FrED: External Data Influence Estimation via Domain Knowledge Graph Grounding

Theodoros Aivalis*
National Centre for Scientific Research "Demokritos", Greece
University of Glasgow, UK

Iraklis A. Klampanos
University of Glasgow
UK

Antonis Troumpoukis
National Centre for Scientific Research "Demokritos"
Greece

Joemon M. Jose
University of Glasgow
UK

Abstract

The rapid deployment of generative AI has amplified the critical need for Training Data Attribution to ensure transparency and accountability. However, current parametric approaches require computationally prohibitive access to model weights, while similarity-based methods ignore deep structural context. We propose a novel probabilistic framework that operates entirely in a black-box setting. Our method fuses continuous feature similarities with discrete, domain-specific Knowledge Graphs (KGs). This approach ensures the attribution is grounded in structural reality, explicitly rewarding highly specific historical samples while preventing generic background data from dominating the results. We evaluate our framework across two distinct domains where linking outputs to data and domain context is inherently complex: abstract artistic image synthesis and high-dimensional physical weather forecasting. Extensive benchmarking demonstrates the robust efficacy of our approach. In the artistic domain, it achieves a strong Linear Datamodeling Score that exceeds standard black-box similarity baselines, while closing much of the gap to gradient-based estimators. We additionally present a cross-domain feasibility case study in environmental forecasting, where we use domain KGs to retrieve physically consistent historical analogs for regional flood forecasts, improving geographic localisation over a latent-only baseline. Operating entirely without internal model access, our approach provides an efficient, interpretable mechanism for post-hoc influence analysis and domain-grounded retrieval.

1 Introduction

In domains of high societal impact, like medicine and Earth observation (EO), researchers want to take advantage of AI's ability to uncover hidden patterns in data to solve crucial problems, like designing novel proteins for targeted therapies or generating high-resolution forecasts grounded in regional geographical metadata. At the same time, in sensitive and creative domains, this technology has intensified the debate over explainability and intellectual property. Because most state-of-the-art models operate as "black boxes,"[1] there is no visibility into the relationship between training samples, domain-specific knowledge and model outputs. Understanding how these systems synthesise content is critical for ensuring that generated artifacts can be reliably traced back to the training data.

*Corresponding Author. teoaivalis@iit.demokritos.gr

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

For creative arts, this need for transparency is highlighted by a growing conflict between AI development and copyright law. High-profile lawsuits involving the New York Times[2] and Disney[3] underscore these risks. Beyond legal disputes, over 200 artists recently signed an open letter [4] urging protection against AI-driven stylistic and voice mimicking without consent, highlighting the need for royalty frameworks. In response, policy bodies have proposed new governance frameworks. The European Union's AI Act [2] outlines rules for AI transparency, while the UK's AI Strategy [3] and UNESCO's guidelines [4] advocate for responsible AI. Notably, Denmark recently proposed legal amendments to grant individuals legal rights over their voice, face, and bodily likeness [5].

Attribution is also vital in EO for global safety. The World Meteorological Organisation's 2025 report [5] underscores that "cascading and compounding" disasters, sequential events impacting agrifood systems and migration, now affect every global region. To process the massive data from programs like the EU's Copernicus [6], meteorological agencies adopt AI-driven models such as Google GraphCast [7] and Huawei's Pangu-Weather [8]. Despite their efficiency, they remain 'black boxes,' making it difficult for decision-makers to distinguish physical precedents from statistical hallucinations. Current interpretability focus on local feature importance, yet mechanistic interpretability requires uncovering the physical relationships and interactions driving predictions. As models scale, traditional explainability tools struggle with complex data. Addressing this remains a challenge: we must move beyond post-hoc visualisations toward methods that translate mathematical outputs into historical and structural contexts, ensuring AI-driven forecasts are faithful to the fundamental physical laws.

To address these limitations, we propose FrED, a probabilistic framework for Training Data Attribution (TDA) in a strictly "black-box" setting. Unlike methods requiring architectural access, we estimate influence by relying solely on the relationship between model's outputs and the training data. Specifically, we bridge the gap between mathematical embeddings and structural understanding by grounding training data and model's outputs in domain-specific Knowledge Graphs (KGs). This allows us to combine the strengths of surface-level similarity with a networked understanding of the domain's logic. By fusing high-dimensional latent features with structured conceptual motifs, we move beyond visual correlations toward a more precise attribution method. This ensures that outputs are traced back to their origins through a combination of semantic closeness and structural relevance, providing a transparent audit trail that is faithful to both the data's content and its domain context.

Our core contributions are the following:

1. We propose FrED, a black-box influence estimator that fuses continuous latent similarity with discrete KG evidence via an asymmetric boosting.
2. We achieve LDS performance on ArtBench that significantly outperforms black-box baselines and approaches gradient-based methods under the D-TRAK protocol.
3. We demonstrate FrED's cross-domain utility in environmental forecasting by retrieving physically consistent historical analogs for flood forecasts.

Paper Structure: Section 2 provides a review of related work in data attribution. Section 3 details our Bayesian framework. Section 4 presents the experimental setup and evaluation across different domains. Sections 5 and 6 summarise the main findings and future research directions. [6].

2 Related Work

We categorise current TDA research into: foundational theoretical models, state-of-the-art parametric methods, and the parameter-free observational approaches that are most relevant to our methodology.

Theoretical methods serve as the "gold standards" for the field [9]. Leave-One-Out quantifies influence by retraining models without a specific sample. While being an exact measure of data impact, it

[2] `https://harvardlawreview.org/blog/2024/04/nyt-v-openai-the-timess-about-face/`, as viewed November 2026.

[3] `https://digiday.com/marketing/the-disney-openai-deal-and-generative-ai-copyright-concerns/`, as viewed November 2026.

[4] `https://www.bbc.com/news/articles/c071elp1rv1o`, as viewed November 2026.

[5] `https://www.theguardian.com/technology/2025/jun/27/deepfakes-denmark-copyright-law-artificial-intelligence`, as viewed November 2026.

[6] Source code and implementation details: `https://github.com/teoaivalis/FrED`

operates at an impossible computational cost for modern foundation models [10]. Similarly, game-theoretic Shapley Values provide a principled framework but suffer from exponential complexity at scale [11]. Local Interpretable Model-agnostic Explanations (LIME) [12] approximates models locally via interpretable surrogates, but prioritises feature importance over tracing statistical lineage to training samples. To avoid full retraining, Influence Functions utilise second-order derivatives to estimate the impact of up-weighting specific samples [9]. While foundational, these methods' high costs have driven a shift toward methodologies that estimate influence without exhaustive retraining. Modern parametric methods leverage internal signals, like gradients and weights, to estimate influence at scale. In Transformers, attention maps visualise specific regions or tokens [13, 14]. Furthermore, Attribution by Unlearning [15] approximates sample impact by directly optimising model weights to "forget" specific images and tracking resulting training loss deviations. TRAK [16] and its diffusion variant D-TRAK [17] enable this by using random projections to make gradient-based attribution computationally tractable. Moving beyond, the Diffusion Attribution Score (DAS) [18] quantifies a sample's contribution by measuring the divergence between predicted noise distributions. The final category includes methods that operate without access to model internals. At the most fundamental level, naive techniques include raw pixel similarity and high-dimensional visualization techniques like t-SNE [19], which project latent representations into low-dimensional maps to observe qualitative semantic clustering. Latent embedding similarity has emerged as the standard black-box baseline, using pre-trained encoders to measure semantic closeness. Extending these, a recent search-based influence analysis framework [20] combines textual retrieval with high-dimensional embeddings to link outputs to training samples. Notably, Fast Data Attribution [21] distill influence scores from a computationally "slow" teacher model into a "fast" student multi-layer perceptron (MLP). Finally, the inherent difficulty of evaluation has led to benchmarking frameworks such as Attribution by Customization [22], that fine-tunes generators toward a specific exemplar and creates "ground-truth" influenced images, providing a controlled environment to evaluate attribution methods.

Observational baselines identify surface-level overlap, but they lack the structural context required for semantic attribution. Recent work instead leverages Knowledge Graphs (KGs) to capture domain characteristics, enabling interpretable, concept-level comparisons. For example, KGs have been employed to trace stylistic influences by comparing the structural representations of generated and training images [23, 24]. Demonstrating the cross-domain adaptability, KGa are used in EO to fuse satellites and reports, enabling structural comparisons and historical retrieval of climate events [25].

3 Proposed Method

3.1 A Unified Modeling Paradigm

We establish a unified modeling paradigm for FrED, abstracting deep generative and predictive architectures as latent space transformations. While we focus on diffusion-based synthesis and weather forecasting, this formulation generalises across representation learning frameworks. Let $\mathcal{S} = \{(\mathbf{X}_j, \mathbf{T}_j)\}_{j=1}^{K}$ be the training dataset, where $\mathbf{X}_j$ is the high-dimensional observation (pixels or atmospheric tensors) and $\mathbf{T}_j$ is the associated conditioning context (text or weather states).

We decompose the model G into two fundamental components [26]:

1. **The Encoder (E):** Maps input pairs into a compressed latent space $\mathcal{Z}$, such that $E(\mathbf{X}, \mathbf{T}) = \mathbf{z}$. This space encodes the essential statistical features of the training distribution.
2. **The Decoder (D):** Projects the latent vector back into the data space to produce a reconstruction or prediction, $\mathbf{X}' = D(\mathbf{z}, \mathbf{T})$.

Target $\mathbf{C}$ is produced by fixing the context and sampling from the latent space: $\mathbf{C} = D(\mathbf{z}_c, \mathbf{T})$. Since $\mathbf{C}$ is a product of the decoded state $\mathbf{z}_c$, we formalise attribution as tracing this state backward to identify the training samples $\mathbf{X}_j$ that most influenced the mapping $E \to D$ during optimisation.

3.2 The Bayesian Influence Framework

The Ideal Influence (Counterfactual) To ground the concept of attribution, we define the generative process as a sampling over the latent manifold $\mathcal{Z}$. The total probability of the model generating a

specific artifact $\mathbf{C}$ is expressed as:

$$P(\mathbf{C}) = \int_{\mathcal{Z}} P(\mathbf{C} \mid \mathbf{z}) P(\mathbf{z}) d\mathbf{z} \tag{1}$$

where $P(\mathbf{z})$ is the latent density learned from the training corpus $\mathcal{S}$ [27]. Ideally, the influence of a specific training sample $\mathbf{X}_j$ is defined counterfactually as the shift in the model's output probability for $\mathbf{C}$ that would occur if $\mathbf{X}_j$ were removed from $\mathcal{S}$. However, in a black-box setting, evaluating this shift directly is computationally prohibitive as it requires retraining for every training sample.

The Bayesian Proxy To address this, FrED approximates the counterfactual impact, approaching attribution as a posterior probability $P(\mathbf{X}_j \mid \mathbf{C})$. We assume that because the model's continuous representation is shaped by the entirety of its corpus, the generation of any target $\mathbf{C}$ can be decomposed into the discrete contributions of the training samples. For the purpose of attribution, we approximate the target as a weighted combination of the training set:

$$\mathbf{C} \approx \sum_{j=1}^{K} P(\mathbf{X}_j \mid \mathbf{C}) \mathbf{X}_j = \sum_{j=1}^{K} \alpha_j \mathbf{X}_j \tag{2}$$

We note that eq. 2 is a ranking-oriented surrogate used to derive a practical scoring rule in black-box settings, rather than a literal generative model of diffusion or forecasting outputs.

Using Bayes' theorem, we expand the attribution weight α_j to identify the primary drivers of the model's density at the target coordinates:

$$\alpha_j = \frac{P(\mathbf{C} \mid \mathbf{X}_j) P(\mathbf{X}_j)}{P(\mathbf{C})} \tag{3}$$

For a fixed target $\mathbf{C}$, the denominator $P(\mathbf{C})$ acts as a constant normalization factor. Thus, the ranking of attribution weights is determined by the product of the likelihood and the prior:

$$\alpha_j \propto P(\mathbf{C} \mid \mathbf{X}_j) P(\mathbf{X}_j) \tag{4}$$

Likelihood $P(\mathbf{C} \mid \mathbf{X}_j)$ (Latent Support): This term serves as a proxy for how strongly X_j supports the region of latent space from which $\mathbf{C}$ is generated. We hypothesise that if $\mathbf{X}_j$ provides the structural instruction for a specific region of the space, any target $\mathbf{C}$ in that region is supported by $\mathbf{X}_j$. In a black-box setting, we model this via latent proximity: if the representations $\mathbf{z}_c$ and $\mathbf{z}_j$ are proximal in $\mathcal{Z}$, we assume $\mathbf{C}$ inherits its specific identity from the features of $\mathbf{X}_j$.

