FAME: Forecasting Academic Impact via Continuous-Time Manifold Evolution

Jianrong Ding, Jianyuan Zhong, Zhengyan Shi, Qiang Xu[†]
Department of Computer Science and Engineering
The Chinese University of Hong Kong
{jrding25,jyzhong,zyshi25,qxu}@cse.cuhk.edu.hk

Abstract

Large Language Models (LLMs) are increasingly used to brainstorm and evaluate research ideas, yet assessing such judgments is fundamentally difficult because the true impact of a new idea may take years to emerge. We address this challenge by using the impact forecasting of human-authored manuscripts as a verifiable proxy task. In a prospective forecasting study, we find that frontier LLMs fail to reliably distinguish high-impact papers from ordinary publications, suggesting that static text-based judging is insufficient for scientific evaluation. To address this limitation, we propose **FAME** (Forecasting Academic Impact via Continuous-Time Manifold Evolution), a spatiotemporal framework for modeling the dynamic trajectories of scientific topics. FAME projects papers into a dynamic latent space informed by textual features and a verified knowledge-flow graph, learning geometric constraints that align impactful manuscripts with the forward momentum of their fields. Experiments on 3,200 arXiv papers across three fast-evolving subfields show that FAME consistently and substantially outperforms state-of-the-art LLM evaluators in prospective multidimensional impact forecasting. Furthermore, integrating FAME's dynamic geometric signals into LLMs significantly improves their forecasting performance. These results support manuscript impact forecasting as a useful, measurable proxy benchmark and position FAME as a strong, trajectory-aware foundation for automated scientific evaluation.

1 Introduction

The integration of artificial intelligence into scientific discovery [19, 35] is accelerating rapidly. Large Language Models (LLMs) are now routinely deployed to navigate literature, brainstorm hypotheses, and assess emerging research directions [19, 28]. As this combinatorial space of possibilities expands, robust automated evaluation becomes essential. Currently, the dominant paradigm relies on LLMs acting as zero-shot or few-shot judges to score novelty and potential impact [19, 28, 20]. Yet, despite their remarkable fluency in textual critique, a fundamental question remains: *Can LLM evaluators actually distinguish genuinely impactful research from ordinary publications?*

To investigate this, we conducted a rigorous prospective forecasting study on state-of-the-art LLMs [26, 17]. Testing an AI's ability to evaluate novel ideas presents a fundamental methodological challenge: true scientific impact takes much time to materialize, meaning there is no immediate ground truth to verify if the AI's judgment of a brand-new idea is correct. To overcome this, we utilize the impact forecasting of human-authored manuscripts as a verifiable proxy task. By rolling back the clock, treating historical papers as brand-new ideas relative to a temporal cutoff, we can validate an AI's predictions against the papers' actual, realized impact. This establishes impact forecasting as a rigorous proxy benchmark for scientific evaluation, operating on the premise that the

[†]Corresponding author

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

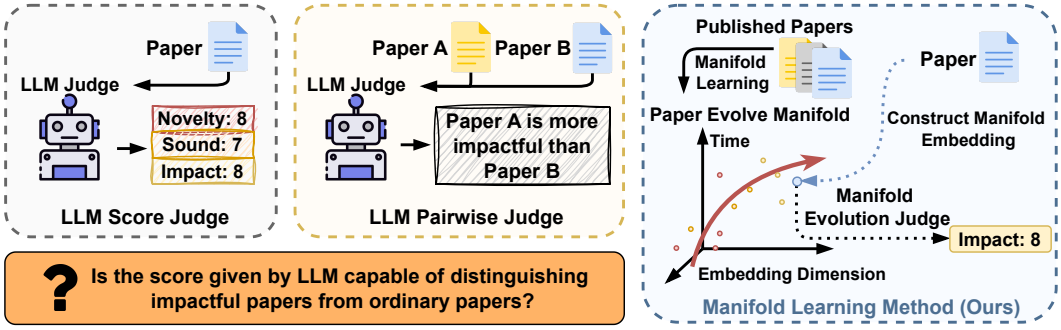


Figure 1: Comparison of standard LLM evaluators with our proposed method, which evaluates research manuscripts based on their geometric alignment with a field's evolutionary momentum.

ability to forecast the trajectory of a fully formalized manuscript serves as a necessary foundation for effectively assessing raw hypotheses.

Our findings reveal significant limitations in the current baseline. When tasked with predicting the multidimensional impact of published research, frontier LLMs, and even domain-specific models like SciJudge [30], exhibit highly volatile predictive correlations. This suggests that static, text-based judging is insufficient for scientific evaluation. Standard LLMs evaluate text in an isolated semantic vacuum; they overlook the macroscopic spatiotemporal dynamics of a scientific field, missing how a piece of research geometrically aligns with the evolutionary momentum of its underlying topic.

To address this limitation, we argue that impact forecasting must move beyond static text to model the temporal evolution of scientific knowledge. We propose **FAME** (Forecasting Academic Impact via Continuous-Time Manifold Evolution), a continuous-time spatiotemporal manifold learning framework. FAME operates in three streamlined stages. First, recognizing that raw citation networks are notoriously noisy and often superficial [10, 16], we construct a high-quality inspiration graph using a retrieve-and-verify pipeline, leveraging LLMs strictly to validate genuine foundational knowledge flow between papers. Second, a dual-branch neural architecture maps discrete papers into a continuous manifold while learning dynamic topic spines that track the evolutionary trajectories of sub-disciplines. Finally, we sculpt this latent space using geometric objectives that anchor papers to their temporal context and force high-impact historical manuscripts to align with their field's forward momentum. By structuring the latent space around evolutionary momentum, evaluating an unseen manuscript becomes highly efficient: its predicted impact is derived directly from the geometric alignment between its positional residual and the instantaneous momentum of its corresponding topic spine. The main contributions of our work are summarized as follows:

- **Empirical Exposure of LLM Limitations:** We systematically quantify the inability of current LLM-as-a-judge paradigms to reliably forecast scientific impact, identifying a critical bottleneck in AI-driven evaluation.
- **Novel Spatiotemporal Framework:** We introduce **FAME**, a continuous-time manifold architecture that tracks the evolutionary trajectory of research topics, shifting the evaluation paradigm from static semantic scoring to dynamic geometric alignment.
- **Robust Predictive Performance:** Across extensive sliding-window evaluations on 3,200 arXiv papers, FAME achieves a Spearman rank correlation exceeding 0.5, consistently and significantly outperforming state-of-the-art LLM baselines. Furthermore, integrating FAME's manifold scores into standard LLMs significantly boosts their performance.

2 Preliminary

Problem Definition. Let $\mathcal{P} = \{p_1, ..., p_N\}$ denote a corpus of N scientific papers published over a continuous time horizon. Each paper p_i is associated with a publication timestamp t_i and a realized ground-truth impact weight w_i. By computing w_i using bibliometric signals (e.g., citations, GitHub stars) aggregated at the study's global conclusion (T_{end}), we frame the task as a calibrated prospective forecast: projecting early manuscript features to their ultimate long-term impact. To

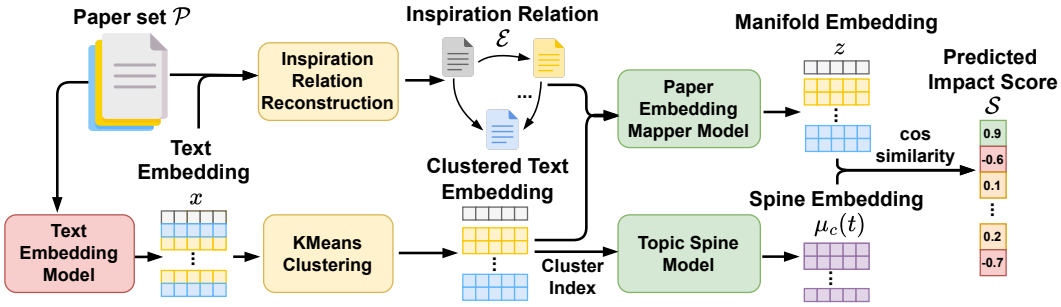


Figure 2: Overall architecture of FAME. First, we construct a high-quality inspiration graph. Second, a dual-branch neural architecture maps discrete papers into a continuous manifold. Finally, we predict the impact score using manifold embeddings.

simulate a realistic forecasting scenario, we temporally partition the corpus based on a threshold timestamp T. Specifically, we define a training set comprising strictly historical publications, $\mathcal{P}_{train} = \{p_i \in \mathcal{P} | t_i \leq T\}$, and a temporally disjoint test set representing future publications, $\mathcal{P}_{test} = \{p_i \in \mathcal{P} | t_i > T\}$. The overarching objective of this task is to leverage the historical data in $\mathcal{P}_{train}$ to learn a model $\mathcal{M}$ capable of estimating the future realized significance of emerging research. Formally, the model aims to maximize the Spearman rank correlation ρ_s between the predicted impact weights $\hat{\mathbf{w}}_{test}$ and the true impact weights $\mathbf{w}_{test}$, evaluated on the test set:

$$\max_{\mathcal{M}} \rho_s(\boldsymbol{w}_{test}, \hat{\boldsymbol{w}}_{test}), \quad \text{where} \quad \hat{\boldsymbol{w}}_{test} = \mathcal{M}(\mathcal{P}_{test}) \tag{1}$$

Impact Weight Formulation. To construct the ground-truth impact weight w_i of each paper, we aggregate multiple real-world signals that capture different facets of a paper's influence. For each paper p_i, let $s_{i,m}$ denote its raw score across $M = 4$ distinct metrics: GitHub stars (software utility), citation counts (broad academic attention), influential citations (foundational impact), and Altmetric scores (social and media traction). To account for the heavy-tailed distribution typical of scientific impact metrics [22, 3], we apply a scale factor a_m and log-normalization with base b_m to each individual signal. The final composite impact weight w_i is formulated as the sum of these four normalized scores: $w_i = \sum_{m=1}^{4} \log_{b_m}(a_m s_{i,m} + 1)$.

