Duke | PRATT SCHOOL of ENGINEERING

Semantic Approach to Quantifying the Consistency of Diffusion Model Image Generation

CVPR SEATTLE, WA JUNE 17-21, 2024

Brinnae Bent, PhD
Duke University

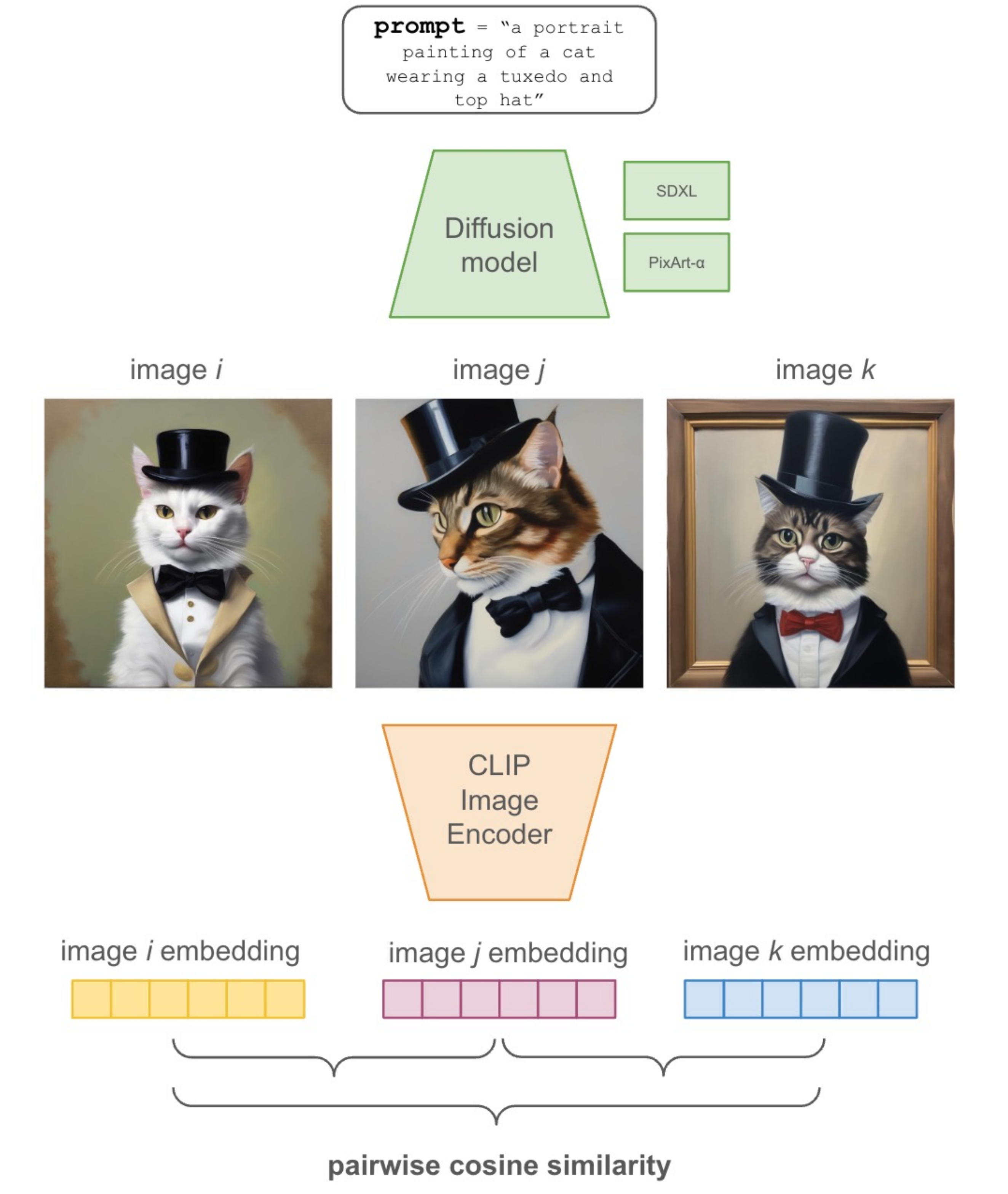
Abstract

In this study, we identify the need for an interpretable, quantitative score of the repeatability, or consistency, of image generation in diffusion models. We propose a semantic approach, using a pairwise mean CLIP (Contrastive Language-Image Pretraining) score as our semantic consistency score. We applied this metric to compare two state-of-the-art open-source image generation diffusion models, Stable Diffusion XL and PixArt-α, and we found statistically significant differences between the semantic consistency scores for the models. Agreement between the Semantic Consistency Score selected model and aggregated human annotations was 94%. We also explored the consistency of SDXL and a LoRA-fine-tuned version of SDXL and found that the fine- tuned model had significantly higher semantic consistency in generated images. The Semantic Consistency Score proposed here offers a measure of image generation alignment, facilitating the evaluation of model architectures for specific tasks and aiding in informed decision-making regarding model selection.

