FrED: External Data Influence Estimation via Domain Knowledge Graph Grounding

Theodoros Aivalis[1,2] Iraklis A. Klampanos[1] Antonis Troumpoukis[2] Joemon M. Jose[1]

[1]University of Glasgow, UK
[2]National Centre for Scientific Research "Demokritos", Greece

Motivation

- AI Models are widely used — but often as **black boxes**.
- Attribution is unclear, raising **legal, ethical and trust concerns**, from **copyright** in art to **physical precedents** in weather forecasting.
- **Current methods:** gradient-based attribution needs **model internals**; black-box similarity ignores **domain structure**.

Can we trace "influence" without opening the black box?

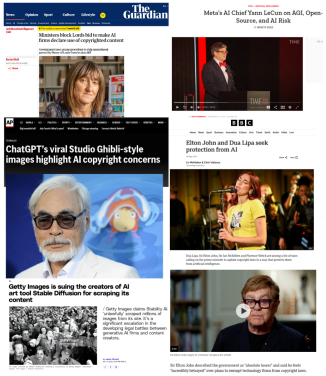


Figure: Legal and ethical concerns around the use of training data in generative AI models.

The Bayesian Influence Framework

The Bayesian Proxy

We approximate the target $\mathbf{C}$ as a weighted combination of the entire training corpus:

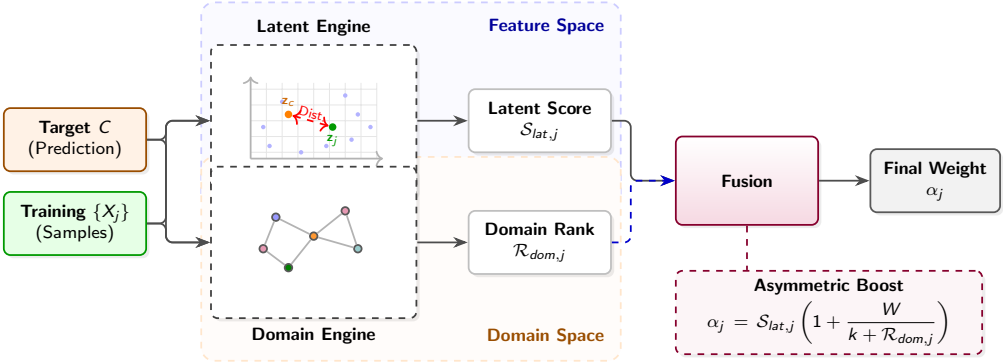
$$\mathbf{C} \approx \sum_{j=1}^{K} P(\mathbf{X}_j \mid \mathbf{C})\mathbf{X}_j = \sum_{j=1}^{K} \alpha_j \mathbf{X}_j$$

Bayesian Expansion for Influence Ranking

Using Bayes' theorem, we compute the influence weight α_j:

$$\alpha_j \propto \underbrace{P(\mathbf{C} \mid \mathbf{X}_j)}_{\text{Latent Support / Similarity}} \times \underbrace{P(\mathbf{X}_j)}_{\text{Informational Scarcity}}$$

The FrED Framework: Dual-Space Attribution



Latent Engine
Ranks samples by feature similarity to the target, down-weighting redundant ones.

Domain Engine
Ranks samples by their KG connections, favouring specific links over generic hubs.

Asymmetric Boost
The latent score leads; the KG rank boosts the samples it supports — samples without KG links are never discarded.

The Artistic Domain: Building the Knowledge Graph

- **ArtBench-10** dataset.
- Metadata from **WikiArt**, **Wikipedia** and **DBpedia**.
- Metadata at the **painting** and **artist** level.
- Generated images are projected into the KG space via **contrastive alignment**.

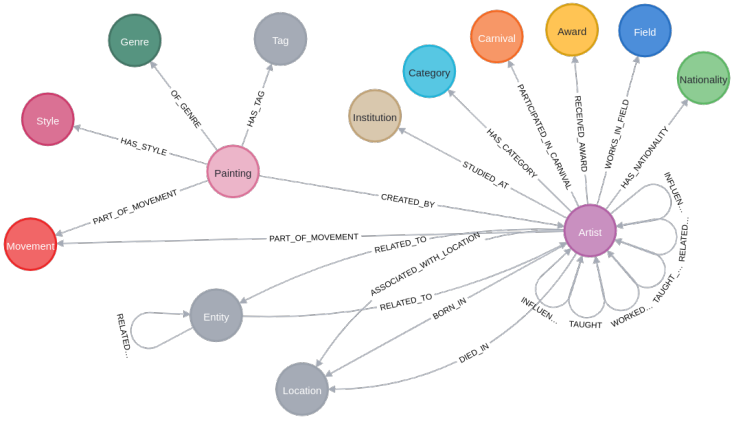


Figure: Ontological schema of the Artistic KG.

Evaluation: Linear Datamodeling Score (LDS)

Setup (D-TRAK protocol)

- Stable Diffusion 256×256, LoRA fine-tuned on **ArtBench-2** and **ArtBench-5**, using the same 1,000 generated targets as the baselines.

LDS: is the attribution counterfactually right?

- Retrain the model on 64 random 50% subsets of the training data.
- Spearman ρ between the *actual* change in the output and the *predicted* (summed) attribution scores.

Only the attribution method changes — model, data and targets are identical.

Experimental Results

Table: LDS results (%) on ArtBench-2 and ArtBench-5 against black-box and parametric baselines.

