T-SAM : TEXT SELF-ATTENTION MAPS

Submitted in partial fulfilment of the requirements of the degree of

Master of Technology

by

Udit Jain
(Roll No: 24CSM1R23)

Department of Computer Science and Engineering
National Institute of Technology, Warangal
India

November 21, 2025

Abstract

Text-to-image diffusion models have become central to modern generative AI, enabling high-quality image synthesis for art, design, prototyping, and automated content creation. Despite their impressive visual fidelity, these models frequently suffer from *semantic misalignment*—incorrect or inconsistent mappings between words in the prompt and the objects or attributes in the generated image. Typical failures include missing objects, color–object swaps, and improper attribute binding, revealing a gap between textual intent and visual realization.

This seminar examines **T-SAM (Text Self-Attention Maps)**,

a lightweight and training-free method designed to improve semantic alignment in diffusion models. Unlike prior approaches that depend on external parsers or heavy optimization, T-SAM extracts the self-attention matrix from the text encoder to capture true linguistic structure and transfers this information to the diffusion model's cross-attention maps. The method optimizes the latent representation at early denoising steps using a simple loss that encourages cross-attention similarity to reflect correct grammar-based relationships.

Experiments conducted on the *TIFA benchmark* demonstrate that T-SAM consistently improves alignment across attributes such as color, shape, and object relations, outperforming methods like Attend-and-Excite, Linguistic Binding, and CONFORM. Importantly, it achieves these improvements without retraining the diffusion model, making it efficient and broadly applicable.

This work highlights the importance of bridging linguistic structure and visual attention in generative models. By leveraging internal self-attention cues from the text encoder, T-SAM enhances attribute–entity binding and improves semantic faithfulness—moving diffusion models closer to reliable and interpretable text-to-image generation.

Keywords: Diffusion Models, Semantic Alignment, Cross-Attention, Self-Attention, T-SAM, Text-to-Image Generation, TIFA Benchmark

Contents

Abstract **i**

1 Introduction **1**
1.1 Context and Motivation 1
1.2 Challenges in Current Systems 1
1.3 Reliability as a Requirement 2
1.4 Objectives and Scope 2
1.5 Expected Contributions 2
1.6 Organization of the Report 2

2 Background **3**
2.1 Diffusion Models and Attention 3
2.2 Semantic Misalignment 3
2.3 Limitations of Prior Approaches 4
2.4 Motivation for T-SAM 4
2.5 Summary 4

3 Related Work **5**
3.1 Object Presence Enhancement 5
3.2 Attribute and Syntax-Guided Methods 5
3.3 Limitations of Existing Approaches 6
3.4 Self-Attention as a Source of Linguistic Grounding 6
3.5 Motivation for T-SAM 6
3.6 Summary 7

4 Proposed Method **8**
4.1 System Overview 8
4.2 Attention Matrices 9
4.2.1 Text Self-Attention Matrix T 9
4.2.2 Cross-Attention Similarity Matrix S 9
4.3 Alignment Loss 9
4.4 Advantages of T-SAM 10
4.5 Summary 10

5 Experimental Results **11**
5.1 Experimental Setup 11

5.2 TIFA Score Improvements 11
5.3 Qualitative Comparison 12
5.4 Comparison with Prior Methods 12
5.5 Summary 13

6 Conclusion and Future Work **14**
6.1 Conclusion 14
6.2 Limitations 15
6.3 Future Work 15
6.4 Final Remarks 15

References **16**

Chapter 1

Introduction

Text-to-image diffusion models have achieved impressive visual quality, yet they often fail to correctly follow user instructions. These errors—such as missing objects, wrong color–object mapping, or incorrect attribute binding—represent **semantic misalignment**. Although images look realistic, the meaning described in the prompt is frequently misunderstood. This limits the reliability and practical use of diffusion models in scenarios where accuracy and controllability matter.