3 Methodology

To address the problem of scientific impact prediction, we propose FAME, a continuous-time framework that models the evolution of scientific topics as dynamic trajectories within a latent space. As shown in Figure 2, the framework operates in three sequential stages: **(1)** distilling noisy academic citations into a high-quality, verified inspiration graph; **(2)** mapping raw textual features into a spatiotemporal manifold using continuous representation learning; and **(3)** leveraging the geometric properties of the learned manifold to infer the future impact of unseen papers.

3.1 Inspiration Graph Construction

Semantic Clustering. We first utilize a pre-trained text embedding model to transform the raw textual content of the corpus into dense semantic representations, denoted as $\{x_i\}_{i=1}^{N}$. To capture the underlying sub-disciplines, we apply KMeans clustering on these embeddings, assigning a base topic label $c_i \in \{1, \ldots, K\}$ to each paper. Each cluster represents a localized scientific topic within the broader academic corpus.

Retrieve-and-Verify Graph Construction. Raw citation networks are notoriously noisy; many citations are superficial and do not reflect genuine intellectual inspiration [10, 16]. To isolate meaningful knowledge transmission, we construct a high-quality inspiration graph $\mathcal{E}$ using a rigorous retrieve-and-verify pipeline. For a given target paper p_j, we first retrieve a candidate precursor pool $\{p_i\}$ using three strict heuristic constraints: **(1) Citation Link:** p_i must be explicitly present in the bibliography of p_j. **(2) Semantic Similarity:** The cosine similarity between the embeddings of the two papers must satisfy a minimum threshold: $\cos(x_i, x_j) > \tau_{sim}$. **(3) Temporal Precedence:** The

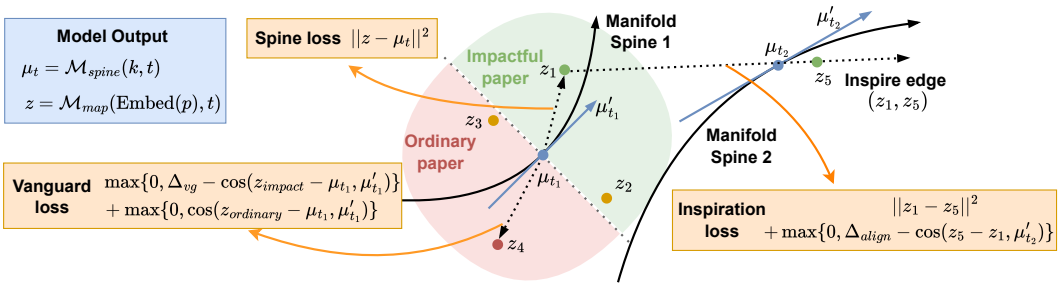


Figure 3: Visualization of the manifold training dynamics and the three applied geometric constraints. The objective functions sculpt the latent space topology by using a spine binding loss to anchor papers to their temporal context, an inspiration loss to ensure knowledge flow alignment, and a vanguard loss to geometrically separate high-impact historical papers from marginal ones.

publication time difference must enforce strict chronological order: $t_j - t_i > \Delta_{days}$. After filtering for the top-K semantically similar candidates, we deploy an LLM to explicitly verify whether p_j was fundamentally inspired by the methodology or findings of p_i. Only LLM-verified pairs (i, j) are retained in the final directional edge set $\mathcal{E}$. By utilizing the LLM solely to evaluate the direct textual evidence of inspiration between two localized documents, we effectively leverage its established strengths in textual critique to sanitize the noisy bibliographic data. This produces a refined graph capturing foundational, idea-level relationships rather than mere bibliographic references.

3.2 Spatiotemporal Manifold Architecture

As shown in Figure 3, we employ a dual-branch neural architecture to map discrete papers into a continuous manifold space and model the overarching temporal development of the topics.

Paper Embedding Mapper ($\mathcal{M}_{map}$): This module parameterized by a Multi-Layer Perceptron (MLP) projects individual papers from the semantic space into the spatiotemporal manifold. It takes the original semantic embedding x_i and the publication timestamp t_i as inputs. To prevent the 1D temporal scalar from being numerically overwhelmed by the high-dimensional text features, we project time into a higher-dimensional space using a continuous temporal encoding function, $\text{TimeEncode}(t_i) \in \mathbb{R}^{d_{time}}$. The features are subsequently fused and projected to yield the final latent spatiotemporal embedding z_i:

$$z_i = \mathcal{M}_{map}\left(x_i, \text{TimeEncode}(t_i)\right) \tag{2}$$

Topic Spine Model ($\mathcal{M}_{spine}$): This MLP-based module generates the continuous evolutionary trajectory, namely *spine*, for each topic cluster. We assign a learnable, static base representation $E_k \in \mathbb{R}^{d_{topic}}$ to each topic $k \in \{1, \ldots, K\}$. The continuous trajectory of the topic is then generated dynamically by passing the base embedding and continuous time t into the model:

$$\mu_k(t) = \mathcal{M}_{spine}\left(E_k, \text{TimeEncode}(t)\right) \tag{3}$$

3.3 Manifold Training Dynamics

To geometrically structure the latent space, we optimize the topology using three distinct objective functions at the micro and macro levels. The visualization of the losses is shown in Figure 3.

Spine Binding Loss: This objective acts as a centripetal structural constraint, anchoring paper embeddings to the evolutionary trajectory of their respective topics at the exact time of publication, thereby preventing arbitrary dispersion in the latent space:

$$\mathcal{L}_{spine} = \sum_{i=1}^{N} \|z_i - \mu_{c_i}(t_i)\|_2^2 \tag{4}$$

Inspiration Loss: To model directional knowledge flow for an inspiration relation $(i, j) \in \mathcal{E}$, this loss aligns the geometric knowledge jump $(z_j - z_i)$ with the inspired topic's forward momentum

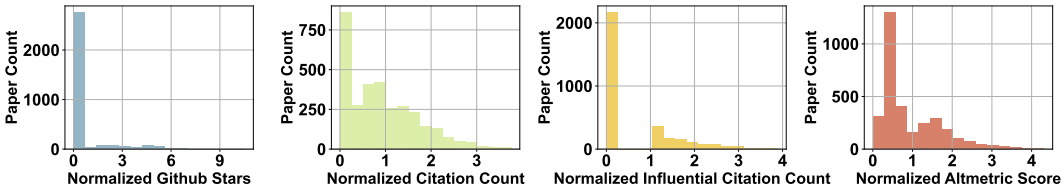


Figure 4: The heavy-tailed distributions of the four raw log-normalized ground-truth metrics used to calculate the composite impact weight for the 3,200 arXiv papers in the dataset.

$\mu'_{c_j}(t_j)$, approximated via finite difference as $\mu'_{c_j}(t_j) \approx \mu_{c_j}(t_j) - \mu_{c_j}(t_j - \Delta t)$. The total loss combines an L2 term for spatial proximity (pulling inspiring and inspired papers closer) and a cosine margin term to ensure the innovation direction aligns perfectly with the field's future trajectory.

$$\mathcal{L}_{inspire} = \sum_{(i,j)\in\mathcal{E}} \max\left(0, \Delta_{align} - \cos\left(z_j - z_i, \mu'_{c_j}(t_j)\right)\right) + \sum_{(i,j)\in\mathcal{E}} \|z_i - z_j\|_2^2 \tag{5}$$

Vanguard Loss: To integrate ground-truth impact w_i, we sculpt the manifold using a localized baseline to isolate genuine breakthroughs from broader temporal drift. For paper p_i, we define a spatiotemporal neighborhood $\mathcal{S}_i = \{j \mid c_j = c_i \text{ and } |t_j - t_i| \leq \tau_{time}\}$ and compute its median impact $m_i = \text{median}(\{w_j \mid j \in \mathcal{S}_i\})$. Using a dynamic margin $\Delta_{vg}^{(i)} = \Delta_{base} + (1 - \Delta_{base})\tilde{w}_i$, where $\tilde{w}_i$ is min-max normalization of w_i to $[0, 1]$, high-impact vanguard papers ($w_i \geq m_i$) are geometrically encouraged to align with the topic's forward momentum $\mu'_{c_i}(t_i)$. Conversely, low-impact papers ($w_i < m_i$) are penalized if their positional residual projects along this derivative:

$$\begin{aligned}\mathcal{L}_{vanguard} = &\sum_{i:\tilde{w}_i \geq m_i} \max\left(0, \Delta_{vg}^{(i)} - \cos\left(z_i - \mu_{c_i}(t_i), \mu'_{c_i}(t_i)\right)\right) \\ &+ \sum_{i:\tilde{w}_i < m_i} \max\left(0, \cos\left(z_i - \mu_{c_i}(t_i), \mu'_{c_i}(t_i)\right)\right)\end{aligned} \tag{6}$$

The total objective function is a weighted sum of these three losses:

$$\mathcal{L}_{total} = \alpha\mathcal{L}_{spine} + \beta\mathcal{L}_{inspire} + \gamma\mathcal{L}_{vanguard} \tag{7}$$

3.4 Downstream Impact Inference

Inference for new papers in FAME is highly efficient. For an unseen paper p_{new} at time t_{new}, we assign its semantic embedding x_{new} to a topic cluster c_{new} via KMeans centroids, and the mapper model projects it to a latent representation z_{new}. To predict impact, we evaluate the paper's positional residual $(z_{new} - \mu_{c_{new}}(t_{new}))$ against its field's instantaneous momentum $\mu'_{c_{new}}(t_{new})$. Crucially, we quantify this using angular alignment rather than spatial distance. Due to the curse of dimensionality, Euclidean distances converge in high-dimensional spaces [2], whereas angular metrics remain robust. Thus, the final predicted impact score $\mathcal{S}$ is calculated as their cosine similarity:

$$\mathcal{S}(p_{new}) = \cos\left(z_{new} - \mu_{c_{new}}(t_{new}), \mu'_{c_{new}}(t_{new})\right) \tag{8}$$

A score near 1 indicates strong directional consistency with the field's future evolutionary trajectory, naturally translating to a high predicted impact weight.

