

Generative Image Models do NOT “Know” how to interpret “No”

Anish Pravin Kulkarni, Computer Science (B.S.)

Mentors: Dr. Bharatesh Chakravarthi, Prof. Yiran Luo
School of Computing and Augmented Intelligence



Introduction and Background

- The Problem:** State-of-the-art (SOTA) Text-to-Image (T2I) models often fail to understand negation, generating the very concepts told to exclude.

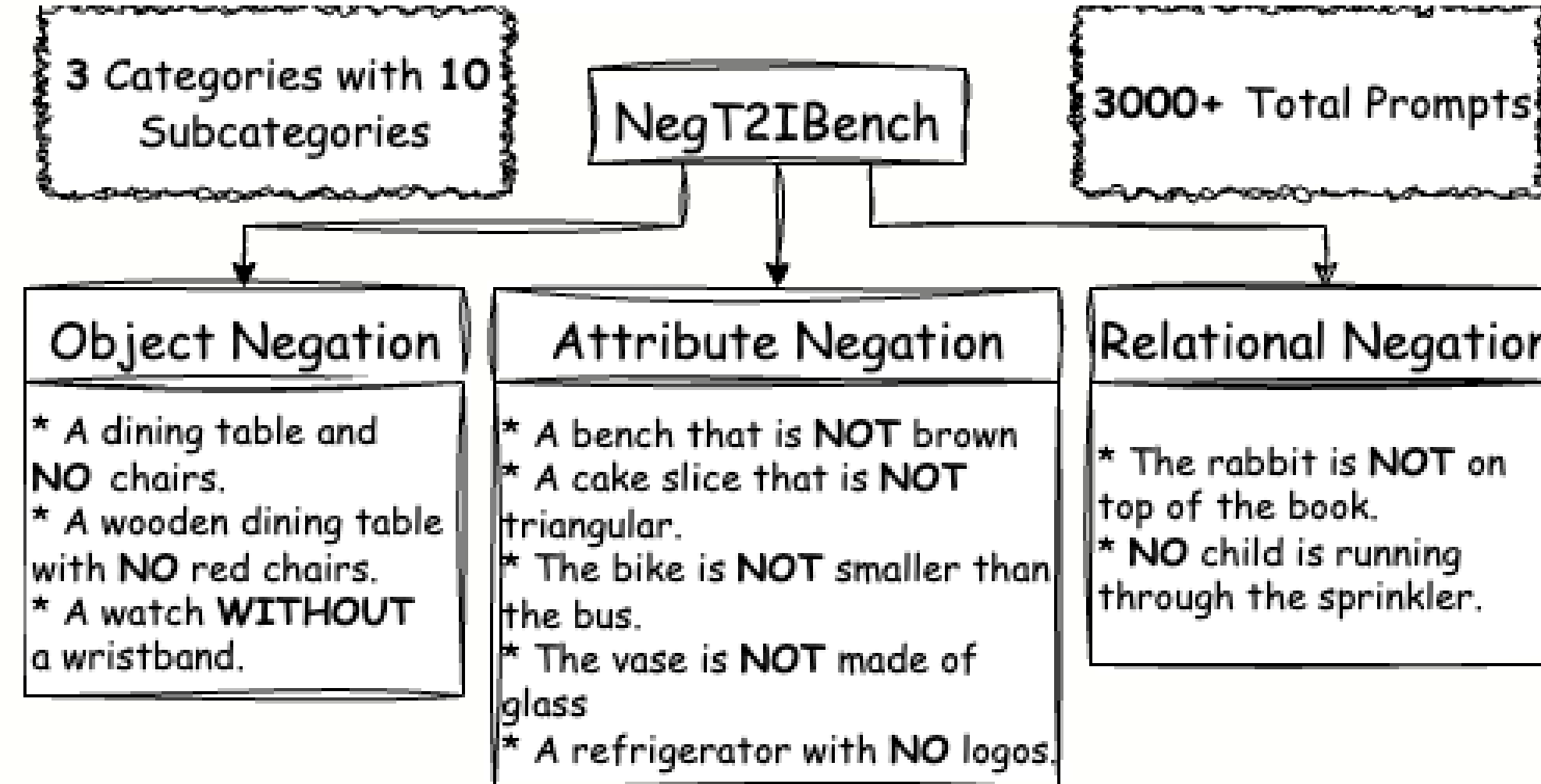


- The Impact:** This critical flaw limits user control, precision, and model reliability for professional applications. It is shown that improving negation understanding boosts overall model understanding.
- Our Goal:** Create a large-scale, reliable benchmark to quantify “negation understanding” in modern T2I models and evaluate current solutions.