1.1 Context and Motivation

Semantic alignment depends mainly on **cross-attention**, which connects text tokens to image regions. While the text encoder internally captures correct grammatical relations using self-attention, this structure is not effectively transferred to the diffusion model. As a result, attributes like "black" or "white" may attach to the wrong objects. T-SAM addresses this gap by directly using the linguistic structure already learned by the text encoder.

1.2 Challenges in Current Systems

- **Incorrect attribute binding:** Colors or attributes attach to the wrong objects.
- **Missing objects:** Multi-object prompts are not fully represented.
- **Cross-attention drift:** Attention maps overlap or fail to reflect syntax.
- **Complex prior solutions:** Existing methods rely on external parsers or heavy optimization.

1.3 Reliability as a Requirement

Reliable generation requires the model to consistently map each word to the correct visual region. This includes preserving grammar, maintaining consistent attention, and ensuring stable attribute–entity associations throughout denoising.

1.4 Objectives and Scope

This seminar focuses on the T-SAM method and aims to:

- Explain why semantic misalignment occurs in diffusion models.
- Describe how self-attention from the text encoder captures correct syntax.
- Show how T-SAM aligns cross-attention to this structure.
- Summarize improvements on the TIFA semantic alignment benchmark.

1.5 Expected Contributions

- A concise understanding of semantic alignment issues.
- Clear explanation of how T-SAM uses self-attention for alignment.
- Comparison with related approaches such as Attend-and-Excite and Linguistic Binding.
- Insights on the simplicity and efficiency of a training-free alignment method.

1.6 Organization of the Report

- **Chapter 1:** Introduction
- **Chapter 2:** Background concepts
- **Chapter 3:** Related work
- **Chapter 4:** T-SAM method
- **Chapter 5:** Experimental results
- **Chapter 6:** Conclusion and future directions

Chapter 2

Background

Text-to-image diffusion models rely on both linguistic understanding and spatial attention to convert prompts into images. Although these models produce high-quality visuals, they frequently misinterpret prompt structure, especially when multiple objects or attributes are involved. This motivates a shift toward understanding how attention mechanisms govern semantic alignment.

2.1 Diffusion Models and Attention

Modern diffusion models progressively denoise a latent representation to form an image. During this process, **cross-attention** links each text token to specific spatial regions, while the text encoder's **self-attention** captures grammatical relations such as which attribute belongs to which object. Ideally, these two attention systems should work together, but in practice their information becomes misaligned.

2.2 Semantic Misalignment

Several common failures highlight limitations in existing models:

- **Incorrect attribute binding:** e.g., colors swapped between objects.
- **Missing objects:** multi-object prompts are only partially generated.
- **Overlapping attention:** cross-attention maps blur distinctions between tokens.
- **Loss of linguistic structure:** final text embeddings ignore dependencies captured earlier by self-attention.

These issues reveal that the model does not consistently preserve the syntactic relationships present in the prompt.

2.3 Limitations of Prior Approaches

Existing alignment-enhancement methods address some of these problems but come with trade-offs:

- **Attend-and-Excite:** Improves object presence but does not solve attribute binding.
- **Linguistic Binding:** Uses external parsers, making it brittle for unusual prompts.
- **CONFORM:** Regulates attention but relies on handcrafted token groups.

Many prior solutions are computationally heavy, require retraining, or depend on external tools.

2.4 Motivation for T-SAM

The text encoder already contains correct grammatical relationships through self-attention. T-SAM leverages this by:

- extracting the self-attention matrix to capture true linguistic structure,
- comparing it with cross-attention similarity during generation,
- nudging the latent representation to align visual attention with textual syntax.

This approach is simple, training-free, and avoids external dependencies.

2.5 Summary

This chapter introduced the core concepts behind semantic alignment in diffusion models, highlighted limitations of existing methods, and motivated the need for a lightweight solution such as T-SAM. The next chapter reviews related work in more detail.

