

Not Every Image is Worth a Thousand Words: Quantifying **Originality** in Stable Diffusion



Adi Haviv, Shahar Sarfaty, Uri Hachohen, Niva Elkin-Koren, Roi Livni and Amit H Bermano

Can Text-to-Image models produce original content and how can we measure originality?

Motivation

- The sine qua non of copyright is **originality** (Supreme Court)
- Quantifying substantial similarity via Kolmogorov complexity (Sceffler, Tromer, Varia 22')

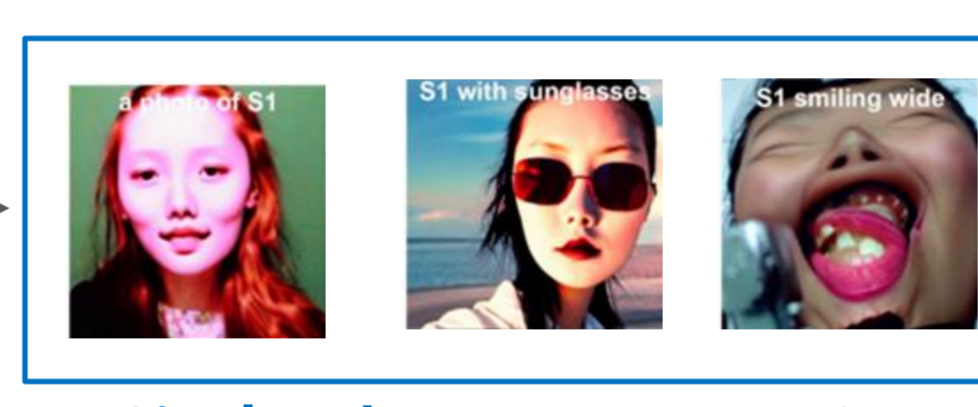
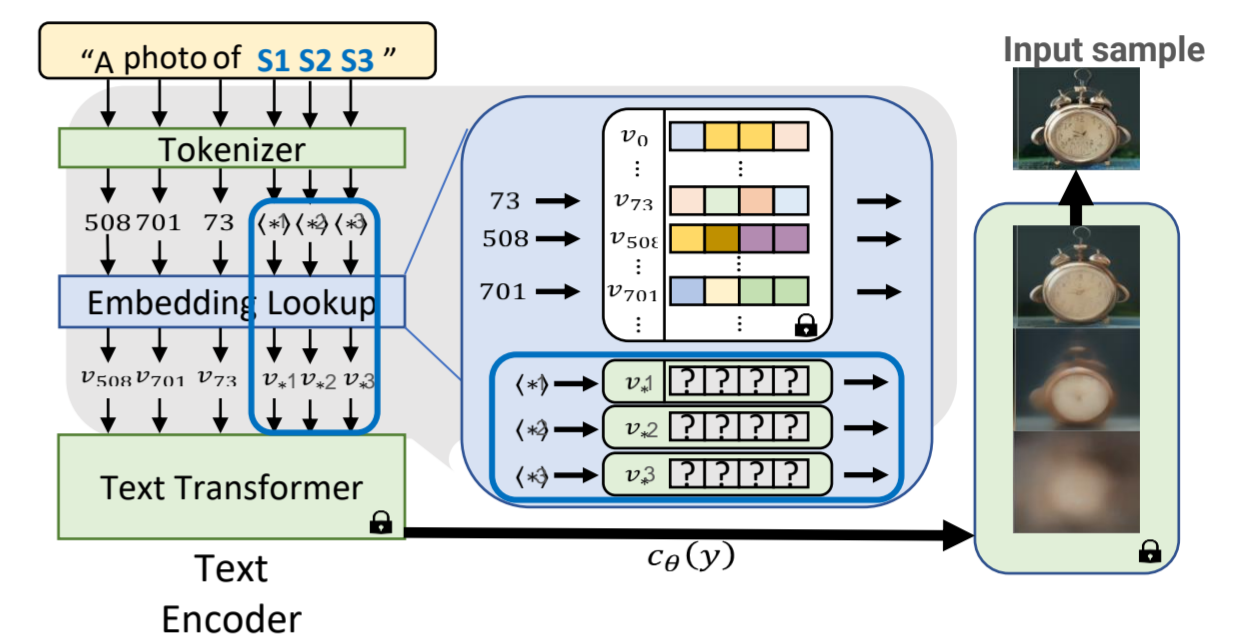
One of the draft provisions in the AI Act which was explicitly copyright-related was a transparency obligation to disclose which data was used for AI training aimed at helping rightsholders to determine if their rights were used (see more on this here). 6 Apr 2024

Lexology
https://www.lexology.com/library/detail/?contentid=8645788&title=The+AI+Act+Calling+for+global+compliance+with+EU+copyright?&context=ai



How many **words** do I need to describe my work?