4 Experiments

Dataset and Metrics. We curate a temporally stamped dataset of 3,200 arXiv papers from three rapidly evolving subfields: Time Series Forecasting, Sparse Attention, and Image Generation. To quantify genuine multi-dimensional influence, we compute a composite ground-truth impact weight w_i for each paper by log-normalizing and aggregating its GitHub stars, citation counts, influential citations, and Altmetric score. The distribution of the four log-normalized ground-truth impact metrics utilized in the dataset is shown in Figure 4.

Table 1: Prospective forecasting performance across an 18-month sliding window evaluation from June 2024 to November 2025. Performance is measured by the Spearman rank correlation ρ_s between predicted and ground-truth composite impact weights. The experiments are carried out three times, and the standard deviation of the performance is reported. Best results are shown in bold.

Model	2024.06	2024.07	2024.08	2024.09	2024.10	2024.11	**Avg**
XGBoost	0.148 ± 0.027	0.321 ± 0.024	0.301 ± 0.030	0.105 ± 0.035	0.182 ± 0.039	-0.067 ± 0.037	0.165
SVR	0.151 ± 0.005	0.225 ± 0.017	0.227 ± 0.001	0.394 ± 0.004	0.281 ± 0.007	0.320 ± 0.003	0.266
Transformer	0.115 ± 0.053	0.410 ± 0.037	0.397 ± 0.043	0.375 ± 0.013	0.243 ± 0.036	0.127 ± 0.027	0.278
TGCN	0.351 ± 0.055	0.383 ± 0.081	0.102 ± 0.063	0.207 ± 0.063	0.295 ± 0.044	0.220 ± 0.084	0.260
GPT-4.1-mini	0.237 ± 0.010	0.069 ± 0.052	0.209 ± 0.033	0.269 ± 0.035	0.314 ± 0.028	0.131 ± 0.103	0.205
GPT-4.1	0.070 ± 0.043	0.078 ± 0.029	0.076 ± 0.066	0.301 ± 0.048	0.220 ± 0.033	0.021 ± 0.047	0.128
FAME (Ours)	$\mathbf{0.521 \pm 0.051}$	$\mathbf{0.633 \pm 0.016}$	$\mathbf{0.483 \pm 0.006}$	$\mathbf{0.531 \pm 0.007}$	$\mathbf{0.562 \pm 0.026}$	$\mathbf{0.482 \pm 0.036}$	**0.535**
Model	2024.12	2025.01	2025.02	2025.03	2025.04	2025.05	**Avg**
XGBoost	0.154 ± 0.074	0.183 ± 0.056	0.298 ± 0.046	0.364 ± 0.046	0.219 ± 0.047	0.362 ± 0.014	0.263
SVR	0.370 ± 0.016	0.319 ± 0.024	0.467 ± 0.001	0.447 ± 0.001	0.309 ± 0.003	0.479 ± 0.004	0.399
Transformer	0.257 ± 0.041	0.377 ± 0.068	0.422 ± 0.035	0.062 ± 0.054	0.297 ± 0.027	0.512 ± 0.018	0.321
TGCN	0.277 ± 0.042	0.248 ± 0.029	0.344 ± 0.022	0.151 ± 0.053	0.212 ± 0.085	0.268 ± 0.056	0.250
GPT-4.1-mini	0.260 ± 0.030	0.048 ± 0.027	0.160 ± 0.045	0.141 ± 0.039	0.051 ± 0.049	0.096 ± 0.024	0.126
GPT-4.1	0.106 ± 0.056	-0.004 ± 0.041	0.103 ± 0.062	0.058 ± 0.037	0.116 ± 0.023	0.093 ± 0.035	0.079
Gemini-3-Flash-Preview	-	0.061 ± 0.027	0.149 ± 0.033	0.209 ± 0.025	0.303 ± 0.022	0.259 ± 0.036	0.196
Gemini-3.1-Pro-Preview	-	0.072 ± 0.053	0.146 ± 0.016	0.243 ± 0.035	0.315 ± 0.015	0.298 ± 0.018	0.215
DeepSeek-V3.2	-	0.043 ± 0.063	0.230 ± 0.050	0.255 ± 0.015	0.306 ± 0.000	0.214 ± 0.060	0.210
SciJudge-4B	-	0.122 ± 0.006	0.156 ± 0.028	0.269 ± 0.005	0.254 ± 0.004	0.312 ± 0.074	0.247
FAME (Ours)	$\mathbf{0.524 \pm 0.010}$	$\mathbf{0.433 \pm 0.055}$	$\mathbf{0.563 \pm 0.025}$	$\mathbf{0.492 \pm 0.023}$	$\mathbf{0.450 \pm 0.005}$	$\mathbf{0.617 \pm 0.029}$	**0.513**
Model	2025.06	2025.07	2025.08	2025.09	2025.10	2025.11	**Avg**
XGBoost	0.347 ± 0.066	0.037 ± 0.068	0.288 ± 0.038	0.169 ± 0.024	0.061 ± 0.049	0.054 ± 0.047	0.160
SVR	0.459 ± 0.003	0.200 ± 0.000	0.270 ± 0.004	0.320 ± 0.000	0.297 ± 0.001	0.238 ± 0.015	0.297
Transformer	0.507 ± 0.028	0.441 ± 0.044	0.454 ± 0.055	0.350 ± 0.051	0.057 ± 0.061	-0.047 ± 0.015	0.294
TGCN	0.077 ± 0.041	0.409 ± 0.071	0.508 ± 0.033	0.534 ± 0.019	0.355 ± 0.031	-0.155 ± 0.026	0.288
GPT-4.1-mini	0.009 ± 0.016	0.151 ± 0.029	0.174 ± 0.060	0.140 ± 0.047	0.251 ± 0.146	0.143 ± 0.028	0.145
GPT-4.1	-0.131 ± 0.064	-0.057 ± 0.017	-0.184 ± 0.070	0.103 ± 0.034	0.220 ± 0.090	0.070 ± 0.036	0.003
Gemini-3-Flash-Preview	0.142 ± 0.050	0.263 ± 0.081	0.060 ± 0.092	0.197 ± 0.023	0.216 ± 0.041	0.231 ± 0.058	0.185
Gemini-3.1-Pro-Preview	0.005 ± 0.034	0.088 ± 0.072	0.104 ± 0.038	0.280 ± 0.044	0.140 ± 0.024	0.168 ± 0.037	0.131
DeepSeek-V3.2	0.156 ± 0.030	0.345 ± 0.084	0.134 ± 0.037	0.237 ± 0.085	0.129 ± 0.078	0.143 ± 0.112	0.191
GPT-5.2	-	-	0.082 ± 0.031	0.250 ± 0.033	0.200 ± 0.024	0.183 ± 0.048	0.179
SciJudge-4B	0.174 ± 0.112	0.419 ± 0.012	0.226 ± 0.055	0.213 ± 0.025	0.274 ± 0.069	0.178 ± 0.136	0.302
FAME (Ours)	$\mathbf{0.616 \pm 0.026}$	$\mathbf{0.676 \pm 0.027}$	$\mathbf{0.582 \pm 0.013}$	$\mathbf{0.552 \pm 0.032}$	$\mathbf{0.399 \pm 0.023}$	$\mathbf{0.370 \pm 0.052}$	**0.532**

Baselines. We benchmark FAME against three distinct categories of evaluators. First, we evaluate ML models, including XGBoost [9], SVR [11, 27], Transformer [31] and TGCN [39], trained directly on raw semantic text embeddings to predict continuous impact weights (we use the inspiration graph as the graph for TGCN). Second, we test general-purpose frontier LLMs (including DeepSeek-V3.2 [17], GPT-4.1-mini [1], GPT-4.1 [1], GPT-5.2 [26], Gemini-3-Flash-Preview, and Gemini-3.1-Pro-Preview), which are evaluated using few-shot calibrated prompting. The prediction of the LLM is an average of predicted novelty, interestingness, and potential impact score, ranging from 1 to 10. Finally, we benchmark against SciJudge-4B [30], a model fine-tuned on Qwen3-4B [36] for academic comparisons. We adapted its pairwise preference outputs into continuous scores using a tournament-style ranking approach (detailed in Appendix A.2).