Approach

$$SCS = \frac{1}{N(N-1)/2}\sum_{i=1}^{N}\sum_{j=i+1}^{N}\max(100\,x\cos(E_i, E_j), 0)$$

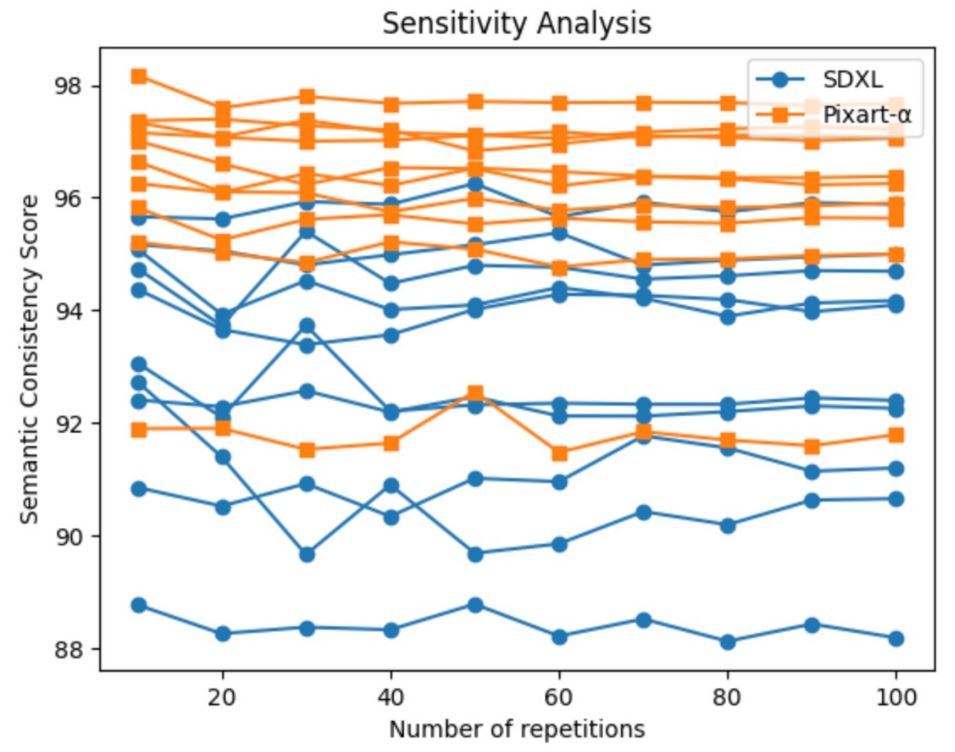
- The Semantic Consistency Score (SCS) is a pairwise mean CLIP score, where N is the number of images, E_i and E_j are the CLIP visual embeddings for images *i* and *j*, respectively.
- For better explainability, the score is bound between 0 and 100, with scores closer to 100 indicating more semantically consistent generated images.
- The summation of all pairwise cosine similarities is divided by the total number of unique image pairs.
- The mean is used to ensure that the metric is sensitive to outliers.



prompt = "a portrait painting of a cat wearing a tuxedo and top hat"