Prior $P(\mathbf{X}_j)$ (Informational Scarcity): The prior represents the informational value of $\mathbf{X}_j$ independent of the target. Moving beyond a naive uniform prior, we define this as informational scarcity. By penalising samples with dense, redundant regions in the latent space, we ensure that specialised, unique data points receive higher weight. This prevents attribution from being diluted by generic states and highlights the samples necessary to reconstruct the identity of $\mathbf{C}$.

3.3 A Dual-Space Attribution Paradigm

Computing α_j solely via the model's latent space has limitations: deep learning architectures optimise for continuous feature similarity but remain fundamentally blind to discrete, real-world semantics. Training typically relies on sparse labels that fail to capture the rich, interconnected domain knowledge inherent in historical sample. However, for domain experts, this broader structural and contextual metadata is the primary metric for evaluating true influence and establishing trust in the model's outputs. To bridge machine optimization and domain context, FrED implements a retrieval-inspired fusion of two independent attribution rankings, each evaluating α_j within a distinct feature space.

Latent Space Attribution. Here we evaluate candidates within a high-dimensional continuous space, mirroring the encoding strategies used by the models themselves. By projecting both $\mathbf{X}_j$ and $\mathbf{C}$ into this space, we capture the continuous distributional features leveraged during the model's training phase. Within this engine, the Likelihood (Feature Similarity) measures the direct mathematical distance (vector similarity) between the training sample and the target output. Conversely, the Prior (Feature Distinctiveness) evaluates the global density of the sample within the training distribution, penalising highly redundant data to boost candidates possessing distinctive motifs.

Discrete Domain Ranking. Simultaneously, we evaluate candidates based on semantic context, which raw latent features ignore. Because real-world domain knowledge is inherently sparse, categorical, and relational, continuous embeddings fail to accurately represent it. Therefore, we can model the Domain Space as a KG, organising this discrete metadata into a verifiable networked topology. Within this KG, the Likelihood measures topological similarity between the training sample and the target artifact. The Prior evaluates structural sparsity, boosting candidates linked to highly specific, non-redundant categorical connections rather than generic, heavily crowded hubs.

This dual-mapping ensures that the final attribution reflects what the model saw (the training distribution), and also the structural reality in which that data exists (the domain context).

3.4 Contextual Boosting for Dual-Space Fusion

To combine evidence from both spaces, we treat attribution as a hierarchical retrieval system. Standard rank aggregation methods (e.g. Reciprocal Rank Fusion (RRF) [28]) treat data sources as equal contributors. However, because generative and predictive models are optimised primarily on continuous latent features, treating discrete metadata with equal contribution can degrade the primary signal. Moreover, this approach provides robustness against incomplete or non-representative domain knowledge. If a sample lacks connections in the KG, an equal-weighting fusion would penalise it. By structuring the domain evidence as an asymmetric boost, system uses foundational latent score when metadata is absent, ensuring that attributions are never discarded due to missing graph topology.

Inspired from retrieval systems, we use the Latent score as primary evidence and the Domain rank as a contextual guide. As Figure 1 shows, the final weight α_j for a training sample $\mathbf{X}_j$ is calculated as:

$$\alpha_j = \mathcal{S}_{lat,j} \times (1.0 + \text{Boost}(\mathcal{R}_{dom,j})) \tag{5}$$

where the boosting term uses the rank-decay principles of RRF, applied to the Domain Engine:

$$\text{Boost}(\mathcal{R}_{dom,j}) = \frac{W}{k + \mathcal{R}_{dom,j}} \tag{6}$$

Here, $\mathcal{S}_{lat,j}$ is the Latent Engine's base probability (combining both the latent similarity and the distinctiveness prior). $\mathcal{R}_{dom,j}$ is the discrete integer rank of the sample in the Domain Engine. W defines the maximum scaling weight, and k is a constant to control how fast the boost drops off. In all experiments we set $k = 5$ and $W = 1.0$. A sensitivity analysis of k is provided in Table 8.

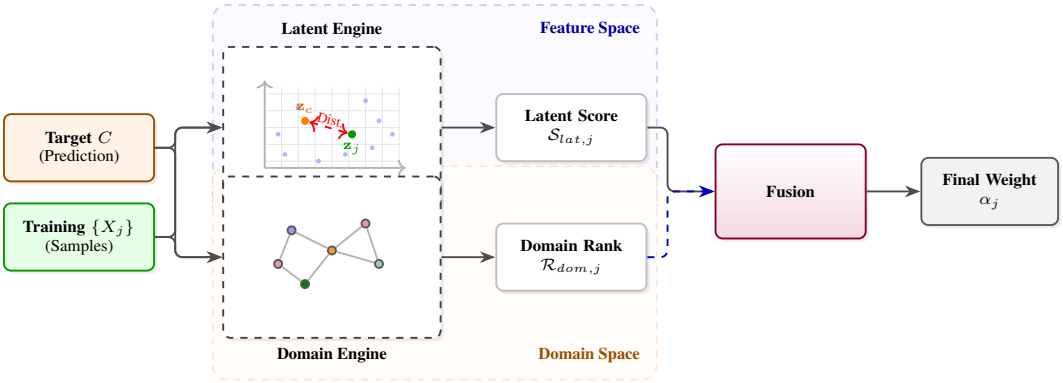


Figure 1: **FrED**'s overview showing the fusion between latent and domain spaces. A hierarchical mechanism applies the discrete domain rank as a contextual boost to the continuous latent score.

4 Experiments

We evaluate **FrED** across two domains with distinct attribution challenges and evaluation paradigms. In Art Generation (Sec. 4.1), we benchmark against parametric methods on ArtBench to trace stylistic

lineages. In Environmental Forecasting (Sec. 4.2), prohibitive retraining costs for global models prevent standard counterfactual evaluation. We instead present a feasibility study for analog retrieval, demonstrating our versatility in aligning high-dimensional physical states with structural knowledge.

4.1 The Artistic Domain

4.1.1 Datasets and Domain Modeling

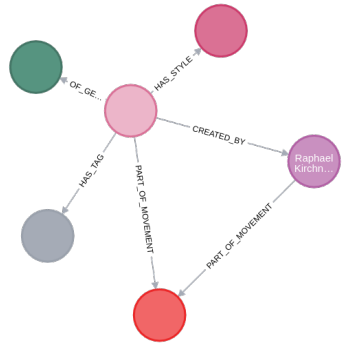
We utilise the ArtBench-10 dataset [29], a standard benchmark for generative modeling consisting of class-balanced artworks categorised into ten artistic styles. While the raw dataset provides the visual modality and basic labels, it lacks the deep historical context necessary for structural attribution. To bridge this gap and establish our domain knowledge, we curate a comprehensive metadata registry for every painting. Using an automated extraction pipeline, we query WikiArt, Wikipedia, and DBpedia to collect historical and semantic information across two primary levels:

1. **Painting-Level Metadata:** Specific categorical details for each artwork, mapping the image to its explicit genre, style, artistic movement and descriptive tags.
2. **Artist-Level Metadata:** Rich biographical and historical data capturing the artist's active years, geographical associations, academic institutions, and complex interpersonal lineages (e.g., who influenced them, who they taught, and their professional collaborators).

By aggregating these heterogeneous data, we construct the Domain KG ($\mathcal{G}_D$), that encodes the historical interconnections between the training samples. The overall domain schema is detailed in Appendix A. Figure 2 presents a visual example paired with its corresponding extracted subgraph.



(a) Visual Exemplar: Scenes of Ancient Greece



(b) Extracted KG for the Exemplar

Figure 2: Artistic exemplar: (a) visual data and (b) its extracted KG, demonstrating the mapping of visual features to topological nodes.

Domain Alignment via Contrastive Learning. A challenge in structural attribution is evaluating generated images ($\mathbf{C}$) against the domain KG. These images lack the metadata required to position them within our domain. Similarly, many training samples ($\mathbf{X}_j$) have missing records. To solve this, we create a mapping that translates visual features into graph coordinates. We curate a representative subset of training images with rich metadata, using them as ground-truth anchors. We then adopt a contrastive learning paradigm inspired by CLIP [30], but instead of aligning images with text, we optimise the alignment between visual embeddings and their graph representations. For methodological consistency with our latent engine, we utilise a pre-trained ViT-g-14 [31] and apply node2vec [32] in the graph modality. By optimising a symmetric InfoNCE loss, we force visual and structural features to align in a shared representation space. Once this space is established, we project any target image in the domain space, enabling us to compute the structural ranking ($\mathcal{R}_{dom,j}$), providing the necessary foundation for our method. Appendix B details data selected, the encoders and training procedure.

4.1.2 Generative Setup and Evaluation Protocol

To ensure direct comparison with state-of-the-art parametric baselines, we follow the experimental protocol established by D-TRAK [17] for ArtBench. We use the same generative pipeline, with

identical model architectures, hyperparameter configurations, and training schedules. We represent our attribution engines using a pre-trained ViT-g-14 for latent features and Node2Vec for structural domain topology. Crucially, we evaluate our framework using the exact same set of 1,000 generated target artifacts ($\mathbf{C}$) synthesised for the baseline's evaluation. By maintaining consistency across training subsets and outputs, we isolate attribution performance as the sole independent variable, ensuring gains reflect our methodology rather than variations in the underlying generative model.

Generative Setup: Stable Diffusion and LoRA. Following the established benchmark, we fine-tune a 256×256 Stable Diffusion model [33] using LoRA [34] ($r = 128$, 25.5M parameters). Training is class-conditional, utilising prompts such as "a {class} painting". To ensure comparison with baselines, we replicate the D-TRAK training and synthesis pipeline. Specifics regarding optimisation schedules, data augmentation, and DDIM solver settings are detailed in Appendix C.

Evaluation Metric: The Linear Datamodeling Score. Evaluating attribution requires the predicted scores accurately reflect actual causal dynamics during the generative process. We validate FrED via the Linear Datamodeling Score (LDS), introduced by TRAK [16] and adapted for diffusion models by D-TRAK [17] and DAS [18]. The premise of LDS is that an effective method must act as a reliable counterfactual predictor. If a framework identifies the important training points for a generation, modifying them should lead to a predictable change in the model's output. LDS calculates the Spearman rank correlation (ρ) between two distinct variables across an ensemble of subsets:

1. **Actual Behavior:** The ground-truth change in model output when the model is trained on random subsets of the training data. Following the benchmark, this comprises an ensemble of models trained across 64 distinct subsets, each containing 50% of the original data.
2. **Attribution Prediction:** The additive proxy, representing the sum of the predicted attribution scores assigned to the specific training samples included in those random subsets.

A high LDS (ρ) indicates a strong relationship between our framework's predicted scores and the actual behavior of the generative model under counterfactual conditions. The final reported LDS is the average score across all evaluated target artifacts. For completeness, we provide the extended mathematical formulation of the LDS ensemble procedure in Appendix D.

4.1.3 Quantitative Benchmarking and Results

Table 1 presents the LDS evaluation across both Validation (real images) and Generation (synthetic images) sets for the ArtBench-2 and ArtBench-5 splits. We categorise baselines in two groups based on their level of model access: Non-Parametric methods, and Parametric methods that require internal access. Detailed implementation specifics for these baseline methods are provided in Appendix E.

Table 1: LDS results (%) on ArtBench-2 and ArtBench-5. **FrED** significantly outperforms black-box baselines and closes the gap to gradient-based estimators without requiring internal model access.