Textual Inversion



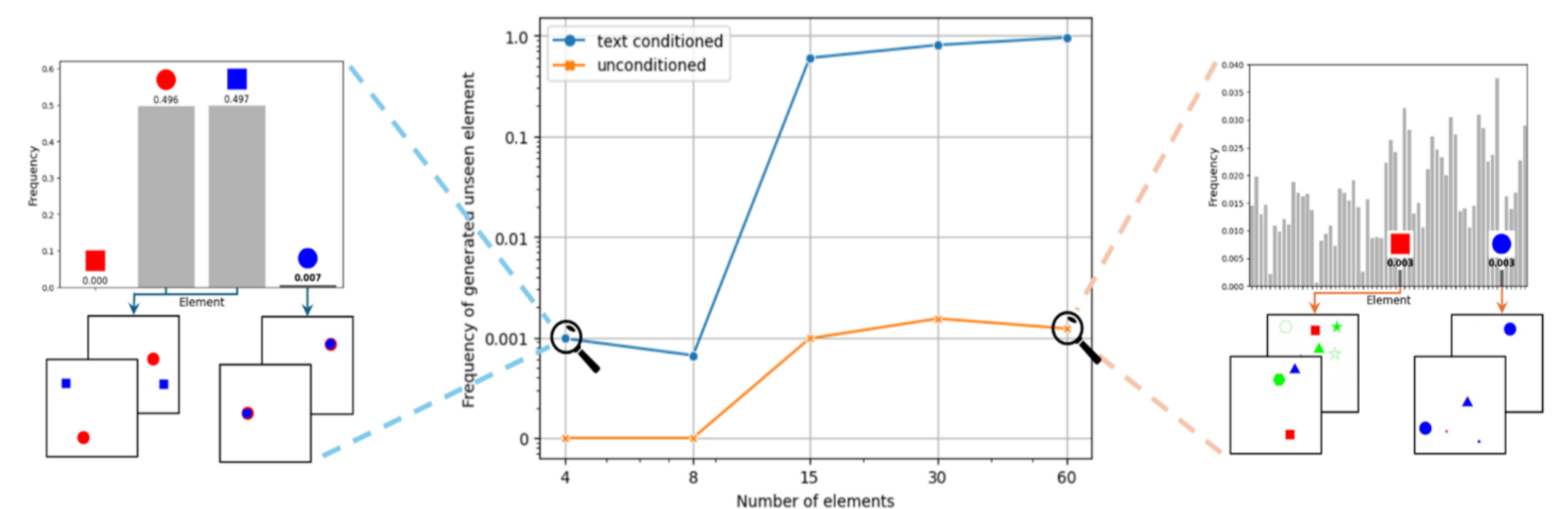
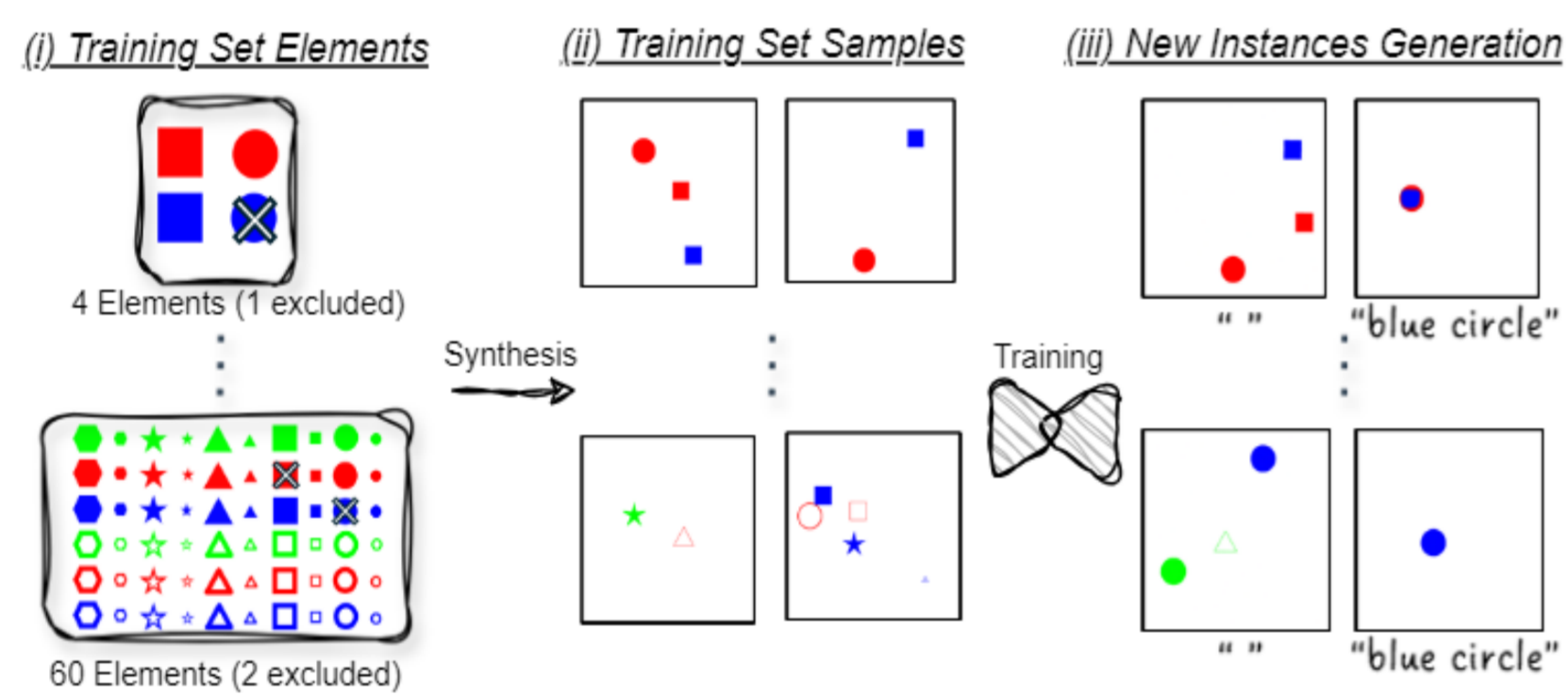
Single tokens Reconstruction