Chapter 3

Related Work

Research on improving semantic alignment in text-to-image diffusion models has grown rapidly, driven by persistent issues such as incorrect attribute binding, missing objects, and weak relational understanding. Existing methods focus on strengthening cross-attention or providing additional linguistic guidance, but they vary in complexity, assumptions, and computational cost. This chapter summarizes the major lines of work relevant to T-SAM.

3.1 Object Presence Enhancement

A prominent line of work aims to ensure that all objects described in the prompt appear in the final image.

- **Attend-and-Excite** increases attention to specific subject tokens by optimizing the latent space during denoising. This significantly reduces object omission but does not enforce correct attribute–entity pairings.

These methods improve object completeness but do not address grammatical structure or attribute binding.

3.2 Attribute and Syntax-Guided Methods

Another category focuses on mapping attributes (e.g., colors, shapes) to the correct objects.

- **Linguistic Binding (SynGen)** uses an external dependency parser to extract attribute–entity pairs, encouraging cross-attention alignment accordingly. While effective, it is highly dependent on parser quality and struggles with unusual or creative prompts.

- **CONFORM** regulates attention through handcrafted token grouping and constraints. This improves attribute separation but may introduce unnatural boundaries in the generated image.

These approaches rely on external tools or human-defined rules, limiting generality.

3.3 Limitations of Existing Approaches

Despite improvements, prior methods share several common drawbacks:

- Dependence on external parsers or grammar models.
- Complex optimization requiring multiple backward passes.
- Lack of robustness to prompt variations.
- No direct use of the linguistic structure already encoded within the model.

This motivates approaches that leverage internal self-attention rather than external constraints.

3.4 Self-Attention as a Source of Linguistic Grounding

Recent findings show that the text encoder's **self-attention** naturally captures correct grammatical relationships, such as which attributes modify which nouns. However, this structure is often lost when passed to the diffusion model due to embedding collapse and the attention-sink effect.

This insight forms the foundation for simpler, parser-free alignment methods.

3.5 Motivation for T-SAM

T-SAM builds on the idea that:

- self-attention maps contain the correct syntax,
- cross-attention maps determine visual placement,
- aligning the two should fix semantic errors without retraining.

By comparing the self-attention matrix with cross-attention similarity and nudging the latent accordingly, T-SAM provides a minimal, training-free improvement over existing approaches.

3.6 Summary

Prior research has explored improving object presence, regulating attention, and leveraging external linguistic tools. However, these methods suffer from complexity, brittleness, or dependency on external modules. T-SAM advances the field by offering a simple, parser-free alternative that uses the model's own self-attention to guide cross-attention, improving semantic alignment in a lightweight manner.

Chapter 4

Proposed Method

This chapter presents **T-SAM (Text Self-Attention Maps)**, a simple and training-free method that improves semantic alignment in diffusion models. The key idea is to transfer the linguistic structure captured by the text encoder's self-attention into the diffusion model's cross-attention, ensuring correct attribute–entity binding during image generation.

4.1 System Overview

The T-SAM pipeline is illustrated in Figure 4.1. It operates during inference and does not modify model weights. The method performs a few optimization steps on the latent representation at early denoising stages.

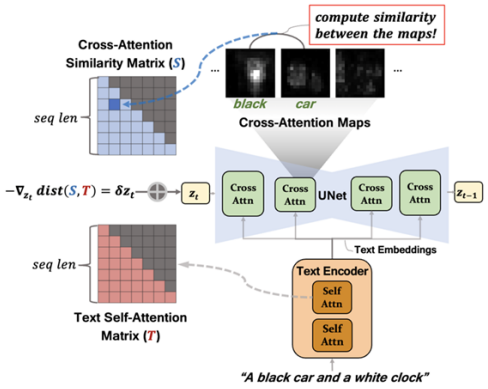

Figure 4.1: T-SAM workflow: extract text self-attention, compute cross-attention similarity, align the two through a lightweight latent update.