Evaluation Protocol. To emulate prospective forecasting, we evaluate our framework under a temporal out-of-distribution (OOD) setting following SciJudge [30]. We employ an 18-month sliding-window evaluation. For each temporal cutoff T, models predict the impact of unseen papers published within a 2-month future testing window. While FAME is trained exclusively on historical data up to T, general LLMs suffer from earlier knowledge cutoffs ($T_{LLM} < T$). To ensure strict parity under our calibrated framework, all evaluators (including FAME, ML baselines, SciJudge, and few-shot LLMs) are anchored to the exact same realized impact labels. For LLMs, we construct a 30-paper few-shot hint window that strictly pairs historical abstracts ($T_{LLM} < t_{hint} \leq T$) with their finalized impact weights collected at T_{end}. Performance is measured via the Spearman rank correlation between predicted and ground-truth impact weights.

4.1 Main Results

As detailed in our comparative evaluations across the 18-month horizon shown in Table 1, FAME significantly and consistently outperforms all baseline models. By mapping papers to continuous

Table 2: Pearson and Spearman correlation evaluating the alignment between the proposed impact metrics and human ratings of the sampled 300 papers. The results demonstrate that the combined Impact Score aligns more strongly with human consensus than any individual normalized component.

Normalized Score Component	**Pearson Correlation ρ**	**Spearman Correlation ρ_s**
Github Stars	0.4032	0.3020
Citation Count	0.5277	0.4289
Influential Citation Count	0.5673	0.4884
Altmetric Score	0.4677	0.4169
Impact score (Sum four scores)	**0.6549**	**0.5283**

spatiotemporal trajectories, FAME achieves average Spearman correlations exceeding 0.50. In contrast, machine learning baselines (XGBoost, SVR, Transformer, and TGCN) show moderate but limited performance. Because they rely entirely on static text embeddings, they fail to capture the temporal momentum of the field. Standard LLMs exhibit poorer and more volatile performance ($\rho_s < 0.25$), frequently fluctuating near random correlation. This highlights the severe limitations of using static, text-based LLM evaluators for prospective impact prediction. This performance gap persists when LLM evaluators are provided with the same historical papers and impact labels as FAME through many-shot prompting (Appendix C). While the domain-specific SciJudge-4B demonstrates improved stability over general LLMs, its reliance on pairwise text comparisons lacks macro-level trajectory awareness, leaving it substantially short of FAME's performance. We also examine the top-5 accuracy of the predicted papers; the results are shown in Appendix B. Furthermore, we demonstrate that integrating our continuous-time manifold scores as explicit context into standard LLM evaluators significantly enhances their predictive performance, as detailed in Appendix D. To assess applicability beyond computer science, we further evaluate FAME on 600 Lattice QCD papers under the same prospective sliding-window protocol, as shown in Appendix H, FAME achieves the highest average Spearman correlation in each of the three six-month periods, providing additional evidence of its effectiveness in a distinct scientific domain.

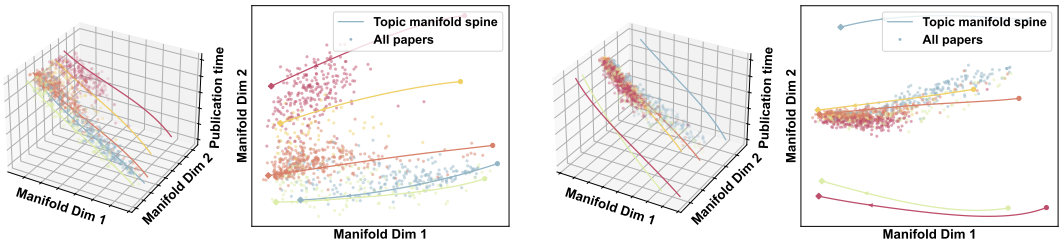
4.2 Ablation Study

To rigorously validate the contribution of each component within our model architecture, we conduct quantitative, qualitative ablation studies and sensitivity analysis of loss weights.

Quantitative Analysis. We compare the full FAME against three variants: removing the inspiration loss ($\mathcal{L}_{inspire}$), the vanguard loss ($\mathcal{L}_{vanguard}$), and the spine loss ($\mathcal{L}_{spine}$). We also test a naive method which trains the scoring model directly on the original text embeddings of the paper produced by a text embedding model. To validate our retrieve-and-verify pipeline, we replace the verified inspiration graph with two baselines: a raw, unfiltered citation graph and a random inspiration graph using eight prior publications as semantic inspirations. The removal of any structural objective leads to a pronounced degradation in predictive performance. Most notably, dropping the vanguard loss yields the most severe performance drop. This confirms that localized impact normalization is critical. It indicates that, without explicitly encouraging high-impact papers to align with the forward momentum of the field, the model cannot effectively rank future contributions. The detailed results are shown in Appendix E and Table 7.

Qualitative Latent Space Dynamics. We used KernelPCA [24] to reduce the dimension of the embeddings. As shown in Figure 5, the fully optimized manifold produces continuous, well-aligned topic trajectories. Removing the spine loss destroys the global temporal anchor. Removing the inspiration loss collapses papers into static clusters devoid of directional knowledge flow. Finally, removing the vanguard loss results in a uniform spatial distribution that utterly fails to differentiate breakthrough research from marginal publications.

Sensitivity Analysis. To evaluate the interplay between FAME's geometric constraints, we conducted a sensitivity analysis on the loss formulation weights (detailed in Appendix F). By systematically varying the spine binding (α), inspiration (β), and vanguard (γ) loss weights, we found that the framework achieves strong, stable predictive performance when β is relatively small, provided α and γ are located within a reasonable range. This confirms that while all three structural objectives are necessary, balancing macroscopic temporal anchoring and localized impact normalization is the



(a) Manifold trained with all losses.

(b) Manifold trained without spine loss.

(c) Manifold trained without inspiration loss.

(d) Manifold trained without vanguard loss.

Figure 5: Qualitative ablation of the spatiotemporal latent manifold. The panels demonstrate how removing individual geometric constraints disrupts the structural integrity of the field's evolutionary trajectory. Within each panel, the left plot illustrates the 3D evolution of the manifold along the publication time axis, while the right plot displays a 2D projection of the embeddings with the temporal dimension collapsed. Different clusters of papers are shown in different colors.

primary driver of the model's forecasting efficacy. As shown in Appendix G, we further examine sensitivity to the number of topic clusters and find that overly coarse clustering reduces forecasting accuracy, whereas average performance largely plateaus at $K = 8$ and remains stable for larger tested values of $K \in 16, 24$, suggesting that FAME does not require a narrowly tuned cluster count.

5 Discussion

Rationality of proposed impact score. Our ground-truth impact weight aggregates GitHub stars, citations, influential citations, and Altmetric scores to capture multi-dimensional scientific influence. To validate this metric, we randomly sampled 300 papers and obtained ground-truth ratings from three independent human evaluators based on novelty, soundness, and potential impact. Comparing our automated metrics to the average human rating, Table 2 shows that the aggregated Impact Score correlates more strongly with human consensus than any individual metric alone. Furthermore, Appendix J illustrates that this strong alignment holds consistently across subfields. This confirms our composite score serves as a robust, human-aligned proxy for genuine scientific impact.

Comparison between raw text embeddings and FAME. Our prospective forecasting results expose a critical flaw in static text evaluations: they ignore a field's macroscopic evolutionary trajectory. To visualize this limitation, we use KernelPCA to reduce the dimensionality of the embeddings. As illustrated in Figure 6 (left), standard text embedding models suffer from a geometric bottleneck, collapsing new paper evaluations directly into the dense semantic centroids of historical data. In this compacted space, novel, high-impact research is severely entangled with ordinary work, causing the baseline model's top predictions (Top-5 predicted, denoted by stars) to land seemingly at random and fail to align with actual breakthrough research (true Top-5 papers, denoted by circles). However, our continuous-time spatiotemporal manifold resolves this centroid collapse, as shown in Figure 6 (right). By projecting textual features into a dynamic latent space structured by continuous topic spines, papers are distributed structurally along their field's evolutionary trajectory. This shifts the evaluation paradigm from static semantic similarity to dynamic geometric alignment, preventing breakthrough ideas from being buried in past literature. As a result, our manifold embeddings allow the predicted vanguard papers (stars) to align closer to the true high-impact ground-truth papers (circles). We provide quantitative analysis in Appendix B and additional visual examples in Appendix I.