	ArtBench-2				**ArtBench-5**			
	Validation		Generation		Validation		Generation	
Method	**10**	**100**	**10**	**100**	**10**	**100**	**10**	**100**
Non-Parametric methods								
Raw pixel (dot prod.)	2.4 ± 1.1		2.5 ± 1.5		1.84 ± 0.42		2.77 ± 0.80	
Raw pixel (cosine)	2.5 ± 1.2		2.6 ± 1.5		1.97 ± 0.41		3.22 ± 0.78	
CLIP similarity (dot prod.)	7.3 ± 1.5		5.3 ± 2.4		5.29 ± 0.45		4.47 ± 1.09	
CLIP similarity (cosine)	8.9 ± 1.3		8.8 ± 2.2		6.57 ± 0.44		6.63 ± 1.14	
FrED (Ours)	$\mathbf{29.1 \pm 0.65}$		$\mathbf{21.3 \pm 1.65}$		$\mathbf{28.9 \pm 0.35}$		$\mathbf{18.4 \pm 1.05}$	
Parametric methods								
Gradient (dot prod.) [35]	7.68 ± 0.43	16.00 ± 0.51	4.07 ± 1.07	10.23 ± 1.08	4.77 ± 0.36	10.02 ± 0.45	3.89 ± 0.88	8.17 ± 1.02
Gradient (cosine) [35]	7.72 ± 0.42	16.04 ± 0.49	4.50 ± 0.97	10.71 ± 1.07	4.96 ± 0.35	9.85 ± 0.44	4.14 ± 0.86	8.18 ± 1.01
TracInCP [36]	9.69 ± 0.49	17.83 ± 0.58	6.36 ± 0.93	13.85 ± 1.01	5.33 ± 0.37	10.87 ± 0.47	4.34 ± 0.84	9.02 ± 1.04
GAS [37]	9.65 ± 0.46	18.04 ± 0.62	6.74 ± 0.82	14.27 ± 0.97	5.52 ± 0.38	10.71 ± 0.48	4.48 ± 0.83	9.13 ± 1.01
Journey TRAK [38]	–	–	5.96 ± 0.97	11.41 ± 1.02	–	–	7.59 ± 0.78	13.31 ± 0.68
Relative IF [39]	12.22 ± 0.43	27.25 ± 0.34	7.62 ± 0.57	19.78 ± 0.69	9.77 ± 0.34	20.97 ± 0.41	8.89 ± 0.59	19.56 ± 0.62
Renorm. IF [37]	11.90 ± 0.43	26.49 ± 0.34	7.83 ± 0.64	19.86 ± 0.71	9.57 ± 0.32	20.72 ± 0.40	8.97 ± 0.58	19.38 ± 0.66
TRAK [16]	12.26 ± 0.42	27.28 ± 0.34	7.78 ± 0.59	20.02 ± 0.69	9.79 ± 0.33	21.03 ± 0.42	8.79 ± 0.59	19.54 ± 0.61
D-TRAK [17]	27.61 ± 0.49	32.38 ± 0.41	24.16 ± 0.67	26.53 ± 0.64	22.84 ± 0.37	27.46 ± 0.37	21.56 ± 0.71	23.85 ± 0.74
DAS [18]	$\mathbf{37.96 \pm 0.64}$	$\mathbf{40.77 \pm 0.47}$	$\mathbf{30.81 \pm 0.31}$	$\mathbf{32.31 \pm 0.42}$	$\mathbf{35.33 \pm 0.49}$	$\mathbf{37.67 \pm 0.68}$	$\mathbf{31.74 \pm 0.75}$	$\mathbf{32.77 \pm 0.53}$

Encoder Choice and Accuracy Boost. The high accuracy of FrED, as evidenced in Table 1, is based on the selection of ViT-g-14 [31]. As demonstrated in our ablation studies (Table 6), this model yields superior predictive faithfulness, justified by the scale and diversity of the LAION-2B

distribution. This high-diversity set allows the encoder to develop a generalised feature space that captures the fine-grained stylistic characteristics necessary for art attribution. Furthermore, we found that applying a non-linear scaling exponent, specifically $p(c|x_j)^{10}$, is critical for mathematical sharpening. Since influence is often concentrated in a sparse subset of training data, this mechanism amplifies causal signals while driving the attribution of non-influential points toward zero.

Engine Validation and Fusion. Our experiments reveal that both the Latent and Domain spaces possess significant individual attribution power (Table 7). However, the comparison of fusion techniques highlights the limitations of linear combinations (Hybrid 50/50), which often introduce categorical noise to the visual signals. Our Boost Technique treats the Domain Space as topological evidence rather than a parallel signal. By using the KG strictly to re-rank candidates from the Latent Space, we ensure the final attribution is both visually acceptable and historically consistent. This asymmetric synthesis allows our framework to bridge the performance gap with gradient-intensive parametric methods while maintaining the efficiency and accessibility of a black-box system.

4.2 Case Study: Domain-Grounded Retrieval of Historical Analogs for Flood Forecasts

Moving beyond art, we apply FrED scoring pipeline strictly for domain-grounded analog retrieval. Since counterfactual evaluation is currently impractical for weather models, our metrics assess regional fidelity and physical consistency, not training-data attribution, and do not imply evidence of causality, memorisation, or training-set membership.

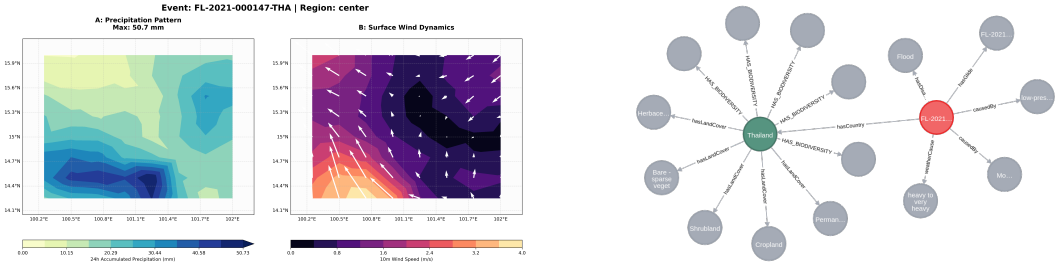
4.2.1 Data Collection and Environmental Mapping

We combine different data in an Environmental KG ($\mathcal{G}_E$), connecting weather patterns to species and ground conditions on the Earth. We utilise 1,372 flood events from [25]. Each record includes a unique ID, exact dates, and impact data (e.g., damage and displacement). Precise GPS coordinates link meteorological records to environmental proxies. To mitigate observation bias, we treat this domain knowledge strictly as a secondary re-ranking signal across two levels:

1. **The Nature Layer (Biological):** Utilising iNaturalist [40] sightings within a 10km radius, we capture regional ecological signatures to integrate biodiversity patterns into our KG. This enables cross-regional comparisons based on ecological habitats.
2. **The Land Cover Layer (Geographic):** Using ESA LandCover satellites [41], we analyse surface composition in a 10km radius. Integrating these geographic signatures, such as urban, agricultural, or water-absorbing zones, allows the framework to evaluate how ground conditions influence forecast severity and link regions with similar physical vulnerabilities.

Integrating these data into a unified graph architecture creates a dense network connecting meteorological events to their surroundings (details in Appendix G). Figure 3 demonstrates this by pairing a specific flood event with its extracted environmental subgraph.

Unlike the art study's contrastive inference, here we utilise Direct Geospatial Matching. Since forecasts provide precise coordinates and timestamps, we query the exact environmental state from the KG. This links current forecasts to historical analogs sharing identical environmental constraints.



(a) Meteorological Exemplar: FL-2021-000147 (b) Extracted KG for the Event

Figure 3: Environmental exemplar: (a) meteorological data and (b) its extracted KG, demonstrating the mapping of atmospheric variables to structural nodes.

4.2.2 Predictive Setup and Evaluation Targets

We prioritise regional meteorological dynamics over global predictions. While foundation models process data globally, effective analog retrieval grounding requires higher resolution to isolate the local extremes and geographic traits. Full-globe embeddings would encode data from unrelated regions, weakening local influence and preventing experts from tracing clear contextual links. Localised filtering instead ensures the framework captures the unique atmospheric signatures of each flood.

Predictive Baseline: Pangu-Weather and WeatherBench 2. We utilise Pangu-Weather [8] via WeatherBench2 [42] for 237 flood events (2018–2022), defining target ($\mathbf{C}$) as the 24-hour forecast at peak intensity. Unlike the 2D pixel grids of the artistic domain, retrieving historical precedents in weather forecasting requires evaluating complex atmospheric tensors. While Pangu operates globally at 0.25° resolution, we extract regional patches centered on event coordinates to capture local signatures. Within these patches, we subset the tensor to isolate primary flood drivers, including surface variables and upper-air profiles across four pressure levels. For representation, we train a 3D Autoencoder for latent features and employ Node2Vec for structural graph topology. Comprehensive details on architecture, variable selection, and training protocols are provided in Appendix H.

Evaluation Protocol: Regional Fidelity and Ecological Consistency. Unlike the artistic domain, where LDS is the gold standard, attribution for global weather foundation models remains unexplored. We evaluate our environmental case-study as domain-grounded analog retrieval, not validated training-data influence. Consequently, we assess the quality of our work through:

1. **Geographic Precision:** We quantify the percentage of top-K historical analogs that share the target forecast's country. This quantifies the model's ability to anchor continuous atmospheric states to specific geographic locations.
2. **Ecological Consistency:** For cross-country matches, we evaluate similarity using Landcover Overlap and Species Category Similarity. This validates whether we retrieve physically consistent analogs when exact geographic matches are unavailable.

4.2.3 Results and Physical Consistency

Table 2 shows that our method significantly improves regional localisation, reaching a Geographic Precision of 65.7% compared to 50.9% for the baseline, demonstrating that the KG situates abstract atmospheric data within correct physical locations. This shift moves from standard retrieval based on numerical similarity, toward a framework grounded in physical consistency. While precision decreases as K increases due to limited regional history, FrED consistently outperforms the baseline.

Table 2: **FrED** improves geographic precision by anchoring atmospheric states to regional contexts. Columns report Geographic Precision (%) at top-K ranks (K = 10, 20, 50, 100).

Retrieval Method	Prec@10	Prec@20	Prec@50	Prec@100
Latent Baseline	50.9%	35.6%	15.2%	6.8%
FrED (Ours)	**65.7%**	**47.8%**	**22.5%**	**12.0%**

Retrievals outside the target country maintain high ecological consistency, averaging **42.1% Landcover Overlap** and **46.2% Species Similarity**. A representative example is the linking of West African savanna profiles in Gambia–Burkina Faso. Similarly, recurring substitutions like South Sudan–Mali, Central African Republic–Sudan, and Paraguay–Chile demonstrate that FrED identifies physically consistent analogs by matching structural signatures across disparate geographic regions.

5 Conclusion

In this paper, we introduced FrED, a probabilistic framework for TDA in black-box settings. Unlike parametric methods, we estimate influence through the hierarchical fusion of continuous latent features and domain-specific KGs. Moving the attribution mechanism outside the model's architecture, we provide a scalable solution for auditing foundation models that are inaccessible due to computational constraints. We demonstrated the versatility of our methodology across two domains.

In the artistic domain, benchmarking on ArtBench revealed that we achieve an LDS competitive with several parametric estimators, while requiring zero internal access. In the environmental domain, we demonstrate feasibility via domain-grounded analog retrieval; establishing counterfactual causal training influence for such models remains future work. By prioritising regional characteristics, we retrieve physically relevant historical analogs for high-dimensional forecasts. Overall, our findings suggest that grounding latent similarity with domain context is a powerful mechanism for producing interpretable, domain-consistent influence hypotheses and historical analogs in black-box settings.

6 Limitations and Future Directions

Key limitations shape our future research directions. The framework's reliance on domain KGs may present an "expert bottleneck" in some fields, suggesting future collaboration to formalise domain knowledge. Additionally, as KGs provide static snapshots, we aim to implement real-time graph updates. Our evaluation is further constrained by the computational intensity of the LDS metric, suggesting that one might develop more efficient, domain-specific benchmarks or try rank-targeted fusion strategies to improve accuracy without full model retraining. Where original training sets are inaccessible, we can utilise representative KG samples or approximate distributions via targeted web-scraping, similarly to [20]. Finally, we will investigate applications in medicine and biology, where tracing sequences back to clinical datasets offers a vital path toward safe, AI-driven discovery.

Acknowledgments and Disclosure of Funding

This work was supported by the Intelligent Information Systems division of the Institute of Informatics and Telecommunications at the National Centre for Scientific Research "Demokritos". We also thank Nikolaos Manginas for his valuable feedback on the theoretical formulation of the proposed method.

References

[1] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pages 1885–1894. Pmlr, 2017.

[2] 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. Official Journal of the European Union, L 1689, 12.7.2024, p. 1–60.

[3] UK Government. UK National AI Strategy, 2021. Published by Department for Science, Innovation and Technology and Office for Artificial Intelligence and Department for Digital, Culture, Media and Sport and Department for Business, Energy and Industrial Strategy.

[4] UNESCO. Recommendation on the Ethics of Artificial Intelligence, 2021. Document code: SHS/BIO/REC-AIETHICS/2021, 21 pages.