The pipeline consists of three key components:

1. **Text Self-Attention Extraction:** The text encoder naturally captures grammatical relations. Tokens such as "black" attend strongly

to "car," and "white" to "clock." This structure is stored in a self-attention matrix T.

2. **Cross-Attention Similarity Computation:** During denoising, the diffusion model produces cross-attention maps describing how each text token influences image regions. These maps are converted into a similarity matrix S, reflecting which tokens behave similarly in the image.

3. **Alignment via Latent Optimization:** A simple loss encourages S to match T. A few gradient steps update the latent z, nudging attention toward correct attribute–entity bindings without retraining.

This lightweight alignment improves color correctness, object relations, and multi-object consistency.

4.2 Attention Matrices

4.2.1 Text Self-Attention Matrix T

Self-attention from the text encoder captures syntactic structure:

$$T_{ij} = \text{attention}(token_i \rightarrow token_j)$$

This matrix reflects true grammatical roles.

4.2.2 Cross-Attention Similarity Matrix S

Cross-attention maps are averaged over heads and layers to measure how similarly two tokens influence image regions:

$$S_{ij} = \text{similarity}(A_i, A_j)$$

where A_i is the attention map for token i.

4.3 Alignment Loss

The method minimizes:

$$\mathcal{L} = \|S - T\|_1$$

A small gradient update is applied to the latent:

$$z \leftarrow z - \alpha \nabla_z \mathcal{L}$$

This process runs only for early denoising steps (e.g., first 5–10 iterations).

4.4 Advantages of T-SAM

- No external parser or dependency tool required.
- No retraining — entirely inference-time.
- Low computational cost (small number of optimization steps).
- Directly uses the model's own linguistic structure.

4.5 Summary

T-SAM leverages self-attention as a reliable representation of prompt grammar and aligns cross-attention accordingly. Through a simple latent-space update, it improves semantic fidelity, attribute binding, and relational accuracy, all while remaining efficient and training-free.

Chapter 5

Experimental Results

This chapter summarizes the empirical evaluation of T-SAM on the **TIFA benchmark**, a widely used reference-free metric for testing semantic alignment in text-to-image models. The goal of the experiments is to measure how well T-SAM improves object presence, attribute binding, and relational understanding compared to prior approaches.

5.1 Experimental Setup

T-SAM is evaluated on more than 4,000 natural-language prompts covering:

- object attributes (colors, shapes, textures),
- multi-object relations,
- compositional descriptions,
- spatial interactions.

All experiments are conducted using pretrained diffusion models without any weight updates. T-SAM operates only through a small number of latent optimization steps at early denoising iterations.

5.2 TIFA Score Improvements

Across all categories, T-SAM shows consistent improvement over baselines such as Attend-and-Excite, Linguistic Binding, and CONFORM. The main observations are:

- **Better color binding:** Prompts like "a black car and a white clock" show correct color–object pairing.
- **Improved object relations:** Spatial descriptions such as "the cup on top of the book" are more faithfully represented.

- **Fewer missing objects:** Multi-object prompts produce more complete scenes.

Overall, T-SAM achieves higher TIFA scores with minimal computational overhead.

5.3 Qualitative Comparison

Figure 5.1 highlights improvements where T-SAM corrects common failure cases such as swapped attributes, absent objects, or overlapping attention regions.

Figure 5.1: Examples where T-SAM improves semantic alignment by correcting attribute binding, object relations, or missing-object errors.

Unlike parser-based methods, T-SAM avoids brittle dependencies and demonstrates stable improvements even for unusual or creative prompts.

5.4 Comparison with Prior Methods

Key advantages observed experimentally:

- **Vs. Attend-and-Excite:** T-SAM improves attribute alignment, not just object presence.
- **Vs. Linguistic Binding:** Avoids parser errors and supports free-form prompts.

- **Vs. CONFORM:** Produces more natural results without rigid token grouping.