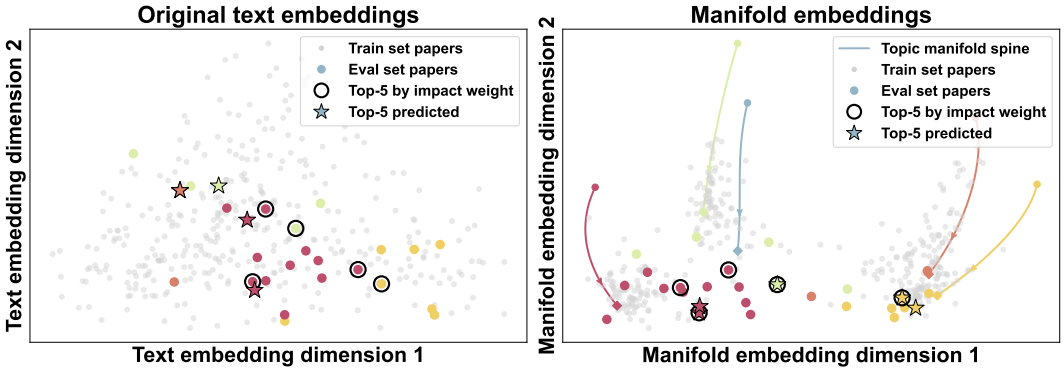


Figure 6: 2D visualizations comparing the latent spaces of standard text embeddings versus the FAME manifold. The left panel illustrates a geometric bottleneck in raw text embeddings, where novel breakthrough ideas are located alongside ordinary publications. The right panel demonstrates how FAME resolves this problem, yielding top-5 impactful predictions that align significantly better with the true ground-truth impact weights.

6 Limitations

While FAME provides a robust framework for automated scientific evaluation, several limitations remain. First, evaluating formalized arXiv manuscripts as proxies leaves a gap in assessing the nascent, early-stage hypotheses required for end-to-end autonomous AI pipelines. Second, while effective in specific computer science subfields, the framework requires adaptation for disciplines with vastly different publication dynamics and time-to-impact horizons. Methodologically, using finalized impact metrics for historical training data deviates from strict real-time forecasting; future work should explore dynamically evolving, time-truncated target labels at the cutoff T. Finally, because citation signals are imperfect proxies for immediate quality, future iterations must model long-term dynamics to accurately capture the delayed impact trajectories of highly innovative research, often termed sleeping beauties.

7 Related Works

LLM-Driven Scientific Discovery and the Evaluation Bottleneck. Automated scientific discovery is shifting from domain-specific models [5, 6] to generalist AI Scientist agents driven by iterative generation-evaluation loops [19, 28, 35, 4, 12]. While LLMs effectively brainstorm [25], navigate literature [14, 29], and write code [34, 37], their evaluation modules—reliant on the LLM-as-a-judge paradigm [40, 7]—remain a critical bottleneck. Despite their fluency, LLMs act as myopic evaluators [18, 33]; even models fine-tuned for academic comparisons, like SciJudge [30], fail to reliably forecast impact because they ignore the macroscopic evolutionary trajectories of research fields.

Spatiotemporal Manifold Learning. Recent advances in representation learning emphasize Continuous-Time Dynamic Graphs [23, 38] and spatiotemporal manifolds [8, 15], but these architectures are primarily optimized for microscopic node interactions [21]. When applied to scientific impact forecasting—which traditionally relies on static bibliometrics [32] or retrospective Graph Neural Networks [13]—they treat network evolution as a passive aggregation of local events, missing the broader macroscopic trajectory of the domain. By explicitly aligning high-impact papers with their field's momentum, our approach streamlines idea evaluation into a direct geometric similarity calculation between a paper's positional residual and its topic's instantaneous trajectory.

8 Conclusion

LLMs are increasingly used to evaluate novel research ideas, but our study reveals they fail to reliably distinguish truly impactful research from ordinary publications. This vulnerability stems from their tendency to evaluate ideas in a static semantic vacuum, overlooking the macroscopic evolutionary trajectory of a scientific field. To address this, we introduce FAME, a continuous-time

spatiotemporal manifold learning framework. By projecting papers into a dynamic latent space structured by continuous topic spines, FAME evaluates ideas based on their geometric alignment with a field's evolutionary momentum. Extensive sliding-window evaluations on 3,200 arXiv papers demonstrate that FAME significantly outperforms state-of-the-art LLM evaluators. Furthermore, integrating our manifold scores directly into standard LLMs as explicit context substantially boosts their predictive reasoning. Ultimately, our study establishes manuscript impact forecasting as a rigorous proxy benchmark, positioning FAME as a robust foundation for scientific evaluation.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Charu C Aggarwal, Alexander Hinneburg, and Daniel A Keim. On the surprising behavior of distance metrics in high dimensional space. In *International conference on database theory*, pages 420–434. Springer, 2001.

[3] Pedro Albarrán, Juan A Crespo, Ignacio Ortuño, and Javier Ruiz-Castillo. The skewness of science in 219 sub-fields and a number of aggregates. *Scientometrics*, 88(2):385–397, 2011.

[4] Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. Researchagent: Iterative research idea generation over scientific literature with large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6709–6738, 2025.

[5] Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023.

[6] Andres M Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. Chemcrow: Augmenting large-language models with chemistry tools. *arXiv preprint arXiv:2304.05376*, 2023.

[7] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

[8] Benjamin Chamberlain, James Rowbottom, Davide Eynard, Francesco Di Giovanni, Xiaowen Dong, and Michael Bronstein. Beltrami flow and neural diffusion on graphs. *Advances in Neural Information Processing Systems*, 34:1594–1609, 2021.

[9] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[10] Arman Cohan, Waleed Ammar, Madeleine Van Zuylen, and Field Cady. Structural scaffolds for citation intent classification in scientific publications. In *Proceedings of the 2019 conference of the North American chapter of the Association for Computational Linguistics: human language technologies, volume 1 (long and short papers)*, pages 3586–3596, 2019.

[11] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[12] Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, et al. Towards an ai co-scientist. *arXiv preprint arXiv:2502.18864*, 2025.

[13] Zhen Han, Zifeng Ding, Yunpu Ma, Yujia Gu, and Volker Tresp. Temporal knowledge graph forecasting with neural ode. *arXiv preprint arXiv:2101.05151*, 2021.

[14] Jakub Lála, Odhran O'Donoghue, Aleksandar Shtedritski, Sam Cox, Samuel G Rodriques, and Andrew D White. Paperqa: Retrieval-augmented generative agent for scientific research. *arXiv preprint arXiv:2312.07559*, 2023.

[15] Xiaona Li, Zhu Wang, Xindong Chen, Bin Guo, and Zhiwen Yu. A hybrid continuous-time dynamic graph representation learning model by exploring both temporal and repetitive information. *ACM Transactions on Knowledge Discovery from Data*, 17(9):1–22, 2023.

[16] Zhentao Liang, Nees Jan van Eck, Xuehua Wu, Jin Mao, and Gang Li. Citation importance-aware document representation learning for large-scale science mapping. *Information Processing & Management*, 63(3):104557, 2026.

[17] Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, et al. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*, 2025.

[18] Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics*, 12:157–173, 2024.

[19] Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

[20] Shambhavi Mishra, Gaurav Sahu, Marco Pedersoli, Laurent Charlin, Jose Dolz, and Christopher Pal. Ainstein: Assessing the feasibility of ai-generated approaches to research problems. *arXiv preprint arXiv:2510.05432*, 2025.

[21] Liang Qu, Huaisheng Zhu, Qiqi Duan, and Yuhui Shi. Continuous-time link prediction via temporal dependent graph neural network. In *Proceedings of the web conference 2020*, pages 3026–3032, 2020.

[22] Filippo Radicchi, Santo Fortunato, and Claudio Castellano. Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences*, 105(45):17268–17272, 2008.

[23] Emanuele Rossi, Ben Chamberlain, Fabrizio Frasca, Davide Eynard, Federico Monti, and Michael Bronstein. Temporal graph networks for deep learning on dynamic graphs. *arXiv preprint arXiv:2006.10637*, 2020.

[24] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Kernel principal component analysis. In *International conference on artificial neural networks*, pages 583–588. Springer, 1997.

[25] Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. *arXiv preprint arXiv:2409.04109*, 2024.

[26] Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

[27] Alex J Smola and Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and computing*, 14(3):199–222, 2004.

[28] Jiabin Tang, Lianghao Xia, Zhonghang Li, and Chao Huang. Ai-researcher: Autonomous scientific innovation. *arXiv preprint arXiv:2505.18705*, 2025.

[29] Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*, 2022.

[30] Jingqi Tong, Mingzhe Li, Hangcheng Li, Yongzhuo Yang, Yurong Mou, Weijie Ma, Zhiheng Xi, Hongji Chen, Xiaoran Liu, Qinyuan Cheng, et al. Ai can learn scientific taste. *arXiv preprint arXiv:2603.14473*, 2026.

[31] 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.

[32] Dashun Wang, Chaoming Song, and Albert-László Barabási. Quantifying long-term scientific impact. *Science*, 342(6154):127–132, 2013.

[33] Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, et al. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, 2024.

[34] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. Autogen: Enabling next-gen llm applications via multi-agent conversations. In *First conference on language modeling*, 2024.

[35] Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist-v2: Workshop-level automated scientific discovery via agentic tree search. *arXiv preprint arXiv:2504.08066*, 2025.

[36] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[37] Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. Large language models as optimizers. In *The Twelfth International Conference on Learning Representations*, 2023.

[38] Le Yu, Leilei Sun, Bowen Du, and Weifeng Lv. Towards better dynamic graph learning: New architecture and unified library. *Advances in Neural Information Processing Systems*, 36:67686–67700, 2023.

[39] Ling Zhao, Yujiao Song, Chao Zhang, Yu Liu, Pu Wang, Tao Lin, Min Deng, and Haifeng Li. T-gcn: A temporal graph convolutional network for traffic prediction. *IEEE transactions on intelligent transportation systems*, 21(9):3848–3858, 2019.