Sensitivity Analysis

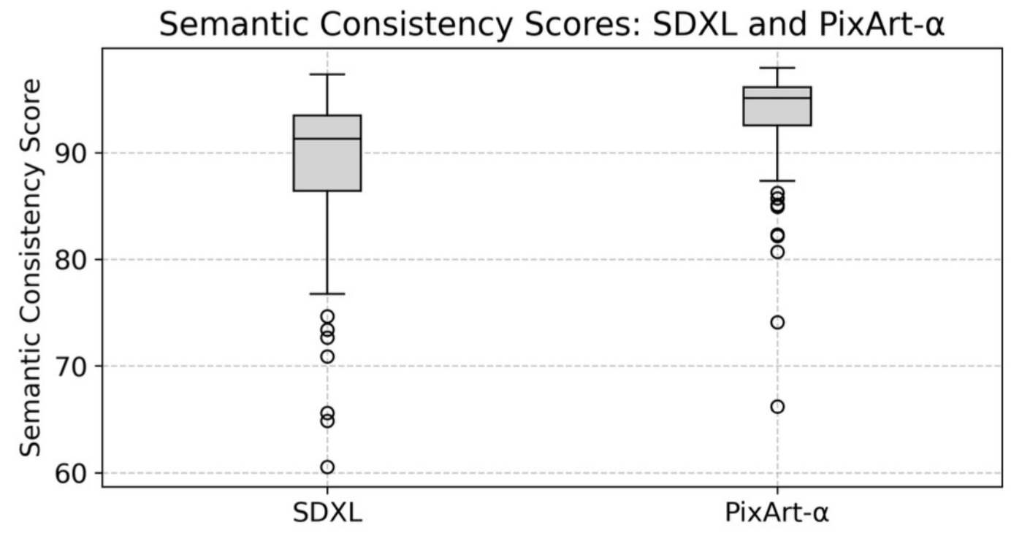
A minimum of 20 repetitions is needed to ensure the score is within 1% of the mean score across all repetitions and within 1% of the score obtained with 100 repetitions.

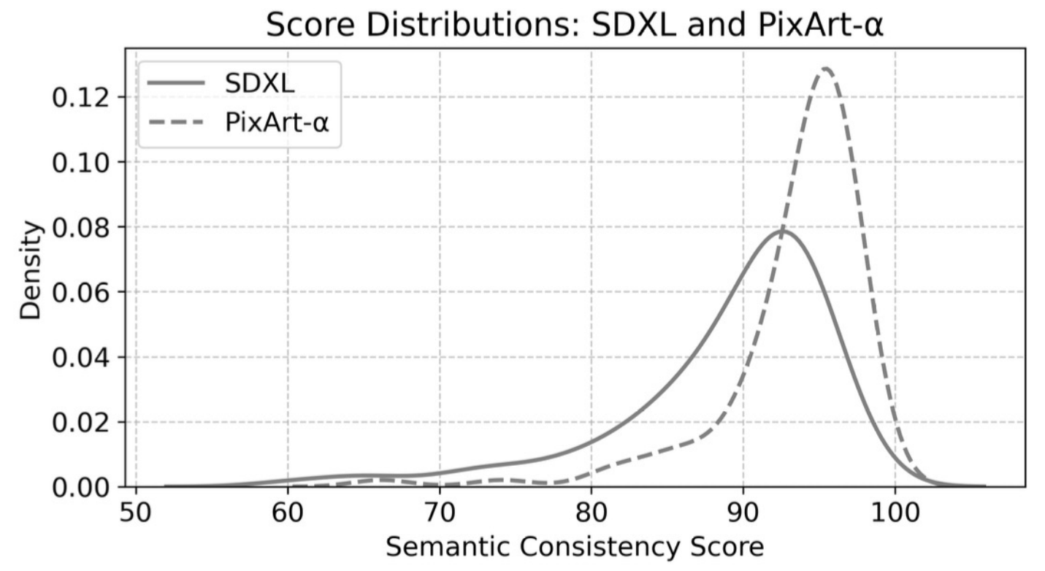


Model Comparison – SDXL and PixArt-α

The distributions of consistency scores across the models were significantly different (KS statistic=0.48, Wilcoxon paired statistic=110.0, p<<0.05)