[5] World Meteorological Organization. State of the global climate 2025. Technical Report WMO-No. 1391, World Meteorological Organization (WMO), Geneva, 2026.

[6] Jean-Noël Thépaut, Dick Dee, Richard Engelen, and Bernard Pinty. The copernicus programme and its climate change service. In *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 1591–1593, 2018.

[7] Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, et al. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023.

[8] Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, 619(7970):533–538, 2023.

[9] Zayd Hammoudeh and Daniel Lowd. Training data influence analysis and estimation: a survey. *Machine Learning*, 113(5):2351–2403, March 2024.

[10] Emily Black and Matt Fredrikson. Leave-one-out unfairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 285–295, 2021.

[11] Mukund Sundararajan, Kedar Dhamdhere, and Ashish Agarwal. The shapley taylor interaction index. In Hal Daumé III and Aarti 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.

[12] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[14] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 782–791. IEEE, 2021.

[15] Sheng-Yu Wang, Aaron Hertzmann, Alexei A Efros, Jun-Yan Zhu, and Richard Zhang. Data attribution for text-to-image models by unlearning synthesized images. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[16] Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Mądry. Trak: attributing model behavior at scale. In *Proceedings of the 40th International Conference on Machine Learning*, pages 27074–27113, 2023.

[17] Xiaosen Zheng, Tianyu Pang, Chao Du, Jing Jiang, and Min Lin. Intriguing properties of data attribution on diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024.

[18] Jinxu Lin, Linwei Tao, Minjing Dong, and Chang Xu. Diffusion attribution score: Evaluating training data influence in diffusion models. In *International Conference on Learning Representations*, volume 2025, pages 22962–22989, 2025.

[19] Laurens van der Maaten and Geoffrey E. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[20] Theodoros Aivalis, Iraklis A. Klampanos, Antonis Troumpoukis, and Joemon 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.

[21] Sheng-Yu Wang, Aaron Hertzmann, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Fast data attribution for text-to-image models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[22] Sheng-Yu Wang, Alexei A. Efros, Jun-Yan Zhu, and Richard Zhang. Evaluating data attribution for text-to-image models. In *ICCV*, 2023.

[23] Theodoros Aivalis, Iraklis A Klampanos, Antonis Troumpoukis, and Joemon M Jose. Training data attribution for image generation using ontology-aligned knowledge graphs. *arXiv preprint arXiv:2512.02713*, 2025.

[24] Theodoros Aivalis, Iraklis Angelos Klampanos, Antonis Troumpoukis, and Joemon Jose. Interpreting generative models through automatically-generated knowledge graphs. In *Proceedings of the 41st ACM/SIGAPP Symposium on Applied Computing*, SAC '26, page 1996–1998, New York, NY, USA, 2026. Association for Computing Machinery.

[25] Theodoros Aivalis, Iraklis A. Klampanos, and Antonis Troumpoukis. Llm-driven knowledge graph construction from earth observation data for extreme events. In Iraklis A. Klampanos, Antonis Troumpoukis, and Manolis Koubarakis, editors, *Proceedings of the Workshop on AI-driven Data Engineering and Reusability for Earth and Space Sciences (DARES 2025) co-located with the 28th European Conference on Artificial Intelligence (ECAI 2025), Bologna, Italy, October 25, 2025*, volume 4128 of *CEUR Workshop Proceedings*, pages 83–95. CEUR-WS.org, 2025.

[26] Yoshua Bengio, Ian Goodfellow, Aaron Courville, et al. *Deep learning*, volume 1. MIT press Cambridge, MA, USA, 2017. Sections 14.1 (Autoencoders) and 15.1 (Representation Learning).

[27] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006. Section 1.2.1 (Sum and Product Rules).

[28] Gordon V. Cormack, Charles L A Clarke, and Stefan Buettcher. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, page 758–759, New York, NY, USA, 2009. Association for Computing Machinery.

[29] Peiyuan Liao, Xiuyu Li, Xihui Liu, and Kurt Keutzer. The artbench dataset: Benchmarking generative models with artworks. *arXiv preprint arXiv:2206.11404*, 2022.

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021.

[31] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2818–2829, 2023.

[32] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864, 2016.

[33] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 10674–10685. ieee, 2022.

[34] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[35] Guillaume Charpiat, Nicolas Girard, Loris Felardos, and Yuliya Tarabalka. Input similarity from the neural network perspective. In *NeurIPS 2019-33th Annual Conference on Neural Information Processing Systems*, 2019.

[36] Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 19920–19930. Curran Associates, Inc., 2020.

[37] Zayd Hammoudeh and Daniel Lowd. Identifying a training-set attack's target using renormalized influence estimation. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pages 1367–1381, 2022.

[38] Kristian Georgiev, Joshua Vendrow, Hadi Salman, Sung Min Park, and Aleksander Madry. The journey, not the destination: How data guides diffusion models. In *ICML Workshop on Challenges in Deployable Generative AI*, 2023.

[39] Elnaz Barshan, Marc-Etienne Brunet, and Gintare Karolina Dziugaite. Relatif: Identifying explanatory training samples via relative influence. In *International Conference on Artificial Intelligence and Statistics*, pages 1899–1909. PMLR, 2020.

[40] iNaturalist. inaturalist research-grade observations, 2026. Accessed: May 2026.

[41] Daniele Zanaga, Ruben Van De Kerchove, Dirk Daems, Wanda De Keersmaecker, Carsten Brockmann, Grit Kirches, Jan Wevers, Oliver Cartus, Maurizio Santoro, Steffen Fritz, et al. Esa worldcover 10 m 2021 v200. 2022.

[42] Stephan Rasp, Stephan Hoyer, Alexander Merose, Ian Langmore, Peter Battaglia, Tyler Russel, Alvaro Sanchez-Gonzalez, Vivian Yang, Rob Carver, Shreya Agrawal, Matthew Chantry, Zied Ben Bouallegue, Peter Dueben, Carla Bromberg, Jared Sisk, Luke Barrington, Aaron Bell, and Fei Sha. Weatherbench 2: A benchmark for the next generation of data-driven global weather models, 2023.

[43] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*, 2016.

[44] Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The era5 global reanalysis. *Quarterly journal of the royal meteorological society*, 146(730):1999–2049, 2020.

A Artistic KG: Schema and Construction

To establish the topological prior for the artistic domain, we constructed a comprehensive Domain KG ($\mathcal{G}_D$). Rather than relying solely on the sparse labels provided by standard generative datasets, we followed an automated data extraction pipeline to build a deeply interconnected historical network.

A.1 Data Collection and Sequential Enrichment

The data ingestion process followed a multi-stage enrichment strategy designed to transform a flat vision dataset into a relational network.

1. **Primary Ingestion:** We initiated the pipeline using painting identifiers and artist names provided by ArtBench-10.
2. **Metadata Expansion (WikiArt):** Using these identifiers as primary keys, we queried the WikiArt API to systematically retrieve all available technical and categorical metadata for each artwork. We then performed a secondary expansion to ingest granular biographical profiles for the associated artists, including active years and thematic tags.
3. **Relational Resolution (Wikipedia & DBpedia):** To capture all the possible historical context, we utilised the Wikipedia URLs embedded within the WikiArt records as bridge identifiers to aggregate data from both Wikipedia and DBpedia. From Wikipedia, we extracted available biographical summaries and historical descriptions of each artist. Simultaneously, we leveraged the structured nature of DBpedia by executing targeted SPARQL queries to ingest specific relational attributes. By combining these sources, we were able to map complex interpersonal networks, effectively recovering the structural context that is inherently missing from the raw vision data.

A visual example of a painting and its corresponding author is illustrated in Figure 4.



(a) Artwork: Ruokokoski Maalaa



(b) Artist: Sallinen Tyko

Figure 4: The artwork Ruokokoski Maalaa (a) presented alongside an image of its creator, Sallinen Tyko (b). This illustrates the real-world visual and historical context underlying the dataset.

A.2 Ontological Schema and Neo4j Modeling

To move beyond flat, tabular metadata, all extracted entities and relationships were instantiated in Neo4j, a native graph database optimised for traversing highly interconnected data. Modeling the domain as a network of nodes (entities) connected by directed edges (relationships) allows us to explicitly capture the historical lineage of the training data.

The ontological schema for our artistic domain centers on two primary anchor nodes: `Painting` and `Artist`. Around these central nodes, we linked various contextual entities. The overall structural design of this ontology is illustrated in Figure 5.

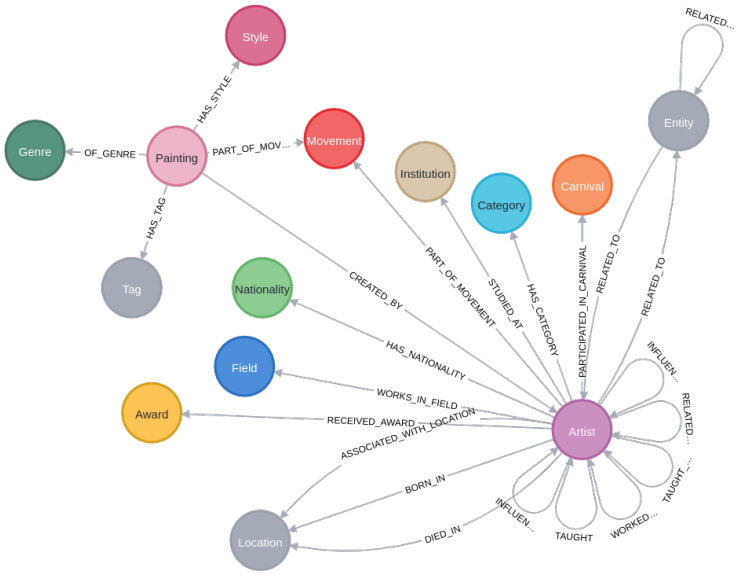


Figure 5: Ontological schema of the Artistic KG ($\mathcal{G}_D$), defining the structural relationships used to map ArtBench to its domain context.

By mapping the extracted data to this schema, we ensure that the resulting topology accurately reflects ground-truth art history. Structuring the domain in Neo4j provides the necessary foundation for our problem. For example, if two visually distinct paintings were created by artists who studied under the same master, the graph structure connects them via a short path length, allowing our embedding algorithms to capture their structural proximity.

Relational Logic. Table 3 outlines the core deterministic relationships utilised to construct the graph. This schema defines how the framework traverses the space between a physical artifact and its historical context.

Table 3: Core Entity Relationships within the Artistic Knowledge Graph ($\mathcal{G}_D$). This schema defines the permissible edges that control the structural topology of the domain.

Source Node	Relationship Edge	Target Node
`Painting`	`CREATED_BY`	`Artist`
`Painting`	`HAS_STYLE`	`ArtStyle`
`Painting`	`HAS_GENRE`	`Genre`
`Painting`	`USES_MEDIUM`	`Medium`
`Artist`	`INFLUENCED_BY`	`Artist`
`Artist`	`STUDIED_UNDER`	`Artist`
`Artist`	`ASSOCIATED_WITH`	`ArtMovement`
`Artist`	`BORN_IN`	`Location`

Scale and Connectivity. To evaluate the complexity and semantic density of our database, we provide a quantitative inventory in Table 4. The high volume of unique tags and artist relationships ensures that the graph defines a rich semantic space, preventing the attribution from collapsing into generic clusters.

Metadata Details. The internal properties of each node type are detailed in Table 5. These fields allow for detailed filtering, such as isolating artists active during specific decades, which is critical for the mathematical sharpening of the attribution scores.

In Figure 6 we can see the KG representation of a painting and its corresponding author (the same as in Figure 4) in Neo4j Database.

Table 4: KG Inventory: Quantitative Analysis of Relationship Density and Semantic Diversity across $\mathcal{G}_D$.