[40] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

A Model Details

A.1 Hyperparameters and Computation Resources

The hyperparameters for constructing the inspiration graph, training the manifold architecture, and constructing the ground truth impact metrics are detailed in Table 3. These parameters govern both the strict retrieval of candidate papers and the subsequent temporal geometric embeddings.

Table 3: Hyperparameter settings for the FAME manifold architecture and inspiration graph.

Inspiration Graph (Retrieve-and-Verify)	
Top-K Candidates per Paper	12
Minimum Time Delta (Δ_{days})	60 days
Semantic Similarity Threshold (τ_{sim})	0.6
LLM Confidence Threshold	0.6
Text Embedding Model	OpenAI text-embedding-3-large
Inspiration Graph Judge LLM	gpt-4.1-mini
Manifold Training Defaults	
Learning Rate	0.0001
Finite Difference Step (Δt)	0.01
Inspiration Alignment Similarity (Δ_{align})	0.5
Base Vanguard Similarity (Δ_{base})	0
Temporal Encoding Dimension (d_{time})	64
Topic Spine Dimension (d_{topic})	128
Spine Loss Weight (α)	1.0
Inspiration Loss Weight (β)	0.1
Vanguard Loss Weight (γ)	1.0
Metrics Construction parameters	
Github Stars (a_1, b_1)	(1, e)
Citation Counts (a_2, b_2)	(1, 8)
Influential Citation Counts (a_3, b_3)	(10, 8)
Altmetric Score (a_4, b_4)	(1, 4)

All the experiments are conducted on an Ubuntu 22.04 server equipped with an AMD EPYC 7K62 CPU and an NVIDIA A100-64G GPU. The running time for each manifold training process is around 2 minutes.

A.2 SciJudge Evaluation Method

Because SciJudge-4B [30] outputs pairwise preferences rather than continuous scores, we adapt it using a tournament-style ranking approach: for a testing window of n papers, we query the model with all $n(n-1)/2$ possible pairs to assess their relative contributions. We maintain a global winning table, and the aggregate win count for each paper serves as its final predicted impact score.

A.3 LLM Prompt Templates

Prompt Template 1: Inspiration Graph Verification. This prompt is used by the judge model gpt-4.1-mini to establish genuine knowledge-flow edges between chronologically ordered paper pairs based on the candidates filtered.

```
You judge whether a later paper is inspired by an earlier paper.
Return strict JSON:
{
  "is_inspired": true,
  "confidence": 0.0,
```

```
  "rationale": "1-2 sentence explanation"
}

Later paper A:
- arXiv ID: {later.paper.arxiv_id}
- Title: {later.paper.title}
- Abstract: {later.paper.abstract}

Earlier paper B:
- arXiv ID: {earlier.paper.arxiv_id}
- Title: {earlier.paper.title}
- Abstract: {earlier.paper.abstract}

Additional signal:
- cosine_similarity: {similarity}
- time_delta_days(A_minus_B): {time_delta_days}
```

Prompt Template 2: Baseline Impact Forecasting. This prompt queries the baseline LLM evaluators to score future papers across metrics.

```
You are a strict research idea evaluator.
Topic: {topic}

Reference: Use the following papers published after your knowledge cutoff
as reference to calibrate your scoring of novelty and potential impact:
{hint_paper.title}{hint_paper.abstract}, impact weight:
{hint_paper.impact_weight}
...

For each idea, provide:

- novelty_score (1-10): Originality relative to existing literature; is
it non-obvious and new?
- interestingness_score (1-10): Intellectual appeal and whether it raises
compelling questions
  or insights
- potential_impact_score (1-10): Expected influence on the field if
successful (e.g., advancing
  theory, shifting paradigms, enabling new research directions, or widely
  adopted applications)

Scoring Guidelines:

- Use only integers from 1 (very low) to 10 (exceptional)
- Be conservative; avoid inflated scores
- Compare against typical research in the field, not absolute ideals

Return concise evaluation_notes (1-3 sentences) per idea.
Output strict JSON only:
{
  "evaluations": [
    {
      "index": 0,
      "novelty_score": 7,
      "interestingness_score": 8,
      "potential_impact_score": 7,
      "evaluation_notes": "..."
    }
  ]
```

```
}

Ideas to evaluate:
{papers}
```

B Top-5 Accuracy Evaluation

We provide a detailed evaluation of the models based on their top-5 prediction accuracy. Specifically, Table 4 illustrates the prospective forecasting performance over an 18-month sliding window from June 2024 to November 2025. This metric measures how accurately a model's top-5 predicted high-impact papers align with the actual ground-truth top-5 papers based on their composite impact weights. The results clearly demonstrate that the proposed FAME framework consistently outperforms all baseline evaluators, including traditional machine learning models and frontier LLMs across every temporal window.

Table 4: Prospective forecasting performance across an 18-month sliding window evaluation from June 2024 to November 2025. Performance is measured by the top-5 accuracy between predicted and ground-truth composite impact weights. The experiments are carried out three times, and the best results are shown in bold.

Model	2024.06	2024.07	2024.08	2024.09	2024.10	2024.11	**Avg**
XGBoost	0.067 ± 0.094	0.200 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.067 ± 0.094	0.056
SVR	0.133 ± 0.094	0.400 ± 0.000	0.400 ± 0.000	0.400 ± 0.000	0.067 ± 0.094	0.133 ± 0.094	0.256
Transformer	0.200 ± 0.000	0.400 ± 0.000	0.133 ± 0.094	0.200 ± 0.000	0.200 ± 0.000	0.200 ± 0.000	0.222
TGCN	0.200 ± 0.000	0.067 ± 0.094	0.133 ± 0.094	0.133 ± 0.094	0.200 ± 0.000	0.133 ± 0.094	0.144
GPT-4.1-mini	0.267 ± 0.094	0.200 ± 0.000	0.133 ± 0.094	0.200 ± 0.000	0.200 ± 0.000	0.267 ± 0.094	0.211
GPT-4.1	0.067 ± 0.094	0.133 ± 0.094	0.133 ± 0.094	0.133 ± 0.094	0.200 ± 0.000	0.200 ± 0.000	0.144
FAME (Ours)	$\mathbf{0.600 \pm 0.000}$	$\mathbf{0.800 \pm 0.000}$	$\mathbf{0.733 \pm 0.094}$	$\mathbf{0.600 \pm 0.000}$	$\mathbf{0.600 \pm 0.000}$	$\mathbf{0.533 \pm 0.094}$	**0.644**

Model	2024.12	2025.01	2025.02	2025.03	2025.04	2025.05	**Avg**
XGBoost	0.133 ± 0.094	0.000 ± 0.000	0.200 ± 0.000	0.000 ± 0.000	0.067 ± 0.094	0.000 ± 0.000	0.067
SVR	0.200 ± 0.000	0.133 ± 0.094	0.400 ± 0.000	0.400 ± 0.000	0.200 ± 0.000	0.200 ± 0.000	0.256
Transformer	0.333 ± 0.094	0.133 ± 0.094	0.333 ± 0.094	0.400 ± 0.000	0.400 ± 0.000	0.200 ± 0.000	0.300
TGCN	0.000 ± 0.000	0.200 ± 0.000	0.267 ± 0.094	0.200 ± 0.000	0.200 ± 0.000	0.133 ± 0.094	0.167
GPT-4.1-mini	0.200 ± 0.000	0.000 ± 0.000	0.067 ± 0.094	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.044
GPT-4.1	0.267 ± 0.189	0.200 ± 0.000	0.200 ± 0.000	0.200 ± 0.000	0.133 ± 0.094	0.000 ± 0.000	0.167
Gemini-3-Flash-Preview	-	0.200 ± 0.000	0.267 ± 0.094	0.333 ± 0.094	0.333 ± 0.094	0.200 ± 0.000	0.267
Gemini-3.1-Pro-Preview	-	0.200 ± 0.000	0.400 ± 0.000	0.267 ± 0.094	0.200 ± 0.000	0.267 ± 0.094	0.267
DeepSeek-V3.2	-	0.133 ± 0.094	0.267 ± 0.094	0.200 ± 0.000	0.200 ± 0.000	0.067 ± 0.094	0.173
SciJudge-4B	-	0.400 ± 0.000	0.200 ± 0.000	0.000 ± 0.000	0.067 ± 0.094	0.000 ± 0.000	0.136
FAME (Ours)	$\mathbf{0.600 \pm 0.000}$	$\mathbf{0.667 \pm 0.094}$	$\mathbf{0.800 \pm 0.000}$	$\mathbf{0.600 \pm 0.000}$	$\mathbf{0.667 \pm 0.094}$	$\mathbf{0.533 \pm 0.094}$	**0.644**