Key Challenges

- Extreme Data Scarcity:** Negation words comprise less than 0.7% of captions and 0.08% of words in the LAION-400M training set.^[8]
- Conceptual Entanglement and Data Bias:** Objects appearing together frequently in the training set are almost inseparable (example: “car” and “wheels”).
- Pattern Matching:** Vision Language Models (VLMs) are *pattern matchers* and not *logicians*. They are trained to associate words with visual features, where “no” has no visual pattern.^[11]
- The Evaluation Gap:** Current popular T2I benchmarks, do not evaluate negation in all forms.

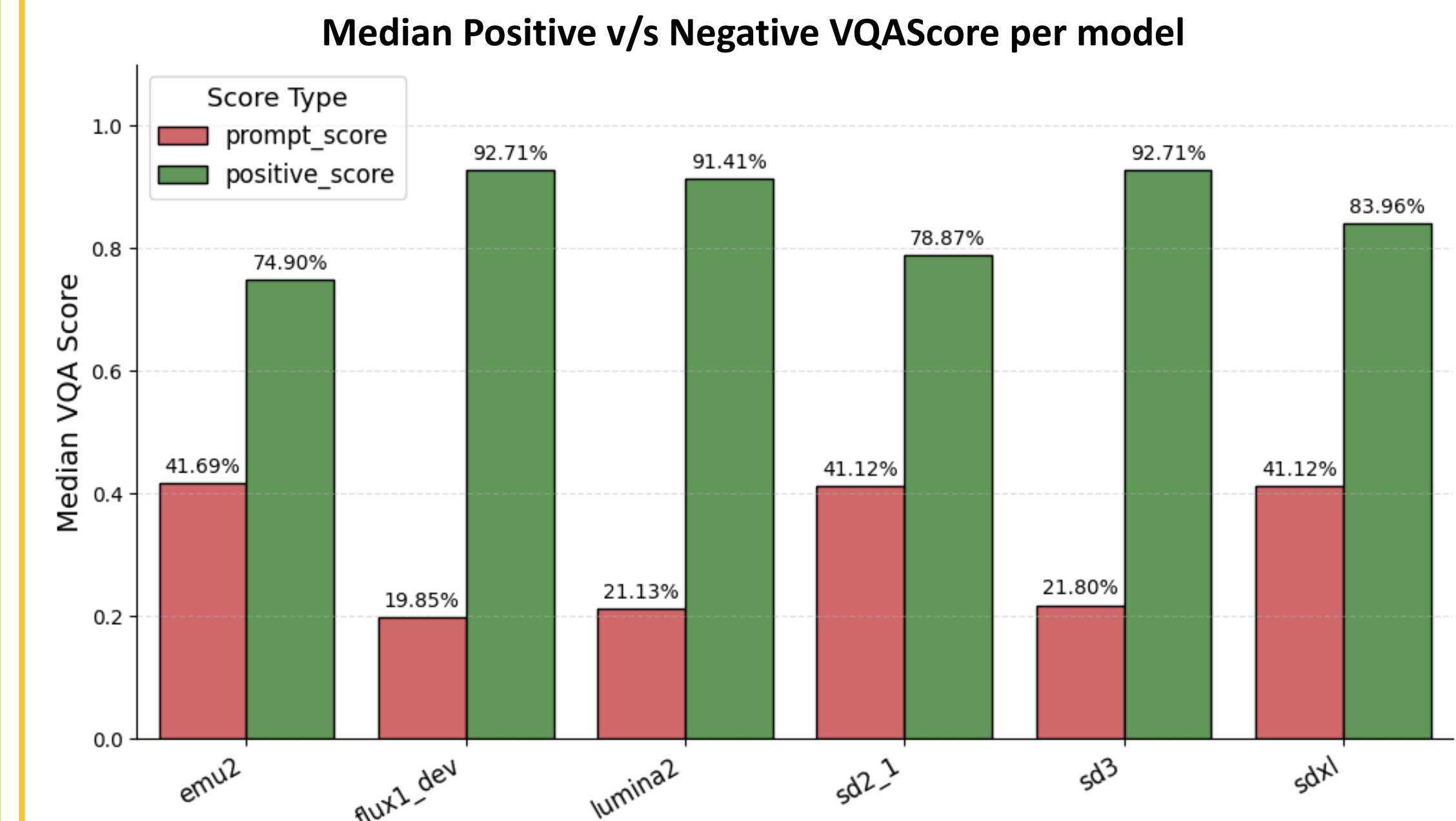
Experimental Setup and Evaluation Metrics



- Benchmark:** Adapted prompts from *T2I-CompBench*^[4] and *MS-COCO captions*^[6] using *Gemini 2.0 Flash*.
- Evaluation:** A dual approach using Automated *VQAScore*^[7] and manual human assessment.
- VQAScore**^[7]: An automated metric that uses a Visual Question Answering model (*CLIP-FlanT5*^[7]) to score image-text alignment (from 0-1).