Source	Relationship	Target	Count	Unique
Painting	HAS_TAG	Tag	122448	3651
Painting	CREATED_BY	Artist	47224	1769
Painting	HAS_STYLE	Style	47224	81
Painting	HAS_GENRE	Genre	47224	339
Painting	PART_OF_MOVEMENT	Movement	47224	73
Artist	HAS_CATEGORY	Category	37483	9776
Artist	ASSOCIATED_WITH_LOCATION	Location	4147	2284
Artist	WORKS_IN_FIELD	Field	3597	241
Artist	PART_OF_MOVEMENT	Movement	3575	393
Artist	STUDIED_AT	Institution	3064	701
Entity	RELATED_TO	Entity	2554	960
Artist	HAS_NATIONALITY	Nationality	2113	93
Artist	BORN_IN	Location	1695	1018
Artist	DIED_IN	Location	1456	627
Artist	INFLUENCED_BY	Artist	796	321
Artist	WORKED_WITH	Artist	652	350
Artist	INFLUENCED	Artist	627	362
Artist	RECEIVED_AWARD	Award	295	231
Artist	TAUGHT_BY	Artist	291	171
Artist	TAUGHT	Artist	250	214
Artist	RELATED_TO	Artist	121	104
Artist	PARTICIPATED_IN_CARNIVAL	Carnival	2	2

Table 5: Internal Node Property Schema: Detailed metadata fields stored within the $\mathcal{G}_D$ topology for filtering and state representation.

Node Label	Internal Properties (Metadata Fields)
Artist	id, name, born_date, died_date, wikipedia_summary, active_years
Genre	id, display_name
MainTarget	id, url
Movement	id, display_name
Painting	id, url
Style	id, display_name
Tag	id, display_name

B Contrastive Learning and Representation Alignment

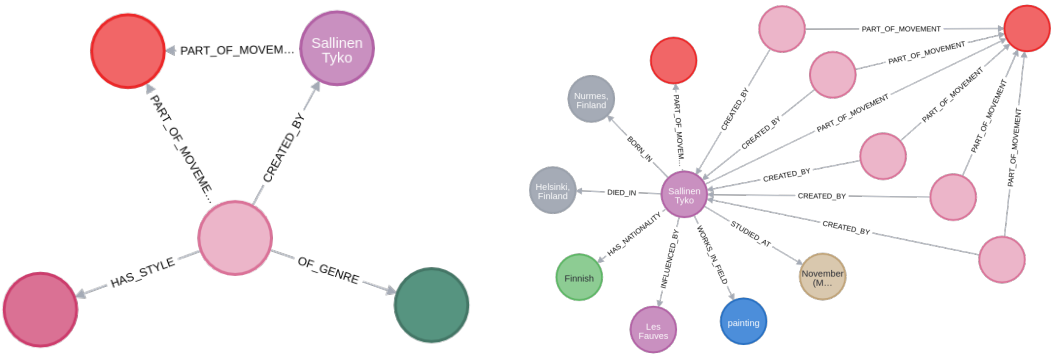
To compute the Bayesian attribution components in the Domain Space, the feature Likelihood $P(\mathbf{C} \mid \mathbf{X}_j)$ and the Prior $P(\mathbf{X}_j)$, the framework requires a unified semantic space. As discussed in Section 4.1.1, generative models synthesise novel artifacts ($\mathbf{C}$) that lack intrinsic metadata. To evaluate these outputs against our Domain KG, we establish a mathematical bridge between continuous visual features and discrete structural knowledge.

Our approach utilises a cross-modal contrastive learning objective. Inspired by the CLIP architecture [30], we align the latent distributions of two different modalities into a shared vector space, allowing for direct cosine similarity measurements between pixels and graph nodes.

B.1 Encoder Architectures and Domain Extraction

To construct this joint space, we must first extract unimodal representations from both the visual and structural domains. We carefully selected state-of-the-art encoders optimised for their respective modalities.

Visual Encoding. For the continuous image modality, we utilise the ViT-g-14 architecture [31]. We selected this model because its large-scale pre-training provides the discriminative depth necessary to capture stylistic characteristics that smaller encoders might overlook. This choice is further validated



(a) Artwork: Ruokokoski Maalaa

(b) Artist: Sallinen Tyko

Figure 6: Graph visualisation of the painting Ruokokoski Maalaa (a) and the artist Sallinen Tyko (b). The network illustrates the central Artist and Painting nodes, detailing their surrounding connections by displaying a subset of the artist's portfolio alongside broader metadata.

in our ablation study (see Table 6), where the ViT-g-14 demonstrated superior predictive faithfulness compared to alternative architectures. In the context of attribution, this high-dimensional capacity is essential for distinguishing between closely related training samples that share surface-level similarities but differ in stylistic characteristics.

Structural Encoding via Neo4j and Node2Vec. The domain metadata is stored natively in Neo4j and encoded using the Node2Vec algorithm [32]. Node2Vec's biased random walks allow us to tune the representation between local stylistic details and global historical context. A key advantage of this approach is its seamless integration with the Neo4j Graph Data Science (GDS) library, which provides optimised, production-ready procedures to extract high-dimensional embeddings directly within the database environment. This efficient implementation ensures that the structural prior $P(\mathbf{X}_j)$ accurately reflects the topological rarity of a node's position within the art-historical network, preventing generic hubs, such as heavily represented styles, from dominating the attribution results.

B.2 Alignment Procedure and InfoNCE Loss

To construct the joint multimodal manifold, raw representations from the ViT and Node2Vec encoders are passed through learned non-linear projection heads. This projection is necessary to map disparate latent distributions into a shared d_e-dimensional space ($d_e = 512$). The procedure is governed by a Symmetric InfoNCE Loss, mirroring the contrastive objective utilised in the CLIP architecture [30], that maximises the lower bound of the mutual information between visual textures and structural graph signatures. By adopting this loss, we ensure that the model learns a discriminative representation where matched pairs are strongly attracted while mismatched pairs are effectively repelled in the joint embedding space.

Architectural Specifics. Each projection head consists of a deep Multi-Layer Perceptron (MLP) designed for high-fidelity alignment. To prevent mode collapse, we incorporate Batch Normalization layers after the first linear transformation. Additionally, a dropout rate of 0.2 is applied to enhance generalization and prevent the model from over-fitting to specific historical samples in the training set. A critical component of our architecture is the Dual-Normalization strategy: input vectors are L2-normalised before the MLP to stabilise the initial gradients, and final embeddings are re-normalised to ensure they lie on a hypersphere, making the cosine similarity calculation mathematically consistent.

Optimization and Training Schedule. The alignment model is optimised using the AdamW optimiser with a weight decay of 10^{-4} to ensure weight sparsity. We utilise a Cosine Annealing learning rate scheduler, which starts at a conservative 10^{-4} and smoothly decays to zero over 200 epochs. This schedule allows the model to explore the complex cross-modal landscape early in training while finely tuning the alignment in the final stages. To calibrate the similarity scores, we employ a learnable temperature parameter τ, initialised at 0.07 and clamped at 100 to prevent gradient explosions during the calculation of logits.

The contrastive training procedure illustrated in Figure 7 establishes the shared representation space required to perform direct semantic comparisons between generated artifacts and the training dataset.

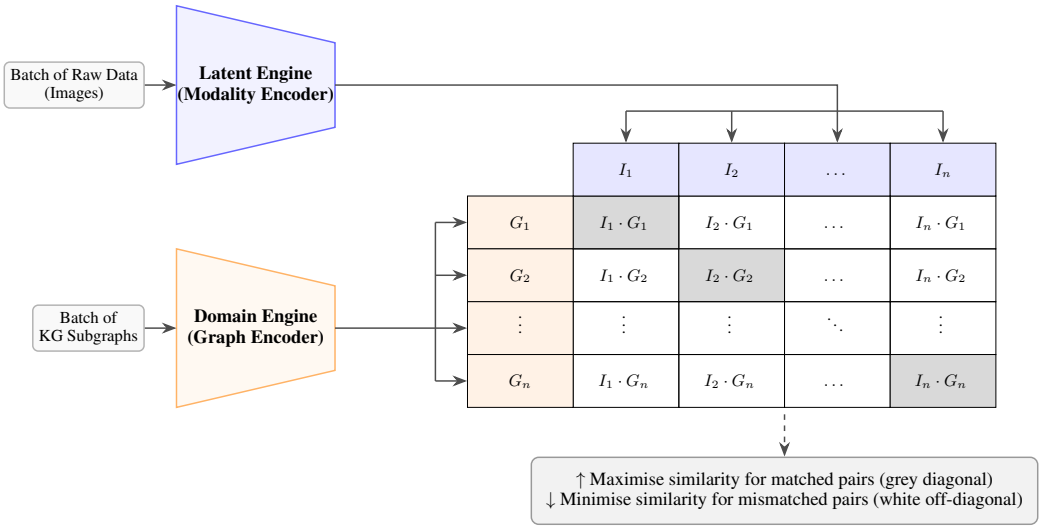


Figure 7: The contrastive multimodal alignment procedure. A batch of raw continuous data (images) and their corresponding subgraphs are passed through parallel stacked encoders. A symmetric InfoNCE loss is applied to the resulting similarity matrix to maximise the alignment of corresponding pairs (the highlighted diagonal) while pushing apart mismatched pairs.

The alignment objective is explicitly optimised through a batch-wise similarity calculation scaled by a learnable temperature parameter τ. This mechanism ensures that the model remains discriminative across a diverse range of artistic styles and historical contexts. The complete algorithmic implementation of this alignment loop, utilising the CLIP-style symmetric loss, is detailed in Figure 8.

```
# image_encoder - ViT-g-14
# graph_encoder - Node2Vec (extracted via Neo4j)
# I[n, c, h, w] - minibatch of raw training images
# G[n, d_g]     - minibatch of topological graph embeddings
# MLP_i         - non-linear projection head for visual features
# MLP_g         - non-linear projection head for graph features
# t             - learned temperature parameter

# 1. Extract feature representations of each modality
I_f = image_encoder(I) #[n, d_i]
G_f = graph_encoder(G) #[n, d_g]

# 2. Project into joint multimodal embedding space [n, d_e]
# Output vectors are strictly L2-normalized within the MLP heads
I_e = MLP_i(I_f)
G_e = MLP_g(G_f)

# 3. Compute scaled pairwise cosine similarities [n, n]
logits = np.dot(I_e, G_e.T) * np.exp(t)

# 4. Compute symmetric loss function (CLIP Objective)
labels = np.arange(n)
loss_i = cross_entropy_loss(logits, labels, axis=0)
loss_g = cross_entropy_loss(logits, labels, axis=1)
loss = (loss_i + loss_g) / 2
```

Figure 8: Algorithmic implementation of the contrastive alignment loop. The procedure leverages parallel projection heads and a symmetric loss function to maximise the cross-modal mutual information required for grounded Bayesian attribution.

C Generative Setup and Hyperparameters

To ensure an exact and rigorous comparison with our parametric baselines, we strictly replicate the diffusion training and generation pipeline utilised by D-TRAK [17] for the ArtBench dataset [29].

Model Architecture and Fine-Tuning. We utilise a pre-trained Stable Diffusion model adapted to a 256×256 resolution. To make training computationally tractable while preserving generative quality, we employ Low-Rank Adaptation (LoRA) on the cross-attention layers of the U-Net. The LoRA rank dimension is set to $r = 128$, resulting in approximately 25.5M trainable parameters. The model is fine-tuned conditionally using textual prompts formatted as "a {class} painting" (e.g., "a ukiyo-e painting").

Optimization Schedule. The model is trained for 100 epochs using a total batch size of 64. We optimise the parameters using the AdamW optimiser with a weight decay of 10^{-6} and a dropout rate of 0.1. Data augmentation is limited to random horizontal flips. The learning rate follows a cosine annealing schedule, initiating with a 0.1 fraction warmup to a peak learning rate of 3×10^{-4}.

Inference and Generation. During the evaluation phase, target images ($\mathbf{C}$) are synthesised utilising the 50-step Denoising Diffusion Implicit Model (DDIM) solver. To enforce strong class conditioning, we apply classifier-free guidance with a fixed scale of 7.5.

D Linear Datamodeling Score (LDS) Formulation

D.1 Conceptual Framework

The Linear Datamodeling Score (LDS), introduced by [16], formalises the evaluation of data attribution as a counterfactual prediction task. The core objective is to determine if an attribution method can accurately predict how a model's output would change if specific training samples were removed.

For a fixed target z, the model output $f(z; \theta^*(S'))$, where $\theta^*(S')$ represents the parameters resulting from training on a subset $S' \subset S$, is viewed as a function of the training subset. An ideal attribution method τ should provide scores that, when summed, approximate the model's behavior on any given subset.