Model	2025.06	2025.07	2025.08	2025.09	2025.10	2025.11	**Avg**
XGBoost	0.000 ± 0.000	0.067 ± 0.094	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.133 ± 0.094	0.033
SVR	0.200 ± 0.000	0.267 ± 0.094	0.200 ± 0.000	0.200 ± 0.000	0.200 ± 0.000	0.200 ± 0.000	0.211
Transformer	0.200 ± 0.000	0.267 ± 0.094	0.200 ± 0.163	0.400 ± 0.000	0.200 ± 0.000	0.200 ± 0.000	0.244
TGCN	0.067 ± 0.094	0.200 ± 0.000	0.200 ± 0.000	0.200 ± 0.000	0.200 ± 0.000	0.133 ± 0.094	0.167
GPT-4.1-mini	0.400 ± 0.000	0.000 ± 0.000	0.133 ± 0.094	0.333 ± 0.094	0.400 ± 0.000	0.200 ± 0.000	0.244
GPT-4.1	0.200 ± 0.000	0.067 ± 0.094	0.067 ± 0.094	0.000 ± 0.000	0.333 ± 0.094	0.133 ± 0.094	0.133
Gemini-3-Flash-Preview	0.200 ± 0.000	0.400 ± 0.000	0.400 ± 0.000	0.133 ± 0.094	0.200 ± 0.000	0.200 ± 0.000	0.256
Gemini-3.1-Pro-Preview	0.333 ± 0.249	0.467 ± 0.094	0.200 ± 0.000	0.133 ± 0.094	0.200 ± 0.163	0.133 ± 0.094	0.244
DeepSeek-V3.2	0.267 ± 0.094	0.200 ± 0.000	0.200 ± 0.000	0.400 ± 0.000	0.333 ± 0.094	0.200 ± 0.000	0.267
GPT-5.2	-	-	0.267 ± 0.094	0.067 ± 0.094	0.200 ± 0.163	0.267 ± 0.094	0.200
SciJudge-4B	0.200 ± 0.000	0.333 ± 0.094	0.133 ± 0.094	0.000 ± 0.000	0.400 ± 0.000	0.267 ± 0.094	0.222
FAME (Ours)	$\mathbf{0.800 \pm 0.000}$	$\mathbf{0.733 \pm 0.094}$	$\mathbf{0.600 \pm 0.000}$	$\mathbf{0.600 \pm 0.000}$	$\mathbf{0.467 \pm 0.094}$	$\mathbf{0.333 \pm 0.094}$	**0.589**

C Many-Shot LLM Comparison

To control for the amount of historical supervision available to the LLM evaluators, we conduct a many-shot comparison in which each LLM receives the same historical paper pool, and thus the same number of papers, as FAME at each temporal cutoff. Each historical example is formatted using its title, abstract, and realized impact label. The resulting prompts contain at most 80k input tokens, which is within the context window of every evaluated model; therefore, no historical examples are omitted or truncated.

As shown in Table 5, the many-shot LLM evaluators remain substantially below FAME despite access to the same historical paper pool. This control is not equivalent to fine-tuning, so our conclusion is limited to the evaluated prompt-based LLM judges.

Table 5: Many-shot LLM comparison under matched historical supervision. Each LLM receives the same historical paper pool as FAME at each cutoff. Performance is measured by the Spearman rank correlation ρ_s between predicted and ground-truth composite impact weights. The experiments are carried out three times, and the standard deviation of the performance is reported. Best results are shown in bold.

Model	2024.06	2024.07	2024.08	2024.09	2024.10	2024.11	**Avg**
GPT-4.1-mini	0.078 ± 0.001	0.155 ± 0.191	0.176 ± 0.124	0.181 ± 0.020	0.167 ± 0.003	0.178 ± 0.123	0.156
GPT-4.1	0.166 ± 0.005	0.110 ± 0.029	0.244 ± 0.124	0.262 ± 0.084	0.268 ± 0.026	0.191 ± 0.040	0.207
FAME (Ours)	$\mathbf{0.521 \pm 0.051}$	$\mathbf{0.633 \pm 0.016}$	$\mathbf{0.483 \pm 0.006}$	$\mathbf{0.531 \pm 0.007}$	$\mathbf{0.562 \pm 0.026}$	$\mathbf{0.482 \pm 0.036}$	**0.535**
Model	2024.12	2025.01	2025.02	2025.03	2025.04	2025.05	**Avg**
GPT-4.1-mini	0.211 ± 0.015	0.128 ± 0.022	0.178 ± 0.007	0.107 ± 0.061	0.133 ± 0.041	0.186 ± 0.019	0.157
GPT-4.1	0.218 ± 0.062	0.061 ± 0.052	0.122 ± 0.065	0.149 ± 0.064	0.127 ± 0.027	0.235 ± 0.007	0.152
Gemini-3-Flash-Preview	-	0.067 ± 0.028	0.148 ± 0.025	0.259 ± 0.020	0.264 ± 0.012	0.277 ± 0.032	0.203
Gemini-3.1-Pro-Preview	-	0.107 ± 0.004	0.128 ± 0.004	0.212 ± 0.058	0.307 ± 0.007	0.305 ± 0.016	0.212
DeepSeek-V3.2	-	0.092 ± 0.029	0.131 ± 0.047	0.151 ± 0.094	0.223 ± 0.079	0.254 ± 0.012	0.170
FAME (Ours)	$\mathbf{0.524 \pm 0.010}$	$\mathbf{0.433 \pm 0.055}$	$\mathbf{0.563 \pm 0.025}$	$\mathbf{0.492 \pm 0.023}$	$\mathbf{0.450 \pm 0.005}$	$\mathbf{0.617 \pm 0.029}$	**0.513**
Model	2025.06	2025.07	2025.08	2025.09	2025.10	2025.11	**Avg**
GPT-4.1-mini	0.087 ± 0.136	0.085 ± 0.006	0.152 ± 0.078	0.117 ± 0.027	0.152 ± 0.019	0.182 ± 0.019	0.129
GPT-4.1	0.072 ± 0.163	0.151 ± 0.035	0.186 ± 0.086	0.164 ± 0.037	0.107 ± 0.038	0.109 ± 0.004	0.132
Gemini-3-Flash-Preview	0.185 ± 0.077	0.226 ± 0.008	0.178 ± 0.082	0.224 ± 0.062	0.188 ± 0.050	0.155 ± 0.038	0.193
Gemini-3.1-Pro-Preview	0.195 ± 0.135	0.209 ± 0.086	0.196 ± 0.127	0.238 ± 0.031	0.213 ± 0.097	0.178 ± 0.038	0.205
DeepSeek-V3.2	0.103 ± 0.098	0.230 ± 0.026	0.043 ± 0.000	0.109 ± 0.043	0.155 ± 0.044	0.038 ± 0.052	0.113
GPT-5.2	-	-	0.227 ± 0.140	0.255 ± 0.032	0.196 ± 0.075	0.082 ± 0.066	0.190
FAME (Ours)	$\mathbf{0.616 \pm 0.026}$	$\mathbf{0.676 \pm 0.027}$	$\mathbf{0.582 \pm 0.013}$	$\mathbf{0.552 \pm 0.032}$	$\mathbf{0.399 \pm 0.023}$	$\mathbf{0.370 \pm 0.052}$	**0.532**

D Boosting LLM Evaluators with Manifold Integration

While standard Large Language Models struggle to independently evaluate scientific impact due to their static semantic context, their reasoning and critique capabilities can be significantly enhanced when grounded with the geometric signals produced by our FAME framework.

D.1 Integration Methodology

To achieve this integration, we modify the standard LLM evaluation prompt to explicitly include the macroscopic evolutionary momentum calculated by our model. Specifically, we inject the fused manifold and naive embedding scores directly into the system prompt structure, allowing the LLM to utilize them as an auxiliary signal.

The baseline evaluation prompt (Prompt Template 2) is augmented by injecting the following `{frontier_section}` into the LLM's context window prior to the instructions:

```
manifold scores for this eval batch (same index order as the ideas list
below):
where score is the trained manifold prediction and z-scores are taken
over this batch. Use as auxiliary signal, judge each idea on its own
merits:
- index 0: {score_0}
- index 1: {score_1}
...
```

D.2 Performance Improvements

By feeding the output of the manifold model back into the LLM as explicit context, the LLM is no longer forced to operate in a temporal vacuum. Instead, it utilizes the continuous-time spatiotemporal trajectory as an authoritative, macro-level reference point. This allows the LLM to effectively combine its inherent, microscopic textual critique capabilities (e.g., evaluating methodological soundness and clarity) with FAME's macroscopic structural awareness. This hybrid evaluation paradigm effectively mitigates the centroid collapse vulnerability. As shown in Table 6, models augmented with FAME exhibit an average Spearman rank correlation increase of over $+0.150$ compared to their unaugmented counterparts. This synergistic integration confirms that while LLMs are currently poor standalone forecasters of scientific evolution, they function exceptionally well as hybrid evaluators when geometrically anchored to the forward momentum of a field.

Table 6: Performance improvements of baseline LLM evaluators when augmented with FAME manifold scores, improvements are shown in green. The experiments are carried out for three times and the standard deviation of the performance is reported.