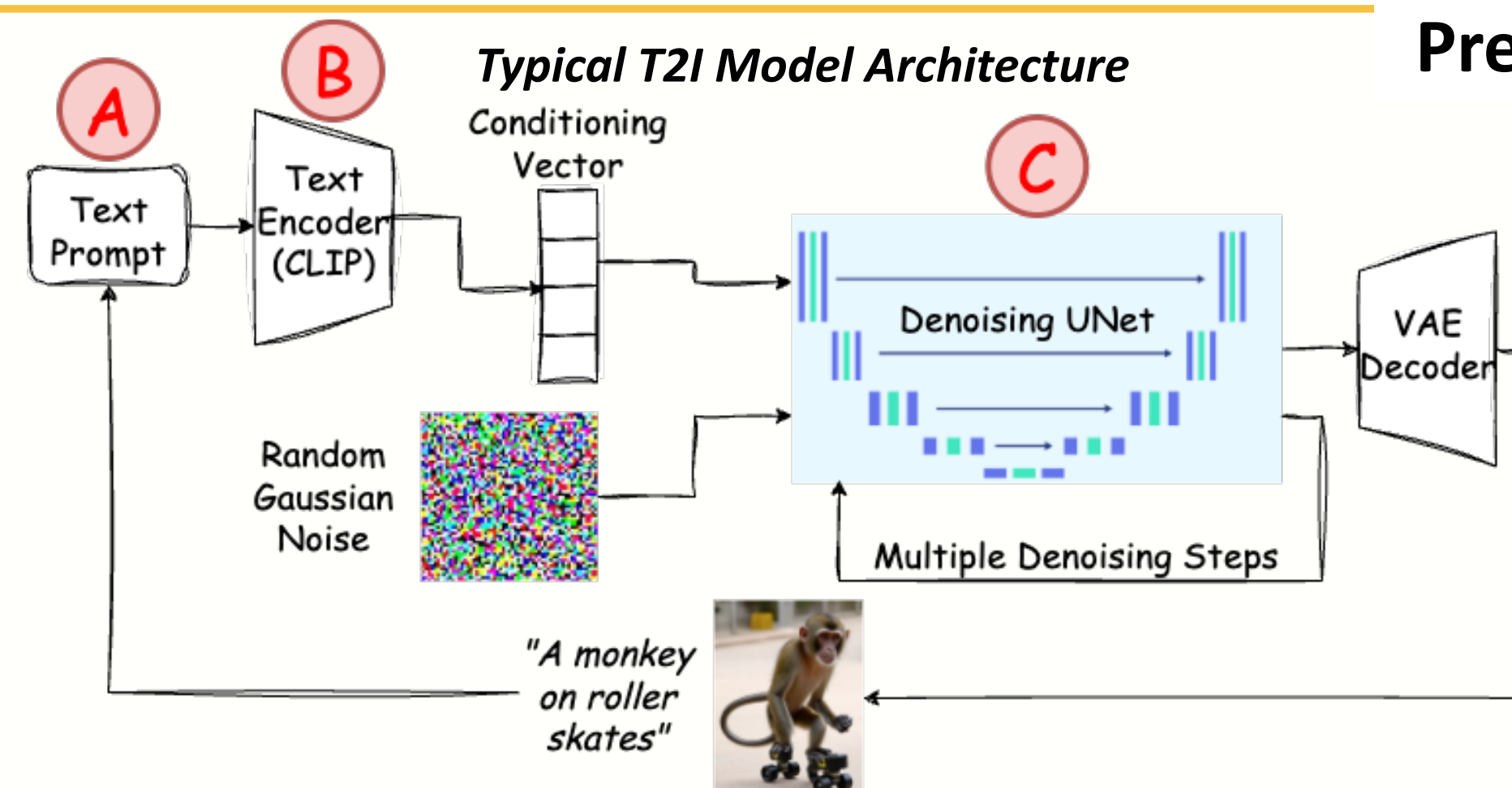
Benchmark Results

- Evaluated 6 SOTA T2I models, revealing a **60-90%** failure rate on the 10 individual subcategories in the benchmark.
- Failure determined via VQAScore:** image fails if it aligns better with the positive prompt than the negative one.