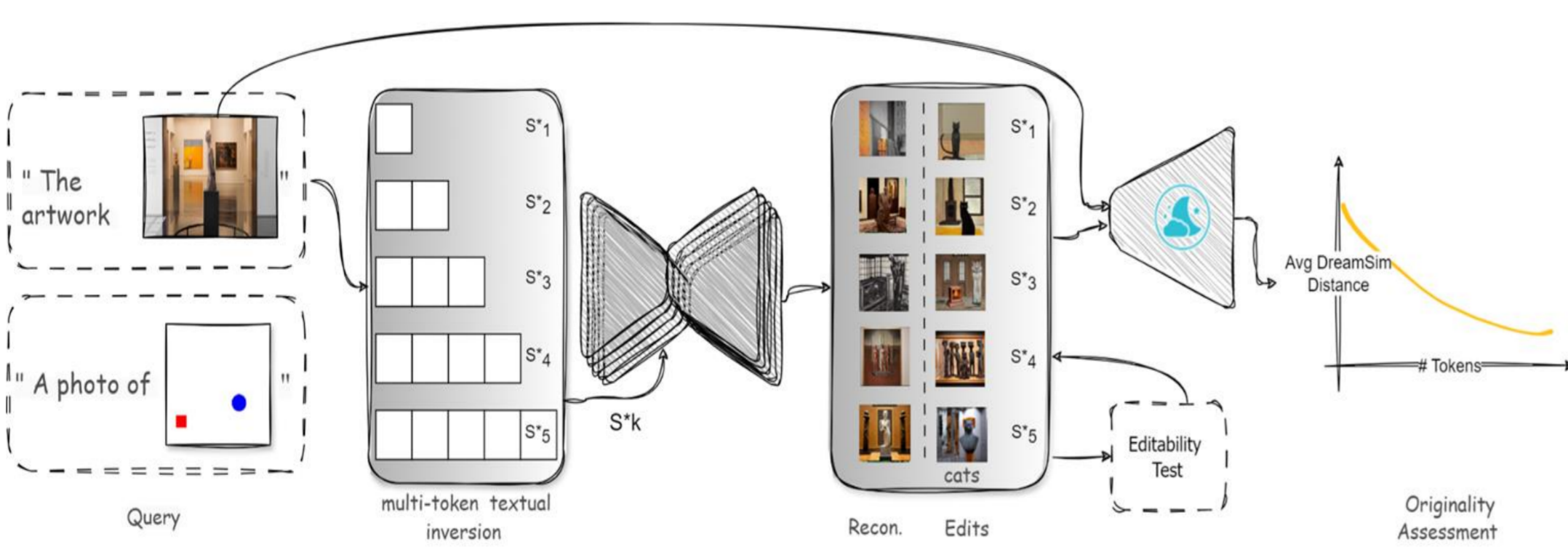
4 tokens Reconstruction

Stable Diffusion Can Create **Original** Content

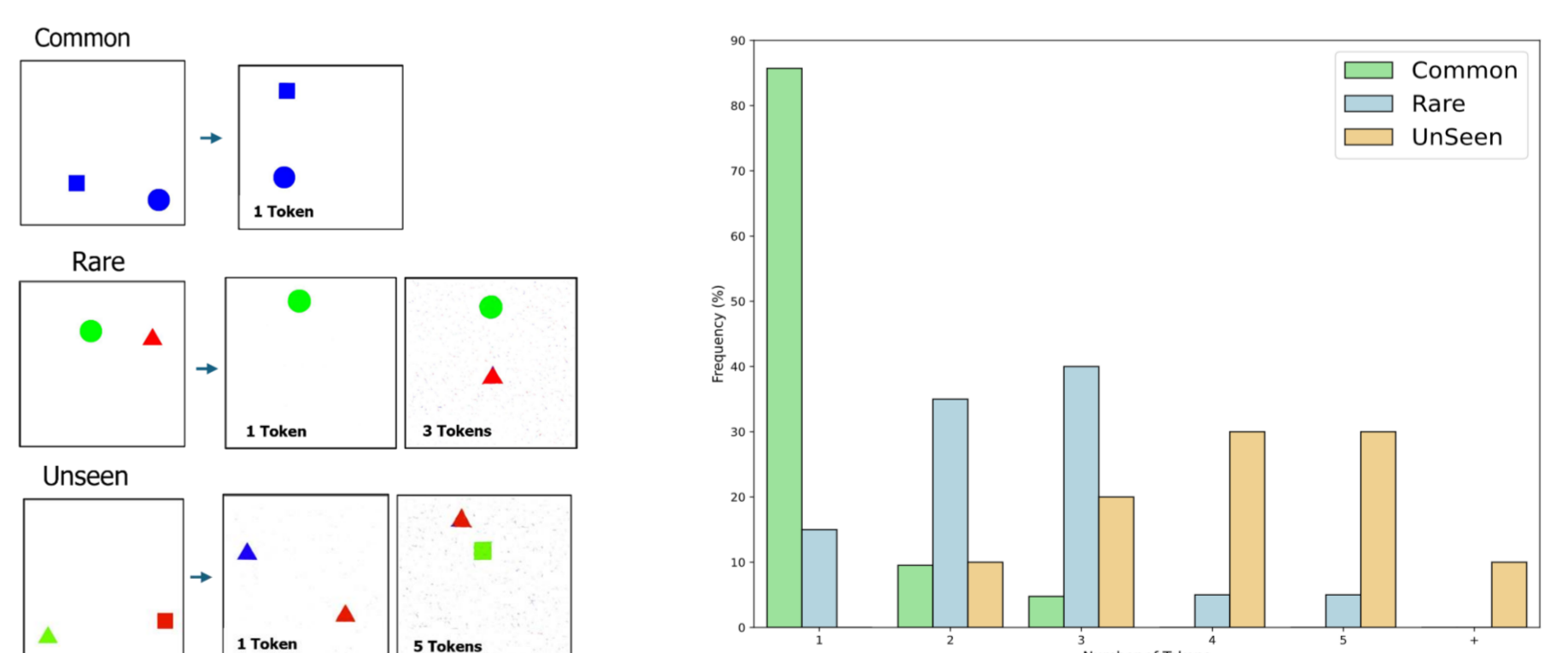
Experiment: We construct a controlled environment where we train Stable Diffusion models from scratch on synthetic data consisting of shapes. By eliminating specific elements from this data, we demonstrate that the model can indeed produce those elements.



Our method to measure **originality**



Controlled Experiment



Real-World Data Experiment

Original:	Domain	Query Image	1 token	2 tokens	3 tokens	4 tokens	5 tokens
Original	Houses						
	Art						
	Sport photography						
	Animals						
	People						
Common:							
Query Image	Houses	Art	Sport Photography	Animals	People		
Single Token Reconstruction							

Takeaways and Conclusions

- Approach is inspired by legal definitions of **originality**, aims to assess whether a model can produce original content **without relying on** specific prompts or having the **training data** of the model.
- Applications in copyrights assessment:
 - Supports **courts** in assessing copyright scope.
 - Aids **copyright owners** in negotiating fair licensing deals.
 - Provides insights for **policymakers** adapting copyright law to GenAI.
- Text-to-Image models generalization relies on **diverse training data** for original content creation.