D.2 Mathematical Definition

Following the formulation of the original TRAK framework [16], the attribution-based prediction for a subset S_j is defined as the additive sum of the individual scores of all training samples z_i contained within that subset:

$$g_\tau(z, S_j) = \sum_{i : z_i \in S_j} \tau(z, S)_i \tag{7}$$

The LDS for a specific target z is then defined as the Spearman rank correlation (ρ) between the actual model outputs across m randomly sampled subsets and the corresponding attribution-based predictions:

$$\text{LDS}(\tau, z) := \rho\left(\{f(z; \theta^*(S_j))\}_{j=1}^m, \{g_\tau(z, S_j)\}_{j=1}^m\right) \tag{8}$$

where ρ denotes the Spearman rank correlation. A high correlation indicates that the attribution scores τ correctly capture the marginal contribution of each training sample to the final output.

D.3 Implementation for Diffusion Models

Following the rigorous protocol established by D-TRAK [17], we implement several approximations to account for the stochasticity inherent in generative diffusion processes. We first perform **ensemble training** by sampling $M = 64$ random subsets of the training set $\mathcal{D}$, each consisting of a fraction $\alpha = 0.5$ of the total data. To mitigate variance from random initialization, we train an ensemble of three independent models with different random seeds for every subset. To evaluate the resulting model output $f(z; \theta^*(S_j))$, we compute the diffusion loss L_{Simple} using a robust **loss approximation** strategy. Specifically, to ensure a stable estimate of the expectation over time ($\mathbb{E}_t$) and noise ($\mathbb{E}_\epsilon$),

we evaluate the loss across 1,000 evenly spaced timesteps within the interval $[1, T]$, sampling three standard Gaussian noise vectors $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ at each step. The **final scoring** is determined by calculating the LDS for each sample of interest across both the validation and generated sets, with the final reported metric being the arithmetic mean of these individual Spearman correlations.

D.4 Statistical Significance and Uncertainty Quantification

To ensure the reported LDS metrics are statistically robust and not biased by specific subset selections, we perform uncertainty quantification following the methodology of prior work. We calculate 95% confidence intervals for all reported correlations using a bootstrapping procedure with $n = 64$ resamples. For each resample, we compute the mean Spearman rank correlation across our evaluation sets. This process allows us to report both the stability and the significance of our attribution results, ensuring that the performance gains of our framework are consistent across different random draws of the training distribution.

E Baseline Implementations

We focus on conducting data attribution in a post-hoc manner, which refers to the application of attribution methods after model training. Post-hoc attribution methods do not add extra restrictions to how we train models and are thus preferred in practice [43].

Raw Pixel. As a naive image-space baseline, this method computes the direct mathematical similarity between the raw pixel values of the images. We represent both the generated artifact $\mathbf{C}$ and the training samples $\mathbf{X}_j$ as flattened RGB vectors. We then calculate the dot product and cosine similarity between these vectors to measure spatial and coloristic overlap.

CLIP Similarity. This baseline serves as a standard similarity-based attribution method by encoding both the generated target and the training samples into a shared multimodal space using a pre-trained CLIP vision encoder [30]. Following established protocols for black-box data attribution, the influence score is calculated as the dot product or cosine similarity between these dense embeddings. While CLIP remains the primary similarity-based reference, we also investigated the impact of various alternative vision encoders on attribution performance. A comprehensive comparison of these architectures and their respective results is detailed in our ablation study, specifically provided in Table 6.

Gradient Similarity. Following the approach of Charpiat et al. [35], this gradient-based estimator quantifies influence by measuring the alignment between projected gradient representations. It captures input similarity from the perspective of the model's internal parameters using the following dot product or cosine similarity formulations:

$$\tau(x, \mathcal{D})_n = \mathcal{P}^\top \nabla_\theta \mathcal{L}_{\text{Simple}}(x; \theta^*)^\top \cdot \mathcal{P}^\top \nabla_\theta \mathcal{L}_{\text{Simple}}(x^n; \theta^*) \tag{9}$$

$$\tau(x, \mathcal{D})_n = \frac{\mathcal{P}^\top \nabla_\theta \mathcal{L}_{\text{Simple}}(x; \theta^*)^\top \cdot \mathcal{P}^\top \nabla_\theta \mathcal{L}_{\text{Simple}}(x^n; \theta^*)}{||\mathcal{P}^\top \nabla_\theta \mathcal{L}_{\text{Simple}}(x; \theta^*)|| \cdot ||\mathcal{P}^\top \nabla_\theta \mathcal{L}_{\text{Simple}}(x^n; \theta^*)||} \tag{10}$$

TracInCP. We utilise the TracInCP estimator [36], which identifies training data influence by tracing the model's performance change along the learning trajectory. Influence is approximated by averaging the inner products of the projected gradients at C evenly spaced checkpoints θ^c:

$$\tau(x, \mathcal{D})_n = \frac{1}{C} \sum_{c=1}^{C} \mathcal{P}_c^\top \nabla_\theta \mathcal{L}_{\text{Simple}}(x; \theta^c)^\top \cdot \mathcal{P}_c^\top \nabla_\theta \mathcal{L}_{\text{Simple}}(x^n; \theta^c) \tag{11}$$

GAS. The Gradient Averaging Similarity (GAS) is a variant of the TracInCP method that normalises the gradient vectors to calculate cosine similarity rather than raw dot products [37].

TRAK. We adapt the standard TRAK estimator [16] to the diffusion setting. It applies random projections to the model's gradients to create a tractable, low-dimensional gradient matrix, calculating attribution via a regularised kernel dot product.

- Retraining-free TRAK, where a damping term λI is included to ensure numerical stability and provide regularization:

$$\Phi_{\text{TRAK}} = [\phi(x^1); \ldots; \phi(x^N)]^\top \tag{12}$$

$$\tau(x, \mathcal{D})_n = \mathcal{P}^\top \nabla_\theta \mathcal{L}_{\text{Simple}}(x; \theta^*)^\top \cdot (\Phi_{\text{TRAK}}^\top \Phi_{\text{TRAK}} + \lambda I)^{-1} \cdot \mathcal{P}^\top \nabla_\theta \mathcal{L}_{\text{Simple}}(x^n; \theta^*) \tag{13}$$

- Ensemble TRAK: To improve robustness, the ensemble variant averages influence scores across S models trained on distinct data subsets:

$$\Phi_{\text{TRAK}}^s = [\phi^s(x^1); \ldots; \phi^s(x^N)]^\top, \quad \text{where } \phi^s(x) = \mathcal{P}_s^\top \nabla_\theta \mathcal{L}_{\text{Simple}}(x; \theta_s^*) \tag{14}$$

$$\tau(x, \mathcal{D})_n = \frac{1}{S} \sum_{s=1}^{S} \mathcal{P}_s^\top \nabla_\theta \mathcal{L}_{\text{Simple}}(x; \theta_s^*)^\top \cdot (\Phi_{\text{TRAK}}^{s\top} \Phi_{\text{TRAK}}^s + \lambda I)^{-1} \cdot \mathcal{P}_s^\top \nabla_\theta \mathcal{L}_{\text{Simple}}(x^n; \theta_s^*) \tag{15}$$

Relative Influence. The Relative Influence, introduced by Barshan et al. [39], seeks to improve the standard influence function estimator by normalising it against the magnitude of the Hessian-Vector Product (HVP). We adapt this methodology to the diffusion setting and apply the same scalability optimizations used in TRAK. The attribution score is calculated as:

$$\tau(x, \mathcal{D})_n = \frac{\mathcal{P}^\top \nabla_\theta \mathcal{L}_{\text{Simple}}(x; \theta^*)^\top \cdot (\Phi_{\text{TRAK}}^\top \Phi_{\text{TRAK}} + \lambda I)^{-1} \cdot \mathcal{P}^\top \nabla_\theta \mathcal{L}_{\text{Simple}}(x^n; \theta^*)}{||(\Phi_{\text{TRAK}}^\top \Phi_{\text{TRAK}} + \lambda I)^{-1} \cdot \mathcal{P}^\top \nabla_\theta \mathcal{L}_{\text{Simple}}(x^n; \theta^*)||} \tag{16}$$

Renormalised Influence. Introduced by Hammoudeh and Lowd [37], the Renormalised Influence baseline seeks to calibrate influence scores by normalising them against the magnitude of the training sample's gradients. The resulting renormalised attribution score is formulated as:

$$\tau(x, \mathcal{D})_n = \frac{\mathcal{P}^\top \nabla_\theta \mathcal{L}_{\text{Simple}}(x; \theta^*)^\top \cdot (\Phi_{\text{TRAK}}^\top \Phi_{\text{TRAK}} + \lambda I)^{-1} \cdot \mathcal{P}^\top \nabla_\theta \mathcal{L}_{\text{Simple}}(x^n; \theta^*)}{||\mathcal{P}^\top \nabla_\theta \mathcal{L}_{\text{Simple}}(x^n; \theta^*)||} \tag{17}$$

Journey TRAK. Georgiev et al. [38] introduced Journey TRAK, which originally focuses on attributing the intermediate noisy images x_t along the diffusion trajectory rather than the final generated output. The formulation is as follows:

$$\tau(x, \mathcal{D})_n = \frac{1}{T'} \sum_{t=1}^{T'} \mathcal{P}^\top \nabla_\theta \mathcal{L}_{\text{Simple}}^t(x_t; \theta^*)^\top \cdot (\Phi_{\text{TRAK}}^\top \Phi_{\text{TRAK}} + \lambda I)^{-1} \cdot \mathcal{P}^\top \nabla_\theta \mathcal{L}_{\text{Simple}}(x^n; \theta^*) \tag{18}$$

where T' represents the total number of inference steps (set to 50 here), and x_t denotes the noisy generated image at timestep t along the sampling trajectory.

D-TRAK. Similar to TRAK, we adapt D-TRAK[17] to our setting. We implement the model output function $f(z, \theta)$ as $\mathcal{L}_{\text{Square}}$. D-TRAK is implemented using the following equations:

$$\boldsymbol{\Phi}_{\text{D-TRAK}} = [\boldsymbol{\Phi}(\boldsymbol{x}^1), \ldots, \boldsymbol{\Phi}(\boldsymbol{x}^N)]^\top, \quad \text{where } \boldsymbol{\Phi}(\boldsymbol{x}) = \boldsymbol{P}^\top \nabla_\theta \mathcal{L}_{\text{Simple}}(\boldsymbol{x}, \theta) \tag{19}$$

$$\tau(z, \mathbb{S})^i = (\boldsymbol{P}^\top \nabla_\theta \mathcal{L}_{\text{Simple}}(\boldsymbol{x}, \theta))^\top \cdot \left(\boldsymbol{\Phi}_{\text{TRAK}}^\top \boldsymbol{\Phi}_{\text{TRAK}} + \lambda \boldsymbol{I}\right)^{-1} \cdot \boldsymbol{P}^\top \nabla_\theta \mathcal{L}_{\text{Simple}}(\boldsymbol{x}^i, \theta) \tag{20}$$

where $\lambda \boldsymbol{I}$ is also included for numerical stability and regularization, as in TRAK. Additionally, the output function $f(z, \theta)$ could be replaced with other functions.

Diffusion Attribution Score (DAS). Introduced by Lin et al. [18], DAS addresses the limitations of standard loss-based attribution. It directly compares the predicted distributions using Kullback-Leibler (KL) divergence. To isolate the effect of removing a specific training sample $z^{(i)}$ on the generated target z_{gen}, the score is defined mathematically as the KL-divergence between the original and retrained models' noise predictors:

$$\tau_{\text{DAS}}(z_{\text{gen}}, \mathcal{S})^{(i)} \approx \mathbb{E}_{\epsilon,t}\left[||\epsilon_\theta(x_t^{\text{gen}}, t) - \epsilon_{\theta\setminus i}(x_t^{\text{gen}}, t)||^2\right] \tag{21}$$

DAS estimates the change in model parameters using a single Newton step and also normalises and averages the gradients and residuals over the entire generation trajectory, and utilises random projection matrices ($\mathcal{P}$) to compress the gradient dimension.

Let $g(x)$ denote the time-averaged, normalised, and randomly projected gradient of the noise predictor, $G(\mathcal{S})$ represent the stacked gradient matrix for the training set, and $r^{(i)}$ denote the averaged residual. The final, tractable DAS formulation is computed as:

$$\tau_{\text{DAS}}(z_{\text{gen}}, \mathcal{S})^{(i)} = \left\| \frac{g(x_{\text{gen}})^\top \left[G(\mathcal{S})^\top G(\mathcal{S})\right]^{-1} g(x^{(i)}) r^{(i)}}{1 - g(x^{(i)})^\top \left[G(\mathcal{S})^\top G(\mathcal{S})\right]^{-1} g(x^{(i)})} \right\|^2 \tag{22}$$

F Ablation Study

In this section, we provide a detailed analysis of the architectural and mathematical choices that define our framework. We first examine how the choice of visual encoder affects attribution accuracy. We then evaluate the impact of our Bayesian components, specifically the non-linear scaling, the dataset prior, and our fusion strategy, to show how each part contributes to the final performance.