Previous Work Analysis

- B (Text encoder fine-tuning):** SOTA methods (ConCLIP^[8], NegCLIP^[11], TNG-CLIP^[2]) fail, achieving a maximum accuracy of 45.65% — worse than random guessing.^[11]
- C (UNet Attention Manipulation):** Even though limited success, our tests show that these methods often hallucinate or fail to follow the negative prompt.



- A (Prompt Engineering):** This can fix basic errors, but it proves largely ineffective for complex negations like absence of a *part* of an object (*part absence*).



Value Sign Flip (VSF)^[10]

Prompt: A simple wall clock.
Negative Prompt: Minute hand

VSF (left): Hallucinating; multiple hands
NAG (right): Failing; minute hand present



Normalized Attention Guidance (NAG)^[3]

REFERENCES

- [1] Alhamoud, Kumail, et al. “Vision-Language Models Do Not Understand Negation.” *arXiv.Org*, 13 May 2025, arxiv.org/abs/2501.09425.
- [2] Cai, Yuliang, et al. “TNG-Clip: Training-Time Negation Data Generation for Negation Awareness of Clip.” *arXiv.Org*, 24 May 2025, arxiv.org/abs/2505.18434.
- [3] Chen, Dar-Yen, et al. “Normalized Attention Guidance: Universal Negative Guidance for Diffusion Models.” *arXiv.Org*, 3 June 2025, arxiv.org/abs/2505.21179.
- [4] Huang, Kaiyi, et al. “T2I-Compbench++: An Enhanced and Comprehensive Benchmark for Compositional Text-to-Image Generation.” *arXiv.Org*, 8 Mar. 2025, arxiv.org/abs/2307.06350.
- [5] Huggingface. “Huggingface/Diffusers: Diffusers: State-of-the-Art Diffusion Models for Image, Video, and Audio Generation in Pytorch and Flax.” *GitHub*, github.com/huggingface/diffusers. Accessed 24 July 2025.
- [6] Lin, Tsung-Yi, et al. “Microsoft Coco: Common Objects in Context.” *arXiv.Org*, 21 Feb. 2015, arxiv.org/abs/1405.0312.
- [7] Lin, Zhiqiu, et al. “Evaluating Text-to-Visual Generation with Image-to-Text Generation.” *arXiv.Org*, 18 June 2024, arxiv.org/abs/2404.01291.
- [8] Park, Junsung, et al. “Know “no” Better: A Data-Driven Approach for Enhancing Negation Awareness in Clip.” *arXiv.Org*, 31 Mar. 2025, arxiv.org/abs/2501.10913.
- [9] Singh, Jaisidh, et al. “Learn ‘No’ to Say ‘Yes’ Better: Improving Vision-Language Models via Negations.” *arXiv.Org*, 29 Mar. 2024, arxiv.org/abs/2403.20312.
- [10] Weathon. “Weathon/VSF: Simple, Efficient, and Effective Negative Guidance in Few-Step Image Generation Models by Value Sign Flip.” *GitHub*, github.com/weathon/VSF/tree/main. Accessed 24 July 2025.
- [11] Yuksekgonul, Mert, et al. “When and Why Vision-Language Models Behave like Bags-of-Words, And...” *Venues*, 29 Sept. 2022, openreview.net/forum?id=KRLUvxh8uax.

Generative Image Models do NOT “Know” how to interpret “No”

Anish Pravin Kulkarni, Computer Science (B.S.)

Mentors: Dr. Bharatesh Chakravarthi, Prof. Yiran Luo
School of Computing and Augmented Intelligence



Introduction and Background

- **The Problem:** State-of-the-art (SOTA) Text-to-Image (T2I) models often fail to understand negation, generating the very concepts told to exclude.