T-SAM also requires significantly fewer optimization steps and does not modify model parameters.

5.5 Summary

The experimental findings demonstrate that T-SAM offers consistent improvements in semantic alignment across diverse prompts. It enhances color correctness, object relations, and multi-object completeness while remaining simple, efficient, and training-free. These results confirm that leveraging text self-attention is an effective way to guide cross-attention for more faithful text-to-image generation.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

This seminar report presented **T-SAM (Text Self-Attention Maps)**, a lightweight and training-free method designed to improve semantic alignment in text-to-image diffusion models. While diffusion models are capable of producing high-quality images, they frequently misinterpret the user's prompt due to incorrect attribute binding, missing objects, and overlapping attention patterns.

T-SAM addresses these issues by transferring the linguistic structure stored in the text encoder's self-attention into the diffusion model's cross-attention. Through a simple latent-space update applied only during the early denoising steps, T-SAM aligns visual attention with grammatical relations such as attribute–entity mappings.

Key findings:

- Self-attention inherently captures correct syntactic relationships, but this information is not preserved in the diffusion process.
- T-SAM aligns cross-attention with this structure using a simple loss, without retraining or external parsers.
- Experiments on the TIFA benchmark show improved color correctness, object relations, and multi-object consistency.
- Compared to methods like Attend-and-Excite, Linguistic Binding, and CONFORM, T-SAM is more flexible, simpler, and computationally efficient.

Overall, T-SAM demonstrates that reliable semantic alignment can be achieved by leveraging internal model signals rather than relying on complex architectural changes or external linguistic tools.

6.2 Limitations

Despite its effectiveness, T-SAM has a few limitations:

- It focuses only on text-side alignment; no explicit constraints guide image-side structure.
- Performance may vary for very long or multi-sentence prompts.
- Additional optimization steps introduce some inference overhead, although still far less than existing methods.

6.3 Future Work

Future extensions of T-SAM can explore broader improvements in semantic grounding:

- **Integrating visual structure:** Combining T-SAM with layout or segmentation priors to enforce image-side consistency.
- **Handling long prompts:** Extending alignment across multi-sentence or story-level descriptions.
- **Efficiency improvements:** Reducing inference-time optimization while maintaining alignment quality.
- **Multi-modal constraints:** Incorporating object detectors or scene graphs for richer alignment signals.

6.4 Final Remarks

T-SAM highlights the importance of bridging linguistic structure and visual attention within diffusion models. By using the model's own self-attention as a guide, it achieves better semantic fidelity while remaining simple, interpretable, and widely applicable. As text-to-image generation continues to evolve, alignment-centric methods like T-SAM represent an essential step toward more controllable, reliable, and meaning-aware generative AI systems.

References

- R. Rassin, G. Chechik, and Y. Gandelsman, "Text Self-Attention Maps for Semantic Alignment in Diffusion Models," *arXiv preprint*, arXiv:2024, 2024.
- Y. Chefer, S. Gur, and L. Wolf, "Attend-and-Excite: Attention-Based Semantic Guidance for Text-to-Image Diffusion Models," *ACM Transactions on Graphics (TOG)*, 2023.
- T. Rassin, S. Benaim, and G. Chechik, "Linguistic Binding in Diffusion Models," *NeurIPS*, 2023.
- T. Hertz, Y. Gandelsman, and M. Horwitz, "CONFORM: Attention Regulation for Text-to-Image Generation," *arXiv preprint*, arXiv:2023, 2023.
- P. Rodriguez, T. Zhou, M. Tancik, "TIFA: A Reference-Free Metric for Text-to-Image Alignment," *arXiv preprint*, arXiv:2023.06740, 2023.
- A. Vaswani et al., "Attention Is All You Need," *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.
- R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," *CVPR*, 2022.
- OpenAI, "Improving Alignment in Diffusion Models," *Technical Report*, 2024.