Visual Encoder Scaling and Data Diversity : As shown in Table 6, we tested several vision backbones to find the best feature extractor for our Latent Space. We found a clear link between the model size and the accuracy of the attribution results. While smaller models like ViT-B/32 struggle to find the specific styles needed for ArtBench, the ViT-g-14 model from OpenCLIP consistently gives the highest results. This performance is likely because ViT-g-14 was trained on the massive LAION-2B dataset. This large-scale, high-diversity training set allows the model to build a more general understanding of images, which is necessary for tracking influences across complex domains.

Table 6: Ablation study of visual encoders. We report the LDS, % for all the selected encoders. The ViT-g-14 model consistently provides the best results across both datasets. This is likely due to the massive scale and diversity of the LAION-2B dataset used for its pre-training, which allows the model to understand complex artistic styles better than smaller-scale alternatives.

	ArtBench-2 LDS (%)		**ArtBench-5 LDS (%)**	
Encoder Architecture	**Validation**	**Generation**	**Validation**	**Generation**
ViT-B/32 (dot prod.)	7.3 ± 1.4	5.3 ± 2.4	5.4 ± 0.8	3.5 ± 1.9
ViT-B/32 (cosine)	8.9 ± 1.3	8.8 ± 2.2	6.7 ± 0.7	5.5 ± 1.8
ViT-L/14@336px (dot prod.)	9.2 ± 1.1	6.2 ± 2.7	5.9 ± 1.0	4.9 ± 2.0
ViT-L/14@336px (cosine)	9.8 ± 1.0	7.7 ± 2.6	7.1 ± 1.0	6.0 ± 1.9
RN50x64 (dot prod.)	9.8 ± 1.2	7.7 ± 2.5	6.4 ± 0.9	6.0 ± 2.0
RN50x64 (cosine)	10.5 ± 1.0	8.2 ± 2.6	7.0 ± 0.9	5.5 ± 2.0
DINOv2 (dot prod.)	10.1 ± 1.1	10.2 ± 2.6	7.4 ± 0.8	8.1 ± 2.1
DINOv2 (cosine)	10.0 ± 1.0	10.1 ± 2.7	7.3 ± 0.8	7.9 ± 2.0
ViT-g-14 (dot prod.)	11.9 ± 0.9	10.9 ± 2.3	9.0 ± 0.8	8.6 ± 2.0
ViT-g-14 (cosine)	$\mathbf{12.3 \pm 1.0}$	$\mathbf{11.1 \pm 2.3}$	$\mathbf{9.3 \pm 0.8}$	$\mathbf{8.6 \pm 1.9}$

Table 7: Ablation study of our proposed attribution framework. LDS scores are reported. Our results show that raising the similarity score to a power (p^5 and p^{10}) significantly improves accuracy by rewarding the most relevant images and driving unimportant scores toward zero. Furthermore, our Asymmetric Rank Boost outperforms standard fusion by using the Domain's ranking to guide the visual comparison, ensuring the final attribution is both visually and semantically correct.

Modality	Formulation	ArtBench-2		ArtBench-5	
		Validation	Generation	Validation	Generation
Visual	$p(c\|x_j)$	12.30 ± 1.00	11.10 ± 2.30	9.30 ± 0.75	8.60 ± 1.90
	$p(c\|x_j)^5$	21.60 ± 0.85	18.00 ± 2.25	21.00 ± 0.45	16.80 ± 1.45
	$p(c\|x_j)^{10}$	27.00 ± 0.70	20.10 ± 1.80	27.00 ± 0.30	17.50 ± 1.05
	$p(c\|x_j)^{10}p(x_j)$	28.50 ± 0.60	20.80 ± 1.70	28.10 ± 0.30	18.00 ± 1.00
Domain	$p(c\|x_j)$	7.80 ± 1.00	6.20 ± 1.50	6.10 ± 0.85	4.60 ± 1.35
	$p(c\|x_j)^5$	16.80 ± 0.65	11.90 ± 1.35	15.30 ± 0.65	10.90 ± 1.35
	$p(c\|x_j)^{10}$	19.70 ± 0.60	12.30 ± 1.30	19.30 ± 0.55	10.70 ± 1.40
	$p(c\|x_j)^{10}p(x_j)$	20.40 ± 0.65	12.40 ± 1.20	19.80 ± 0.50	10.80 ± 1.35
Hybrid (50/50)	$p(c\|x_j)$	10.90 ± 0.85	9.30 ± 2.00	8.60 ± 0.75	7.10 ± 1.25
	$p(c\|x_j)^5$	20.80 ± 0.70	16.00 ± 1.80	20.00 ± 0.60	15.20 ± 1.25
	$p(c\|x_j)^{10}$	25.60 ± 0.65	17.20 ± 1.70	25.70 ± 0.45	15.20 ± 1.15
	$p(c\|x_j)^{10}p(x_j)$	26.67 ± 1.20	17.73 ± 1.14	26.37 ± 1.24	15.38 ± 1.19
Hybrid (70/30)	$p(c\|x_j)$	12.10 ± 0.85	10.70 ± 2.15	9.40 ± 0.75	8.30 ± 1.35
	$p(c\|x_j)^5$	21.70 ± 0.75	17.30 ± 2.00	21.00 ± 0.50	16.30 ± 1.30
	$p(c\|x_j)^{10}$	26.80 ± 0.75	18.90 ± 1.85	27.20 ± 0.40	16.70 ± 1.10
	$p(c\|x_j)^{10}p(x_j)$	28.00 ± 0.70	19.40 ± 1.55	26.40 ± 0.45	15.40 ± 1.10
Proposed	**Asymmetric Rank Boost**	$\mathbf{29.10 \pm 0.65}$	$\mathbf{21.30 \pm 1.65}$	$\mathbf{28.90 \pm 0.35}$	$\mathbf{18.40 \pm 1.05}$

Mathematical Sharpening and Fusion Strategy. : Table 7 provides a detailed breakdown of our framework's components, where our tests highlight two critical findings. First, we observe that non-linear scaling is essential for improving the attribution results. By raising the similarity score to a high power, such as $p(c|x_j)^{10}$, the framework effectively rewards a small group of training images that hold actual influence in generative models while pushing the scores of unrelated images toward zero. Second, our proposed Asymmetric Rank Boost consistently performs better than standard linear fusion methods like 50/50 or 70/30 weighting. By using the Domain KG as a guide to re-rank the top visual matches, we maintain a clear visual signal while leveraging historical evidence to resolve ties. This ensures the final attribution is both visually plausible and historically accurate, allowing our black-box system to match the performance of more complex, gradient-based methods.

Boost Decay Parameter k. The parameter k in Eq. (6) controls the decay rate of the domain-based boost and therefore determines how rapidly lower-ranked domain matches lose influence on the final attribution weight α_j. Table 8 reports a sensitivity analysis across $k \in \{1, 5, 10, 20, 30, 60\}$. We observe that very small values (e.g., $k = 1$) concentrate the boost on the few top-ranked domain matches, over-emphasising the domain evidence and overriding the primary latent signal. This effect is most pronounced on the ArtBench-5 Generation set, where performance drops to 13.2%. Conversely, larger values flatten the boost across ranks, so that the KG can no longer discriminate between strongly and weakly related candidates, and performance decreases steadily as k grows. A moderate value of $k = 5$ provides the best balance, achieving the highest LDS across all four evaluation settings. Notably, the Generation sets are more sensitive to the choice of k than the Validation sets. We therefore adopt $k = 5$ for all reported experiments.

G Environmental KG: Schema and Construction

To establish the necessary domain knowledge for the attribution in the environmental domain, we constructed a comprehensive KG ($\mathcal{G}_E$) that links physical flood events to their ecological and geographic surroundings. This allows the framework to evaluate weather predictions not just as pixel grids, but as events with real-world context.

Table 8: Sensitivity analysis of the boost decay parameter k in the Asymmetric Rank Boost (Eq. (6)). We report the LDS (%) on ArtBench-2 and ArtBench-5. Small values over-emphasise the domain evidence, while large values weaken the contribution of the KG. A moderate value of $k = 5$ consistently provides the best results across both datasets.

	ArtBench-2 LDS (%)		ArtBench-5 LDS (%)	
Parameter k	Validation	Generation	Validation	Generation
$k = 1$	27.7 ± 0.50	16.4 ± 1.15	26.4 ± 0.75	13.2 ± 0.95
$k = 5$	$\mathbf{29.1 \pm 0.65}$	$\mathbf{21.3 \pm 1.65}$	$\mathbf{28.9 \pm 0.35}$	$\mathbf{18.4 \pm 1.05}$
$k = 10$	28.5 ± 0.35	19.2 ± 1.25	26.3 ± 0.50	16.1 ± 1.20
$k = 20$	27.5 ± 0.55	19.1 ± 1.35	25.1 ± 0.50	16.4 ± 1.20
$k = 30$	26.7 ± 0.55	18.8 ± 1.40	24.1 ± 0.50	16.1 ± 1.30
$k = 60$	25.0 ± 0.60	18.0 ± 1.40	22.2 ± 0.50	16.0 ± 1.30

G.1 Data Collection and Source Integration

To construct the environmental KG ($\mathcal{G}_E$), we developed a multi-source data pipeline anchored around specific historical weather events. Our foundational reference is based on a comprehensive registry of global extreme flood events [25]. This dataset provides the critical spatiotemporal anchors for our study, including unique event identifiers, the precise geographical coordinates of the region's event, temporal footprints, and macroscopic impact metrics such as human displacement and damage severity. While this foundational dataset establishes the macro-level context of extreme floods, evaluating and grounding predictions from generative meteorological models introduces a significant scale mismatch. Standard climate models typically generate predictions at a relatively coarse spatial resolution (e.g., a $0.25^\circ \times 0.25^\circ$ grid). At this macroscopic scale, it is impossible to directly map predictions to smaller, localised geographical features. To enable accurate, high-resolution data attribution, we must project these grid-level predictions into a richer semantic space. By augmenting the regional profiles with fine-grained external knowledge, we construct a domain representation that explicitly models the topological similarities, physical characteristics, and ecological interconnections between different disaster zones. This allows us to meaningfully compare historical events and trace potential lineages via retrieval at a much higher resolution than the models' native output grids.

To build this highly interconnected representation, we extended the geographical profiles of each flood region using an enrichment-based approach. We integrated the following high-resolution data sources:

- **Biological Enrichment: iNaturalist[40].** Using the GPS anchors from the primary flood dataset, we enriched each event with localised ecological data. We queried the iNaturalist database to retrieve biological observations within a 10km radius of each flood center. This process provided verified taxonomic data (Species, Genus, and Kingdom), allowing the KG to encode the specific flora and fauna native to the region, which often serve as highly sensitive indicators of specific climate micro-zones and regional similarities.
- **Geographic Enrichment: ESA LandCover[41].** To capture the detailed physical characteristics of the disaster zones, we integrated the European Space Agency's (ESA) 10m-resolution LandCover data. By calculating zonal statistics for the 10km buffer around each epicenter, we extracted the physical characteristics of the land. This layer provides the percentage distribution of primary land-cover classes, such as tree cover, built-up urban areas, and mangroves.

G.2 Ontological Schema and Neo4j Modeling

All extracted data points were instantiated in a Neo4j graph database. By modeling these relationships structurally, we can ask complex questions about environmental vulnerability and historical context that raw, continuous meteorological data cannot answer.

The ontological schema for the environmental domain centers on the **Disaster** event as the primary temporal anchor. This event is spatially linked to a **Region** node, which in turn acts as a node

connecting to localised biological (**Species**) and geographic (**LandCoverClass**) nodes. The structural design of this ontology is illustrated in Figure 9.

