Source: eugeneyan.com

---


https://t-sam-diffusion.github.io/static/images/a%20black%20car%20and%20a%20white%20clock_combined_heatmaps_visualization.png

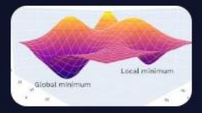
Source: t-sam-diffusion.github.io

---

https://blogger.googleusercontent.com/img/b/R29vZ2xl/AVvXsEiFxnCVzO-u9lvvEnImJunmuQeNI2uSdNqAPlgk9PAobrMA1eyUH6avACRDN5jCw6AdlOeSZEb_pZ9t-K_1ZNjiQfyziMZhz7cFvye3o3ZAXsi6ONhCuKGrV-DjaNdF9A3Ds_61d8DZMAQj67PhGDHCXA6GjgMSFuphD-WMiXVGU80a275_I-MUhDenjGI/w1200-h630-p-k-no-nu/syntactic%20tree.png

Source: www.datatechnotes.com

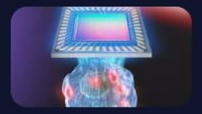
---



https://miro.medium.com/v2/resize:fit:1400/0*i6_mIuTC45rcBSl7.png

Source: medium.com

---



https://scitechdaily.com/images/Biochemical-Quantitative-Phase-Imaging.jpg

Source: scitechdaily.com

# Image Sources

https://www.f1authentics.com/cdn/shop/files/2024_05_30-10884.jpg?v=1717593412&width=2000

Source: www.f1authentics.com

---

https://i.ytimg.com/vi/20xeGkqIZeE/hq720.jpg?sqp=-oaymwEhCK4FEIIDSFryq4qpAxMIARUAAAAAGAElAADIQj0AgKJD&rs=AOn4CLDNzaDvTHmIvSmmg-QCBPzQvt7rwA

Source: www.youtube.com

---

http://www.reptilecentre.com/cdn/shop/articles/3d2e35983dbfde3102b2309f8b70dbee.jpg?v=1718288624

Source: www.reptilecentre.com

---

https://img.freepik.com/premium-photo/vibrant-gradient-abstract-art-with-fluid-swirling-colors-cyan-accent-generative-ai_1149279-7532.jpg

Source: www.freepik.com