Watermarking Diffusion Graph Models

GUISE: Graph GaUssIan Shading watErmark

Renyi Yang[1]

[1] R.Yang-7@student.tudelft.nl

28-06-2024 Final Poster

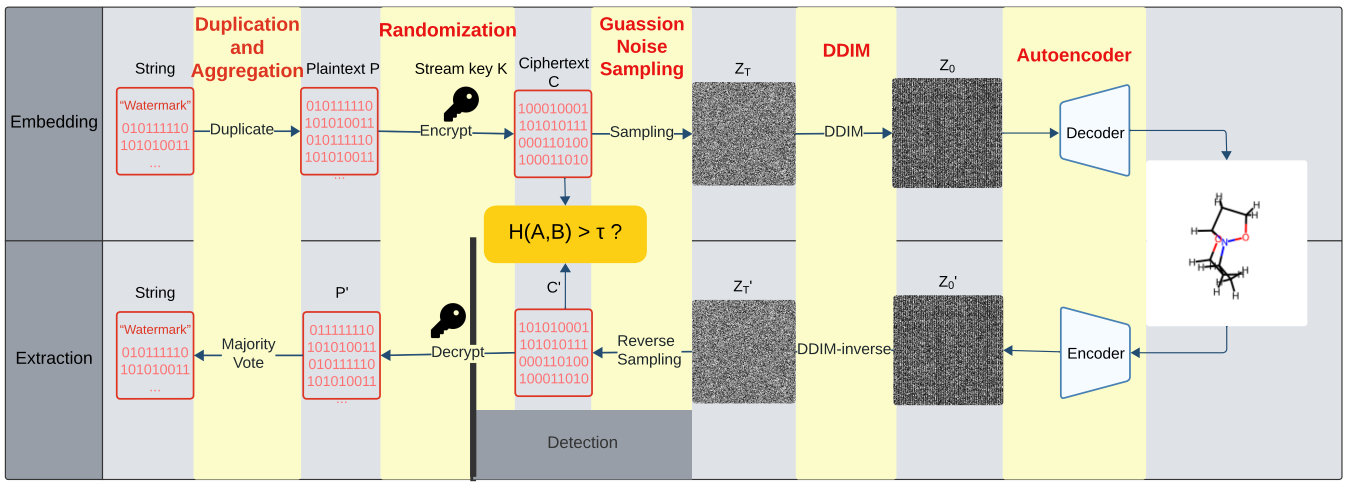
1 Introduction

Background In the expanding field of generative artificial intelligence, the integration of robust watermarking technologies is essential to protect intellectual property and maintain content authenticity.

Research Gap Watermarking techniques have been developed primarily for rich information media such as images [2] and audio [1]. However, these methods have not been adequately adapted for graph-based data, particularly on molecular graphs.

Research Question How can we develop a watermarking method for the graphs generated by diffusion models?

2 Methodology



The watermark is duplicated and encrypted to generate a random bitstream, then we use Gaussian noise sampling to generate the latent, DDIM-sampling, and decode it to create a watermarked molecule. The watermark is extracted by reversing these operations and detected by comparing Hamming distances between bitstreams