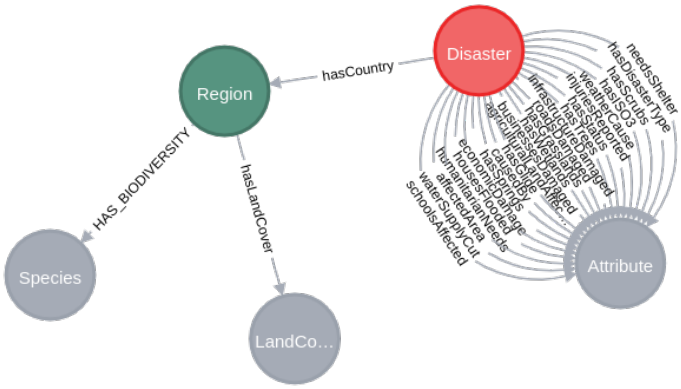


Figure 9: The ontological schema of the Environmental KG ($\mathcal{G}_E$). The schema defines the core nodes and directed edges used to map flood events through their regional hubs to their specific ecological and physical surroundings within Neo4j.

Table 9 provides an inventory of the relationships currently within the graph, showcasing the semantic diversity of the dataset. This setup ensures that if a generative model predicts a flood in a specific geographic grid, our framework can retrieve historical events that share nearly identical environmental and topological contexts.

Table 9: KG Relationship Inventory and Semantic Diversity. We report the total edge count and unique target entities to demonstrate the breadth of the environmental context.

Source	**Relationship**	**Target**	**Count**	**Unique**
Region	Has_Biodiversity	Species	18,065	9,910
Region	hasLandCover	LandCoverClass	432	11
Disaster	hasCountry	Region	1,592	149
Disaster	hasDisasterType	Attribute	1,523	3
Disaster	hasGlide	Attribute	1,468	1,468
Disaster	causedBy	Attribute	1,197	471
Disaster	weatherCause	Attribute	896	249
Disaster	hasScrubs	Attribute	751	29
Disaster	hasWetlands	Attribute	744	20
Disaster	hasGrasslands	Attribute	703	22
Disaster	affectedArea	Attribute	665	479
Disaster	housesFlooded	Attribute	462	311
Disaster	needsShelter	Attribute	407	11
Disaster	economicDamage	Attribute	176	116

Table 10 details the internal properties and metadata fields associated with each node and edge type. These properties allow for precise filtering and the calculation of structural ranking scores within our pipeline.

Ultimately, this Neo4j structure allows for Direct Matching. Because each prediction has a location and time, we can query the graph to find the exact environmental footprint of the predicted event and compare it to historical records without the need for contrastive mapping.

Table 10: Entity and Relationship Property Schema of the Knowledge Graph $\mathcal{G}_E$.

Node Label	**Internal Properties (Metadata Fields)**
Disaster	name, date, status, latitude, longitude, iso3, displacedPeople, housesCollapsed, powerOutage, waterSupplyCut, mountainsCount, watersCount, woodsCount
Region	name, latitude, longitude, springsCount, treesCount
Species	scientificName, category, inaturalistId, wikipediaUrl, photoUrl
LandCoverClass	name
Edge Type	Relationship Metrics (Edge Properties)
hasLandCover	area_km2
Has_Biodiversity	observations

H Environmental Evaluation Framework

H.1 Experimental Setup and Meteorological Data

To evaluate our attribution framework on physical weather events, we leverage the pre-computed predictions of the Pangu-Weather model hosted by WeatherBench 2 [42]. Pangu-Weather optimises for both surface and upper-air variables by training on the ERA5 reanalysis dataset.

Temporal Scope and Forecast Intervals. Our evaluation leverages predictions from WeatherBench 2 dataset spanning a five-year period from 2018, to 2022. For each documented extreme flood event within this period, our primary experiments strictly utilise the 24-hour forecast target corresponding to the event's peak intensity. This specific lead time was chosen to effectively capture short-term, high-impact flood dynamics without the signal degradation.

Variables and Pressure Levels. The full Pangu-Weather output in WeatherBench 2 contains a comprehensive suite of atmospheric variables at a global $0.25^\circ \times 0.25^\circ$ spatial resolution. The available surface variables include 10m U and V wind components, 10m wind speed, 2m temperature, and mean sea-level pressure. The available upper-air variables include geopotential, specific humidity, temperature, U and V wind components, and wind speed. These upper-air variables are distributed across 13 distinct pressure levels ranging from 1000 hPa to 50 hPa.

Selected Features for Flood Analog Retrieval Case Study. To optimise the Latent Distribution Engine for our specific environmental case study, we subset this massive output tensor to isolate the physical drivers most relevant to extreme precipitation and surface flooding. We restrict the surface variables to 2m temperature and mean sea-level pressure. For the upper-air profile, we utilise temperature, specific humidity, geopotential, and the U/V wind components exclusively at the 1000, 850, 500, and 250 hPa levels. This targeted subset is physically justified: it captures essential low-level moisture transport and boundary layer dynamics (1000 and 850 hPa), mid-level steering flows (500 hPa), and upper-level divergence (250 hPa) necessary to attribute heavy rainfall events, effectively reducing computational overhead while preserving the primary meteorological signal.

H.2 Multi-Modal Representation Learning

Environmental Representation Learning. To bridge the gap between high-dimensional atmospheric tensors and discrete domain knowledge, we establish two specialised encoding pipelines.

Latent Representation (3D Denoising Autoencoder): For the continuous meteorological modality, we developed a 3D ResNet-based Denoising Autoencoder. The encoder architecture utilises a series of 3D convolutional residual blocks to process the multi-level atmospheric cubes (surface and upper-air variables). We trained the model for 100 epochs on a dataset comprising ERA5 reanalysis data [44] centered on extreme flood events and regional historical baselines. The resulting 512-dimensional embeddings provide a robust representation of the latent space for both ERA5 ground truth and Pangu-Weather predictions.

Structural Representation (Node2Vec): The environmental metadata is encoded via the Node2Vec algorithm [32] implemented within the Neo4j Graph Data Science (GDS) library. By utilising biased random walks over $\mathcal{G}_E$, we extract high-dimensional embeddings that capture the topological relationships between geographical regions, ecological biomes, and land-cover signatures. This ensures that the structural prior $P(\mathbf{X}_j)$ reflects the physical and ecological rarity of a region's position within the global environmental network, preventing common geographical hubs from biasing the analog retrieval.

I Computational Resources and Efficiency

To ensure the reproducibility of our results while maintaining a low computational footprint, we utilise a combination of pre-computed baseline assets and efficient graph-based infrastructure.

I.1 Domain-Specific Assets

- **Artistic Domain:** We utilised pre-generated images and model weights provided by the D-TRAK baseline. This allowed us to evaluate attribution performance without the computational cost of retraining diffusion models from scratch and also ensure a fair comparison with baseline methods.
- **Environmental Domain:** We leveraged pre-computed Pangu-Weather predictions directly from the WeatherBench 2 benchmark, ensuring our physical-consistency evaluations are grounded in standardised meteorological AI outputs.

I.2 Graph Infrastructure and Embedding Generation

For the construction, hosting, and analysis of our domain-specific KGs, we utilised the Neo4j Desktop. A key advantage of this setup is that graph management and embedding generation did not require GPU access, relying instead on standard CPU compute.

To generate embeddings natively based on our graph's topology and properties, we employed the Neo4j Graph Data Science (GDS) Library. The GDS library provides highly optimised algorithms that turn nodes and relationships into vector representations without leaving the database, allowing for efficient representation learning within the KG environment.

I.3 Hardware and Implementation

- **Local Compute:** Local training for our contrastive MLPs and domain-specific encoders was performed on a single node equipped with one **NVIDIA RTX 5000 Ada Generation GPU (32GB VRAM)**.
- **Software Environment:** Our framework was implemented using PyTorch, the Diffusers library, and the D-TRAK requirements.

J Asset Licenses and Attribution

We provide the licenses for all external assets used in this study in Table 11.

Table 11: Licenses and sources for external datasets and models.

Asset	License	Sources
D-TRAK Codebase	MIT License	[17]
ArtBench Dataset	MIT License	[29]
Environmental Data	MIT License	[25]
WeatherBench 2	Apache 2.0	[42]
Stable Diffusion	RAIL-M	[33]

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes].

   Justification: We introduce a probabilistic framework in Section 3 that enables training data attribution without requiring internal model parameters. The claims regarding a model-agnostic, domain-aware attribution framework are supported by quantitative LDS results in the artistic domain (Section 4.1.3) and a Geographic Precision Analysis at top-K ranks in the environmental domain (Section 4.2.3). We explicitly state the scope, the fact that we do not require access to model internals and the assumption of having access to a representative domain KG.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes].

   Justification: We discuss performance dependency on encoder quality and KG density (Section 3.4), the lack of established environmental benchmarks (Section 4.2.2), computational scalability compared to gradient-based methods, and potential membership inference risks (Section 6).

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes].

   Justification: The mathematical foundation of our framework is derived from a probabilistic formulation detailed in Section 3. We explicitly state the assumptions regarding generative necessity and the counterfactual evidence gap, providing the relevant derivations and citations to support the attribution decomposition.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes].

   Justification: We provide a GitHub repository containing all code for data collection, KG construction, and evaluation scripts. All the details that are necessary to reproduce the main experimental results of the paper in both domains are also described in the Appendixes. For the artistic domain, we utilise models and generated images provided by the authors of our primary baselines to ensure direct comparability. For the environmental domain, we rely on the publicly available WeatherBench 2 benchmark for Pangu-Weather predictions and ERA5 ground truth data. Representative data samples are included in the repository and also some examples in the Appendixes.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example

(a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes].

   Justification: We provide a GitHub repository link in the supplemental material containing all source code for our framework, KG construction, and evaluation pipelines. The repository includes a comprehensive README with instructions for environment setup and the exact commands needed to reproduce the results of the LDS evaluation. While the full environmental datasets are hosted on external storage due to size constraints, we provide representative samples and clear instructions for accessing the public WeatherBench 2 and ERA5 datasets used in our study.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes].

   Justification: The core experimental setup is described in Section 4. Detailed training and testing specifications, including data splits, hyperparameter selection, and optimiser configurations, are provided in the Appendix for each domain. Specifically, Appendix C details the LoRA fine-tuning and artistic setups, while Appendix H covers the environmental benchmarking data and encoding parameters.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes].

   Justification: All LDS in our tables are reported as the mean value accompanied by 95% confidence intervals (Section D.4). These intervals were calculated using a bootstrapping procedure with 64 resamples, following the standard evaluation protocol for data attribution established in prior work like D-TRAK.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes].

   Justification: We specify the hardware and external assets used in our experiments in Appendix I. For the artistic domain, we utilised pre-generated images and model weights provided by the D-TRAK baseline, avoiding the need for retraining. For the environmental domain, we leveraged Pangu-Weather predictions directly from the WeatherBench 2 benchmark. Local training for our contrastive MLPs and encoders was performed on a single NVIDIA RTX 5000 Ada Generation GPU (32GB VRAM).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes].

   Justification: We have reviewed the NeurIPS Code of Ethics and confirm that our research conforms to its principles. Our study utilizes publicly available, established benchmarks and focuses on enhancing model transparency and accountability through data attribution.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes].

    Justification: We discuss broader impacts in Section 1. Positive impacts include protecting artist copyrights by identifying influential training samples and improving the reliability of environmental models for disaster prevention. A potential negative impact is that attribution/analog retrieval outputs could be misinterpreted as definitive proof of training set usage or causality; we recommend using the method alongside complementary audits.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A].

Justification: Our work introduces a diagnostic framework for training-data attribution and does not involve the release of new generative models or primary datasets with a high risk for misuse. We utilise established, publicly available benchmarks such as WeatherBench 2 and ArtBench and we collect domain knowledge from well defined and popular sources extensively refered in the paper.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes].

Justification: We cite the original authors of all datasets used, including ERA5, WeatherBench 2, and ArtBench. All software libraries and pre-trained models (e.g., Stable Diffusion, Pangu-Weather) are credited in our references and Appendix. We provide a detailed list of asset versions, URLs, and specific licenses, including the MIT License for the D-TRAK code, ArtBench, and the environmental datasets, and the Apache License for WeatherBench, in the Appendix J to ensure full compliance with their respective terms of use.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes].

Justification: We provide comprehensive documentation for the **FrED** framework and the newly constructed domain-specific KGs in the Appendix and our anonymised GitHub repository. This includes a clear repository structure, environment requirements, KG schema, and a license file (Apache License 2.0) to facilitate reproducibility and future research.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A].

    Justification: Our research does not involve crowdsourcing or human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A].

    Justification: Our research does not involve crowdsourcing or human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A].

    Justification: LLMs were used only as a writing aid to improve the clarity and grammar of the manuscript and did not play a role in the core scientific method, data collection, or experimental analysis.

    Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.