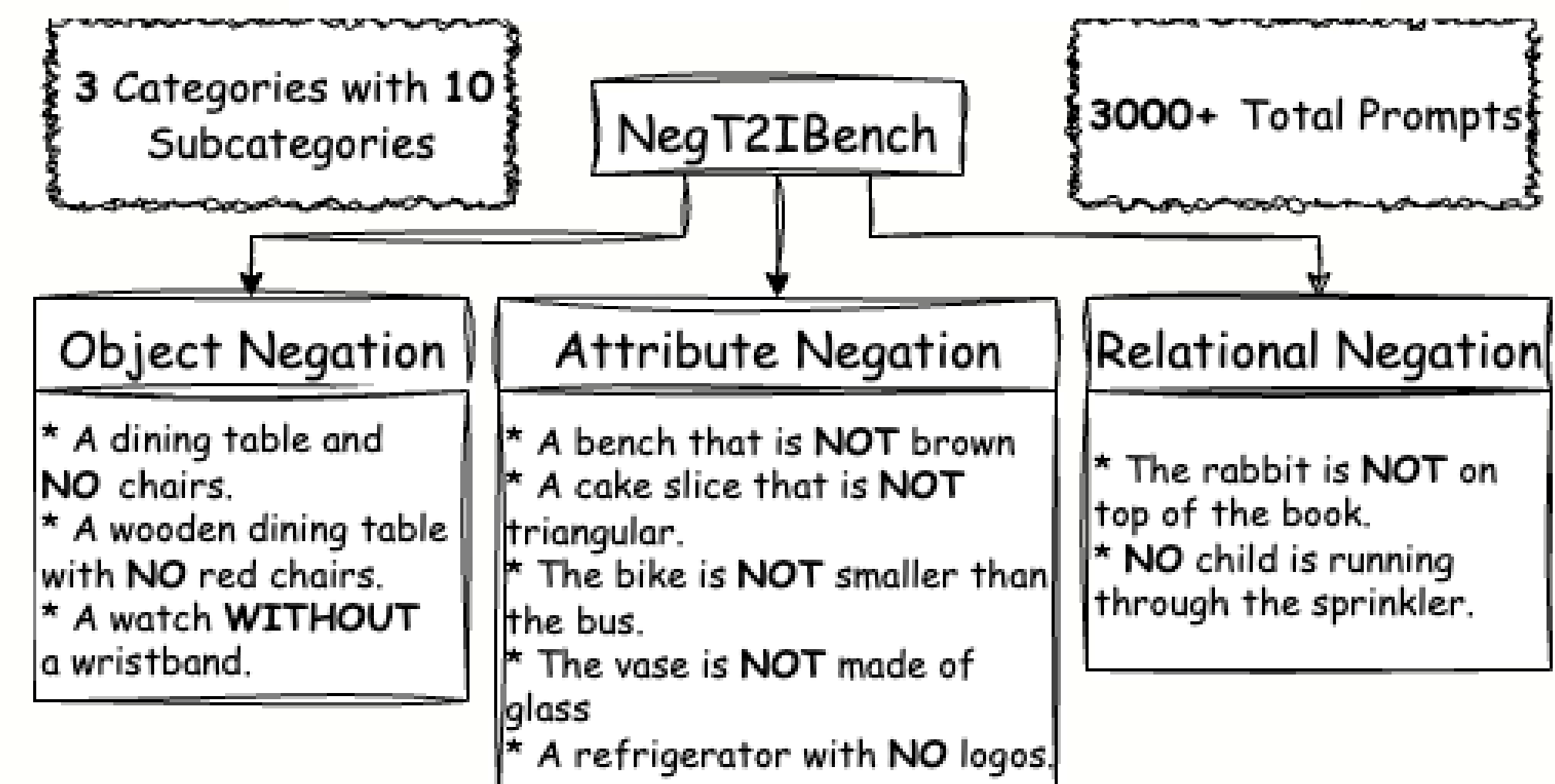


- **The Impact:** This critical flaw limits user control, precision, and model reliability for professional applications. It is shown that improving negation understanding boosts overall model understanding.
- **Our Goal:** Create a large-scale, reliable benchmark to quantify “negation understanding” in modern T2I models and evaluate current solutions.

Key Challenges

- **Extreme Data Scarcity:** Negation words comprise less than 0.7% of captions and 0.08% of words in the LAION-400M training set.
- **Conceptual Entanglement and Data Bias:** Object appearing together frequently in the training set are almost inseparable (example: “car” and “wheels”).
- **Pattern Matching:** Vision Language Models (VLMs) are *pattern matchers* and not *logicians*. They are trained to associate words with visual features, where “no” has no visual pattern.
- **The Evaluation Gap:** Current popular T2I benchmarks, do not evaluate negation in all forms.

Benchmark and Evaluation



- **Benchmark:** Adapted prompts from “T2I-CompBench” and “MS-COCO captions” using Gemini 2.0 Flash.
- **Evaluation:** A dual approach using Automated VQAScore and manual human assessment.
- **VQAScore:** An automated metric that uses a Visual Question Answering model (CLIP-FlanT5) to score image-text alignment (from 0-1).

Benchmark Results

- Evaluated 6 SOTA T2I models, revealing a 60-90% failure rate on the 10 individual subcategories in the benchmark.
- **Failure determined via VQAScore:** image fails if it aligns better with the positive prompt than the negative one.