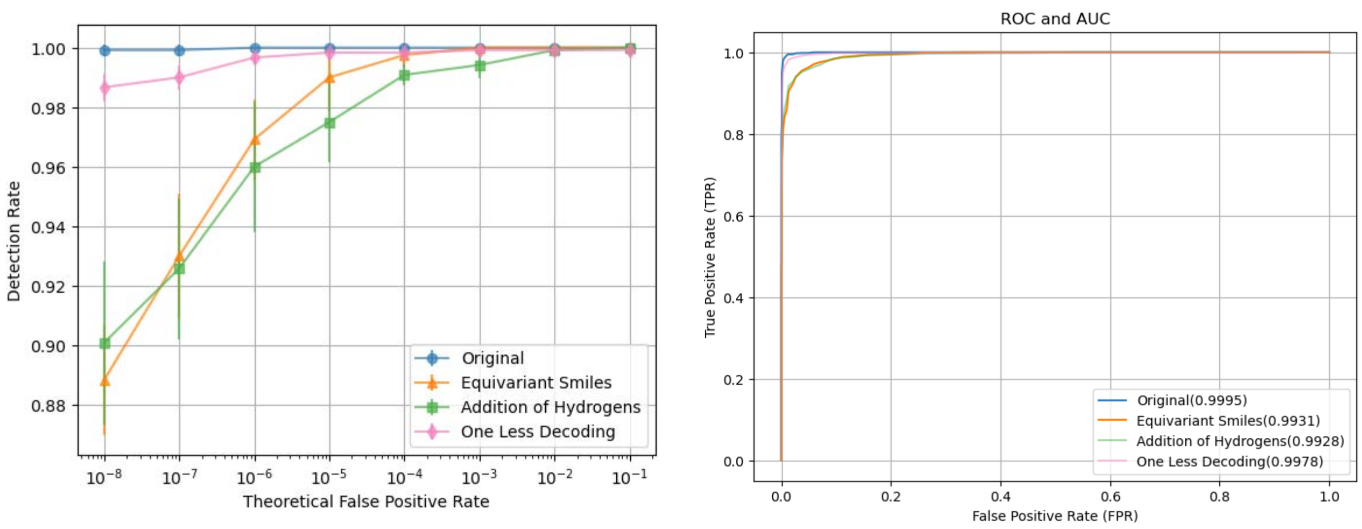
3 Results

Watermarked Model Performance

Methods	QM9				GEOM-drugs			
	Valid ↑	Valid&Uni ↑	AtomSta ↑	MolSta ↑	Valid ↑	Valid&Uni ↑	AtomSta ↑	MolSta ↑
Original	1.00(0)	97.83(0.04)	94.5(0.20)	81.01(0.30)	1.00(0)	99.99(0)	79.75(0.11)	4.21(0.27)
Watermarked	1.00(0)	98.10(0.22)	94.41(0.27)	80.84(0.24)	1.00(0)	1.00(0)	79.64(0.11)	4.23(0.40)
t-statistic	-	2.09	0.46	0.77	-	-	1.23	0.07

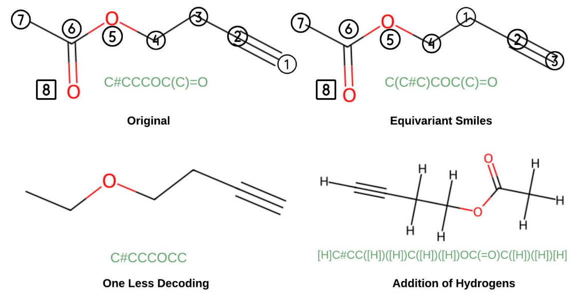
We benchmark molecules generated under two conditions: with and without the "Watermark" string embedded. The watermarked molecules maintain statistical parity in performance metrics compared to the original.

Watermark Robustness



Despite the detection rate being negatively affected due to the attacks, the watermark detection rates remained distinctly higher compared to the control group(not watermarked molecules)

4 Attack Methods



Conclusion

Research Question Answer We adapt Gaussian Shading [3], originally designed for image diffusion models, to graph diffusion models.

Limitations 1. The uniqueness of watermarked molecules depends on the uniqueness of key and nonce. 2. The watermarking process is relatively slow and bottlenecked by encryption speed.

Future Work We plan to develop more sophisticated and realistic attack methodologies that are tailored to different domains of graph structures.

Source code

References

[1] X. Cao, X. Li, D. Jadav, Y. Wu, Z. Chen, C. Zeng, and W. Wei. "Invisible watermarking for audio generation diffusion models". In: *arXiv preprint arXiv:2309.13166* (2023).

[2] Y. Wen, J. Kirchenbauer, J. Geiping, and T. Goldstein. "Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust". In: *arXiv preprint arXiv:2305.20030* (2023).

[3] Z. Yang, K. Zeng, K. Chen, H. Fang, W. Zhang, and N. Yu. "Gaussian Shading: Provable Performance-Lossless Image Watermarking for Diffusion Models". In: *arXiv preprint arXiv:2404.04956* (2024).

Delft University of Technology