Towards Interpretable Generative AI via Search and Knowledge Graphs

Theodoros Aivalis[a, b,*]

[a]University of Glasgow, UK
[b]National Centre for Scientific Research “Demokritos”, Greece
ORCID (Theodoros Aivalis): https://orcid.org/0009-0005-4452-9402

Abstract. As generative AI models become increasingly influential across creative and scientific domains, concerns about transparency and data attribution have grown significantly. Our research aims to develop a model-agnostic, interpretable framework for estimating the influence of training data on generated outputs, without requiring access to internal model parameters. Initial experiments used text and embedding similarity, while current work introduces structured knowledge graphs (KGs) to provide concept-level explanations. We evaluate these methods in both local and black-box settings. Looking ahead, we will further enhance reasoning through user and expert studies, and explore disentangled representations as an interpretable alternative to dense embeddings. The overarching goal is to build practical tools that increase transparency in generative AI and support applications in copyright and responsible deployment.

1 Introduction and Research Objectives

The rapid advancement of generative AI has led to widespread adoption across creative and scientific domains, but it has also intensified concerns about transparency, ownership, and accountability. These models are often treated as black-boxes, offering limited visibility into how specific training samples influence the generated outputs. This raises critical issues in contexts where attribution matters, especially in domains involving copyrighted content such as art.

At the heart of these concerns lies a fundamental ethical and legal question: who should receive credit or ownership when a model generates content influenced by protected works? Artists and content creators are increasingly speaking out against the use of their work in training AI models without their permission. Lawsuits, such as those by the New York Times[1] and Getty Images[2], highlight the growing conflict between AI development and intellectual property rights. In response, regulatory frameworks such as the EU’s AI Act [3] and UNESCO’s ethical guidelines [8] are beginning to outline principles for responsible AI development that respect ownership and transparency. Despite these developments, most generative models, especially the large-scale and commercially deployed, remain opaque, making it difficult for creators to assess if and how their work has influenced outputs. Existing influence estimation techniques, such as influence functions or retraining-based attribution [4, 5, 7], require access to internal model parameters and gradients, which makes them impractical for black-box scenarios. Moreover, these approaches often fail to provide interpretable, human-readable justifications of how and why particular training samples affect the outputs.

In this work, we investigate the development of a general method for analysing the influence of training data that can be applied across different generative models, including those treated as black boxes. The approach is model-agnostic and data-centric, aiming to uncover which training samples are most likely to have influenced a given output, without requiring access to model internals. The central idea is to estimate the contribution of specific training samples to a given output by leveraging techniques from information retrieval, semantic similarity, and structured representations. The goal is to make generative models more interpretable, providing creators, researchers, and regulators with clearer insights into how outputs are shaped by training data, and supporting more ethical use of AI systems.

2 Overview of the Proposed Framework

The central goal of our research is to develop a general, interpretable framework for analysing how training data influences outputs from generative models. This framework is designed to be model-agnostic and applicable even in black-box settings, meaning it does not rely on access to internal parameters, training pipelines, or gradients. Instead, it focuses on external observations: the user’s prompt, the generated output, and the accessible training data.

Figure 1 illustrates the central idea behind the proposed framework. Given a generated image produced from a user’s prompt, without requiring access to the internals of the model, the aim is to trace potential influences from the training dataset. Assuming access to training data, the framework compares the generated output to candidate samples in order to identify those most likely to have shaped the result. The core research challenge lies in designing an effective and interpretable comparison strategy.

Our work explores different strategies for this step. An initial approach, presented in earlier work [2], combines text-based retrieval with image embeddings to identify influences. This paper highlighted that removing the most similar training samples led to noticeable changes in the output, validating their influence and supporting a model-agnostic pipeline for copyright analysis. More recently, we have introduced a structured comparison method based on KGs, which enables semantic-level matching and interpretability. In both cases, the outcome is an influence report that lists the training samples that may have contributed to the generated image in decreasing probability of influence. Importantly, we also consider cases where

[*] Email: teoaivalis@iit.demokritos.gr

[1] https://harvardlawreview.org/blog/2024/04/nyt-v-openai-the-timess-about-face/ as viewed July 2025.