	ArtBench-2		ArtBench-5	
Method	**Validation**	**Generation**	**Validation**	**Generation**
Black-Box / Non-Parametric Methods				
Raw pixel (cosine)	2.5 ± 1.2	2.6 ± 1.5	1.97 ± 0.41	3.22 ± 0.78
CLIP similarity (cosine)	8.9 ± 1.3	8.8 ± 2.2	6.57 ± 0.44	6.63 ± 1.14
FrED (Ours)	$\mathbf{29.1 \pm 0.65}$	$\mathbf{21.3 \pm 1.65}$	$\mathbf{28.9 \pm 0.35}$	$\mathbf{18.4 \pm 1.05}$
White-Box / Parametric Methods				
TracInCP	9.69 ± 0.49	6.36 ± 0.93	5.33 ± 0.37	4.34 ± 0.84
TRAK	12.26 ± 0.42	7.78 ± 0.59	9.79 ± 0.33	8.79 ± 0.59
D-TRAK	27.61 ± 0.49	24.16 ± 0.67	22.84 ± 0.37	21.56 ± 0.71
DAS	$\mathbf{37.96 \pm 0.64}$	$\mathbf{30.81 \pm 0.31}$	$\mathbf{35.33 \pm 0.49}$	$\mathbf{31.74 \pm 0.75}$

FrED is the best **black-box** *method and is competitive with gradient-based estimators —* **without model access**.

Ablation Study

Table: Ablation study comparing foundational vs. full framework configurations (LDS, %).

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
	$p(c\|x_j)^{10}p(x_j)$	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
	$p(c\|x_j)^{10}p(x_j)$	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
	$p(c\|x_j)^{10}p(x_j)$	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
	$p(c\|x_j)^{10}p(x_j)$	28.00 ± 0.70	19.40 ± 1.55	26.40 ± 0.45	15.40 ± 1.10
Proposed	**Asymmetric Boost**	$\mathbf{29.10 \pm 0.65}$	$\mathbf{21.30 \pm 1.65}$	$\mathbf{28.90 \pm 0.35}$	$\mathbf{18.40 \pm 1.05}$

Linear Fusion (50/50)
Equal weighting lets KG noise degrade the visual signal — below visual-only.

Weighted Fusion (70/30)
Favouring the visual space helps, but still does not surpass visual-only.

Asymmetric Boost
The only fusion that improves over visual-only in **all** settings.

The Environmental Domain: Flood Forecasting Feasibility Study

Weather models cannot be retrained for counterfactuals, so we test **domain-grounded analog retrieval** on **237 floods** forecast by **Pangu-Weather**.

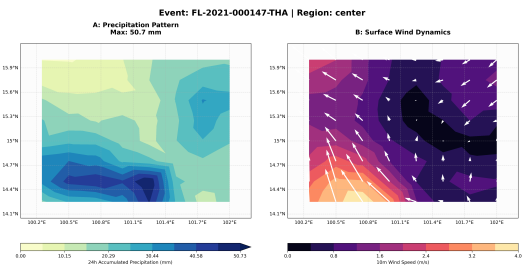


Figure: (a) Atmospheric Physics (Forecast)

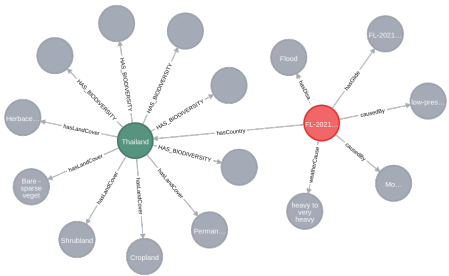


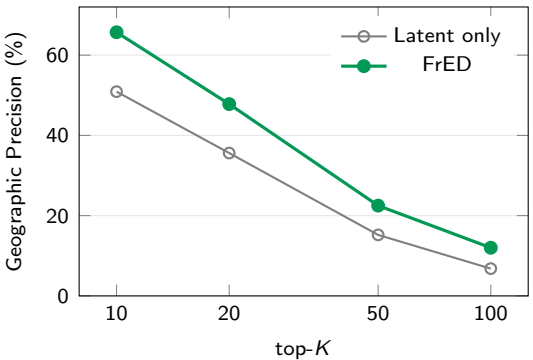
Figure: (b) Structural Context (Knowledge Graph)

Target C
24h forecast at peak intensity, regional patch around the event.

Latent Space
3D Autoencoder over surface & upper-air variables.

Domain Space
Environmental KG queried by **direct geospatial matching**.

Environmental Results: Sharper Localisation



- **Prec@10: 50.9% → 65.7%**, better at every K.
- Cross-country analogs stay **ecologically consistent**:
  - 42.1% land-cover overlap
  - 46.2% species similarity
- e.g. Gambia ↔ Burkina Faso, Paraguay ↔ Chile.

Takeaways & Future Directions

Takeaways

- Black-box attribution grounded in **domain knowledge**.
- Best black-box LDS on ArtBench, competitive with gradient-based methods.
- Transfers to Earth observation as analog retrieval.

Future Directions

- **Expert collaboration:** co-design KGs and validate attributions.
- **New evaluation methods** for both domains.
- **No training data:** approximate with KG samples.
- Extend FrED to new **modalities & domains**.

Thank you!

Code https://github.com/teoaivalis/FrED

Demo https://huggingface.co/spaces/teoaivalis/fred_art_demo

Mail teoaivalis@iit.demokritos.gr