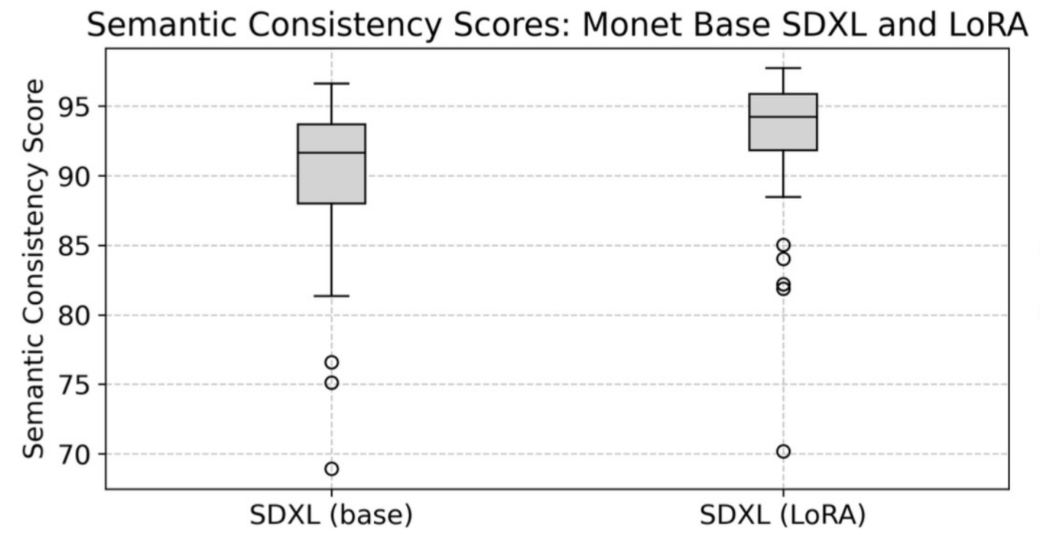
For each pair of results from the models, human annotators labeled which image was the most consistent. The model with the highest semantic consistency score matched the most common selection among annotators 94% of the time. Across all annotators, there was an average agreement rate of 90.9% [range 86%-94%].

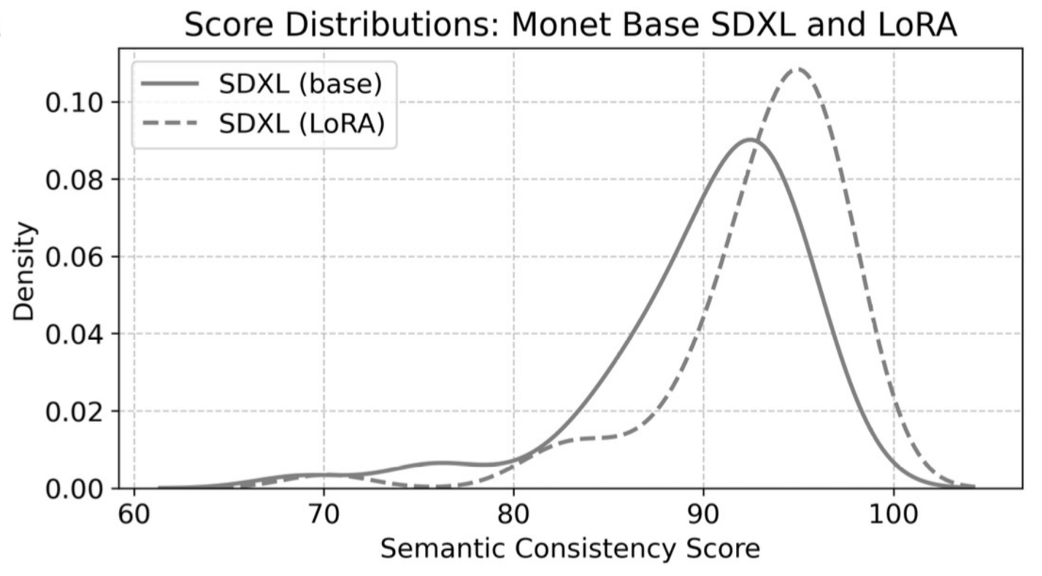




Model Comparison – SDXL and fine-tuned SDXL

The distributions of consistency scores across base SDXL and SDXL fine-tuned on Monet paintings using LoRA were significantly different (KS statistic=0.38, Wilcoxon paired statistic=95.0, p<0.05)





Why do we need this?

- Inherent variability in diffusion model image generation
- Variability differences across different models
- In real world problems, how do we reconcile the desire for diversity and creativity in generated outputs with the need for consistency and coherence relative to the input prompt?
- By quantifying consistency, we can enable decisions reconciling creativity and consistency for particular applications or tasks
- Can provide an assessment of model consistency, detect unintended bias, validate interpretations of model outputs, and enhance user understanding

Future Directions

It is important to note that there exists biases in embedding models like CLIP. Alternative models should be explored.

Quantifying the consistency of generative model outputs could be extended beyond image generation to other modalities, such as evaluating consistency of generated text or audio-based outputs.

Have other ideas? Please reach out and let me know: brinnae.bent@duke.edu

Code and Data Availability

GitHub: https://bit.ly/3QTVnro

Hugging Face: https://bit.ly/4awsSHg