[2] https://www.theverge.com/2023/1/17/23558516/ai-art-copyright-stable-diffusion-getty-images-lawsuit as viewed July 2025.

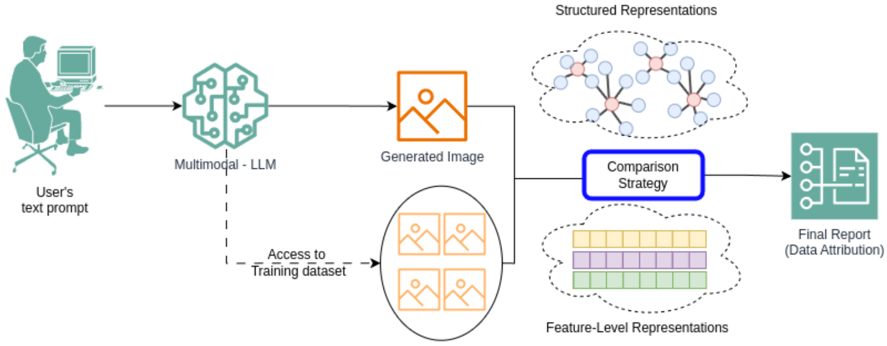
Structured Representations
User's text prompt
Multimodal - LLM
Generated Image
Comparison Strategy
Final Report (Data Attribution)
Access to Training dataset
Feature-Level Representations

Figure 1: Overview of the proposed framework. Given a user prompt, a model produces an image. The system compares the output with the training samples using structured or feature-based representations, generating a report on likely influences to support copyright auditing.

the original training dataset is not directly accessible. In such scenarios, we either use web-scale retrieval to approximate the original training data or construct domain-specific knowledge bases to enable influence estimation and semantic comparison.

3 Initial Steps

In the early stages of our research, we have developed a model-agnostic framework for identifying the influence of training data on the outputs of generative models [2]. This approach was grounded in two principles: (1) avoiding any reliance on model internals, and (2) combining textual and visual similarity to locate the most relevant training examples for a given output.

Specifically, we have a two-step pipeline. The first step performs text-based retrieval using TF-IDF and cosine similarity between the user's prompt and the training data descriptions. The second step refines the candidate set by evaluating similarity between the generated image and the retrieved samples, using raw pixel comparisons and embeddings, e.g. as derived from ResNet50. We define the most influential images as those that scored highest in this ranking.

Focusing on a fashion application scenario, we validated our framework by conducting unlearning experiments using a locally trained model, based on the DALL-E[3] and a fashion dataset of 44K items [1]. The source code for the experiments is available in the GitHub repository[4]. The results show that generated images changed significantly when the selected training samples were removed, confirming that the retrieval-based method identified true contributors. We have also extended the approach to a black-box setting by using Stable Diffusion generations and retrieving web images as a proxy for the unknown training set. The framework consistently surfaced highly similar retrieved images, indicating promising results in this setting. An initial demo for large-scale, black-box scenarios is available on Hugging Face[5].

4 Current Work

While the initial approach provides promising results, it lacks human-understandable explanations. To address this limitation, we are currently working on enhancing influence estimation by incorporating structured representations. In particular, we choose to work with KGs as a way to semantically describe the content of each image. We have developed a framework that automatically extracts structured representations from images using multimodal-LLMs, guided by a domain-specific ontology[6]. The system generates semantic triples that are stored in Neo4j[7] and used to construct interpretable graphs. These graphs enable meaningful comparisons between generated and training images by focusing on specific conceptual axes, such as object types, visual attributes, materials, or stylistic motifs that may have influenced the generation process.

This KG-based approach is being evaluated in both local and black-box settings. In the local case, we applied the method to identify influential training samples and validated the results through unlearning experiments. The results are comparable to those from embedding-based retrieval, with the added advantage of more interpretable explanations. In the black-box case, we are focusing on style-specific generations by collecting reference images that capture a target style. We extract KGs from the reference and generated images to compare their semantic content and identify overlaps. This enables a transparent analysis of stylistic influence. Although our experiments focus on fashion and style-specific generations, the method is domain-agnostic. It can be adapted to other fields by using embeddings from domain-relevant models and adopting an appropriate ontology.

5 Future Directions

In the next phase of our research, we plan to focus on KG-based influence analysis, as it shows strong potential to provide human-understandable explanations. The goal is to evaluate how effectively these representations help users interpret the generation process. To this end, we intend to conduct user studies that assess the usefulness of the influence reports. In parallel, we will engage with domain experts to assess the quality of the produced graphs. This will help shape a more general framework that supports both the KG community and practitioners seeking interpretable generative models.

A second direction involves exploring disentangled representations as a promising solution. Unlike dense embeddings, disentangled features aim to separate meaningful generative factors, such as colour, shape, or style, which are easier for users to understand and reason about [6, 9, 10]. In our research, we plan to use KGs as a structured way to support the extraction and organisation of these factors from images. This strategy will offer a practical and explainable alternative to black-box embeddings, supporting the analysis of what aspects of training data contribute to specific components of a generated image.

Acknowledgment: This work was supported by the Intelligent Information Systems division of the Institute of Informatics and Telecommunications at National Scientific Research Centre "Demokritos".

[3] https://github.com/lucidrains/DALLE-pytorch, as viewed July 2025.

[4] https://github.com/teoaivalis/Search-Based_Data_Influence_Analysis as of July 2025.

[5] https://huggingface.co/spaces/teoaivalis/InfluenceAnalyzerDemo as of July 2025.

[6] https://fashionpedia.github.io/home/, as viewed July 2025.

[7] https://neo4j.com/, as viewed July 2025.

References

[1] P. Aggarwal. Fashion product images dataset, 2019. URL https://www.kaggle.com/ds/139630.

[2] T. Aivalis, I. A. Klampanos, A. Troumpoukis, and J. M. Jose. Enhancing interpretability in generative AI through search-based data influence analysis. In *ICLR 2025 Workshop on Navigating and Addressing Data Problems for Foundation Models*, 2025. URL https://openreview.net/forum?id=2nbFLVTHcF.

[3] European Parliament and Council. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), 2024. URL https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng. Official Journal of the European Union, L 1689, 12.7.2024, p. 1–60.

[4] Z. Hammoudeh and D. Lowd. Training data influence analysis and estimation: a survey. *Machine Learning*, 113(5):2351–2403, Mar. 2024. ISSN 1573-0565. doi: 10.1007/s10994-023-06495-7. URL http://dx.doi.org/10.1007/s10994-023-06495-7.

[5] P. W. Koh and P. Liang. Understanding black-box predictions via influence functions, 2020. URL https://arxiv.org/abs/1703.04730.

[6] X. Ren, T. Yang, Y. Wang, and W. Zeng. Learning disentangled representation by exploiting pretrained generative models: A contrastive learning view, 2022. URL https://arxiv.org/abs/2102.10543.

[7] M. Sundararajan, K. Dhamdhere, and A. Agarwal. The shapley taylor interaction index. In H. D. III and A. Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9259–9268. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/sundararajan20a.html.

[8] UNESCO. Recommendation on the Ethics of Artificial Intelligence, 2021. URL https://unesdoc.unesco.org/ark:/48223/pf0000380455. Document code: SHS/BIO/REC-AIETHICS/2021, 21 pages.

[9] X. Wang, H. Chen, S. Tang, Z. Wu, and W. Zhu. Disentangled representation learning, 2024. URL https://arxiv.org/abs/2211.11695.

[10] Q. Wu, Y. Liu, H. Zhao, A. Kale, T. Bui, T. Yu, Z. Lin, Y. Zhang, and S. Chang. Uncovering the disentanglement capability in text-to-image diffusion models, 2022. URL https://arxiv.org/abs/2212.08698.