Model	2024.06 - 2024.12	2025.01 - 2025.07	2025.08 - 2025.11	**Avg**
GPT-4.1-mini	0.213 ± 0.079	0.094 ± 0.055	0.177 ± 0.045	0.158 ± 0.083
w/ FAME	0.391 ± 0.099 (+0.178 ↑)	0.251 ± 0.067 (+0.157 ↑)	0.304 ± 0.155 (+0.127 ↑)	0.317 ± 0.122 (+0.159 ↑)
GPT-4.1	0.125 ± 0.092	0.025 ± 0.086	0.052 ± 0.147	0.070 ± 0.114
w/ FAME	0.394 ± 0.100 (+0.269 ↑)	0.366 ± 0.092 (+0.341 ↑)	0.290 ± 0.137 (+0.238 ↑)	0.360 ± 0.113 (+0.290 ↑)
Gemini-3-Flash-Preview	-	0.198 ± 0.079	0.176 ± 0.068	0.190 ± 0.076
w/ FAME	-	0.399 ± 0.083 (+0.201 ↑)	0.410 ± 0.105 (+0.234 ↑)	0.403 ± 0.092 (+0.213 ↑)
Gemini-3.1-Pro-Preview	-	0.167 ± 0.111	0.173 ± 0.066	0.169 ± 0.097
w/ FAME	-	0.362 ± 0.105 (+0.195 ↑)	0.382 ± 0.104 (+0.209 ↑)	0.369 ± 0.105 (+0.200 ↑)
DeepSeek-V3.2	-	0.221 ± 0.092	0.161 ± 0.044	0.199 ± 0.084
w/ FAME	-	0.367 ± 0.129 (+0.146 ↑)	0.409 ± 0.124 (+0.248 ↑)	0.382 ± 0.128 (+0.183 ↑)
GPT-5.2	-	-	0.179 ± 0.061	0.179 ± 0.061
w/ FAME	-	-	0.394 ± 0.103 (+0.215 ↑)	0.394 ± 0.103 (+0.215 ↑)

E Ablation Study on Manifold Learning

To rigorously validate the contribution of each geometric objective within the FAME architecture, we systematically evaluated the model's predictive performance after independently removing the inspiration loss ($\mathcal{L}_{inspire}$), vanguard loss ($\mathcal{L}_{vanguard}$), and spine loss ($\mathcal{L}_{spine}$). Furthermore, we assessed the importance of our retrieve-and-verify pipeline by substituting the refined inspiration graph with either the full citation graph or a random inspiration graph. For the random inspiration graph variant, we explicitly dismantle true knowledge flow by randomly selecting 8 papers published earlier than the target paper to serve as its inspiration papers. Meanwhile, to separate the effects of label supervision and publication time from the proposed geometric formulation, we train an MLP on the same impact labels after concatenating the raw text embedding with a temporal embedding. The normalized publication time and its learned projection are defined as

$$\tau = \frac{t - t_{\min}^{\text{train}}}{t_{\max}^{\text{train}} - t_{\min}^{\text{train}}}, \quad e_t = W_t \tau + b_t. \tag{9}$$

The 32-dimensional temporal embedding is concatenated with the unchanged text embedding. The resulting representation is passed through an MLP, using LayerNorm and ReLU in each hidden layer. All time-normalization statistics are computed exclusively from training papers.

As shown in Table 7, removing any geometric constraint degrades performance over the 18-month evaluation, while FAME maintains average Spearman correlations above 0.510 in all temporal phases. The vanguard loss ablation causes the largest drop, confirming the importance of localized impact normalization and momentum alignment for impact ranking. Ablating the inspiration or spine loss confirms that directional knowledge flow and temporal anchoring are essential to manifold structure and prediction. Replacing the verified inspiration graph with full citations or random links also reduces performance, underscoring the need to filter citation noise and capture genuine semantic inspiration.The MLP + temporal embedding baseline also underperforms FAME despite using the same impact labels, suggesting that temporal features and label supervision alone are insufficient to explain the gains from our geometric formulation.

Table 7: Quantitative ablation study evaluating the impact of each geometric objective. Removing the inspiration loss, vanguard loss, or spine loss results in pronounced degradation in Spearman rank correlation between predicted score and true impact weight compared to the full FAME model. The experiments are carried out for three times and the standard deviation of the performance is reported.

Model	2024.06	2024.07	2024.08	2024.09	2024.10	2024.11	**Avg**
MLP (raw text embedding)	0.058 ± 0.110	0.204 ± 0.050	0.238 ± 0.018	0.278 ± 0.076	0.255 ± 0.040	0.125 ± 0.019	0.193
MLP + temporal embedding	0.210 ± 0.064	0.301 ± 0.143	0.318 ± 0.106	0.384 ± 0.032	0.281 ± 0.042	0.167 ± 0.104	0.277
w/ full citation graph	0.160 ± 0.040	0.011 ± 0.109	0.193 ± 0.020	0.316 ± 0.058	0.167 ± 0.008	0.345 ± 0.050	0.199
w/ random inspiration graph	0.077 ± 0.091	0.099 ± 0.051	0.057 ± 0.121	0.258 ± 0.055	0.157 ± 0.040	0.128 ± 0.034	0.129
w/o inspiration loss	0.160 ± 0.026	0.177 ± 0.045	0.185 ± 0.045	0.330 ± 0.037	0.228 ± 0.036	0.177 ± 0.039	0.209
w/o vanguard loss	0.061 ± 0.103	0.073 ± 0.061	0.223 ± 0.062	0.235 ± 0.047	0.235 ± 0.054	0.108 ± 0.061	0.156
w/o spine loss	0.186 ± 0.036	0.210 ± 0.055	0.276 ± 0.016	0.396 ± 0.036	0.307 ± 0.027	0.234 ± 0.021	0.268
FAME	**0.521 ± 0.051**	**0.633 ± 0.016**	**0.483 ± 0.006**	**0.531 ± 0.007**	**0.562 ± 0.026**	**0.482 ± 0.036**	**0.535**
Model	2024.12	2025.01	2025.02	2025.03	2025.04	2025.05	**Avg**
MLP (raw text embedding)	0.324 ± 0.035	0.281 ± 0.050	0.457 ± 0.022	0.274 ± 0.083	0.250 ± 0.052	0.437 ± 0.051	0.337
MLP + temporal embedding	0.262 ± 0.083	0.259 ± 0.038	0.386 ± 0.024	0.309 ± 0.070	0.317 ± 0.041	0.459 ± 0.033	0.332
w/ full citation graph	0.340 ± 0.028	0.014 ± 0.027	0.263 ± 0.003	0.360 ± 0.002	0.328 ± 0.026	0.474 ± 0.008	0.296
w/ random inspiration graph	0.167 ± 0.047	0.214 ± 0.020	0.383 ± 0.006	0.062 ± 0.083	0.279 ± 0.027	0.440 ± 0.034	0.257
w/o inspiration loss	0.295 ± 0.080	0.261 ± 0.034	0.421 ± 0.031	0.216 ± 0.121	0.332 ± 0.039	0.485 ± 0.024	0.335
w/o vanguard loss	0.256 ± 0.037	0.115 ± 0.011	0.266 ± 0.036	0.145 ± 0.091	0.179 ± 0.037	0.224 ± 0.086	0.197
w/o spine loss	0.346 ± 0.035	0.253 ± 0.042	0.436 ± 0.014	0.261 ± 0.052	0.372 ± 0.009	0.529 ± 0.024	0.366
FAME	**0.524 ± 0.010**	**0.433 ± 0.055**	**0.563 ± 0.025**	**0.492 ± 0.023**	**0.450 ± 0.005**	**0.617 ± 0.029**	**0.513**
Model	2025.06	2025.07	2025.08	2025.09	2025.10	2025.11	**Avg**
MLP (raw text embedding)	0.361 ± 0.018	0.248 ± 0.056	0.363 ± 0.095	0.445 ± 0.026	0.204 ± 0.064	-0.062 ± 0.082	0.260
MLP + temporal embedding	0.410 ± 0.039	0.451 ± 0.080	0.468 ± 0.044	0.372 ± 0.047	0.253 ± 0.043	0.121 ± 0.104	0.346
w/ full citation graph	0.345 ± 0.042	0.313 ± 0.032	0.356 ± 0.015	0.361 ± 0.027	0.224 ± 0.098	0.235 ± 0.078	0.306
w/ random inspiration graph	0.364 ± 0.009	0.194 ± 0.025	0.275 ± 0.067	0.317 ± 0.012	0.131 ± 0.104	0.107 ± 0.050	0.231
w/o inspiration loss	0.389 ± 0.021	0.292 ± 0.085	0.357 ± 0.050	0.367 ± 0.036	0.295 ± 0.124	0.206 ± 0.053	0.317
w/o vanguard loss	0.260 ± 0.034	0.213 ± 0.034	0.315 ± 0.029	0.414 ± 0.036	0.343 ± 0.123	0.156 ± 0.134	0.283
w/o spine loss	0.463 ± 0.042	0.367 ± 0.026	0.454 ± 0.012	0.436 ± 0.014	0.365 ± 0.047	0.269 ± 0.075	0.392
FAME	**0.616 ± 0.026**	**0.676 ± 0.027**	**0.582 ± 0.013**	**0.552 ± 0.032**	**0.399 ± 0.023**	**0.370 ± 0.052**	**0.532**

F Sensitivity Analysis of Loss Weights

To rigorously evaluate the stability of the FAME framework and the interplay between its constituent geometric constraints, we conduct a comprehensive sensitivity analysis of the loss formulation hyperparameters. Specifically, we systematically vary the relative weights of the spine binding loss (α), inspiration loss (β), and vanguard loss (γ) within the total objective function to explore the combinatorial hyperparameter space. The resultant predictive performance, quantified via the Spearman rank correlation between the predicted and ground-truth impact weights, is mapped across these varying weight distributions. Specifically, the heatmap indicates that the framework consistently achieves strong predictive performance when the inspiration loss weight (β) is kept relatively small, provided that the spine binding loss (α) and vanguard loss (γ) are located within a reasonable range. This analysis elucidates the optimal balance required among the three structural objectives, further corroborating the findings from our ablation studies that the concurrent application of macroscopic temporal anchoring, directional knowledge flow modeling, and localized impact normalization is imperative for maximizing the model's forecasting efficacy.

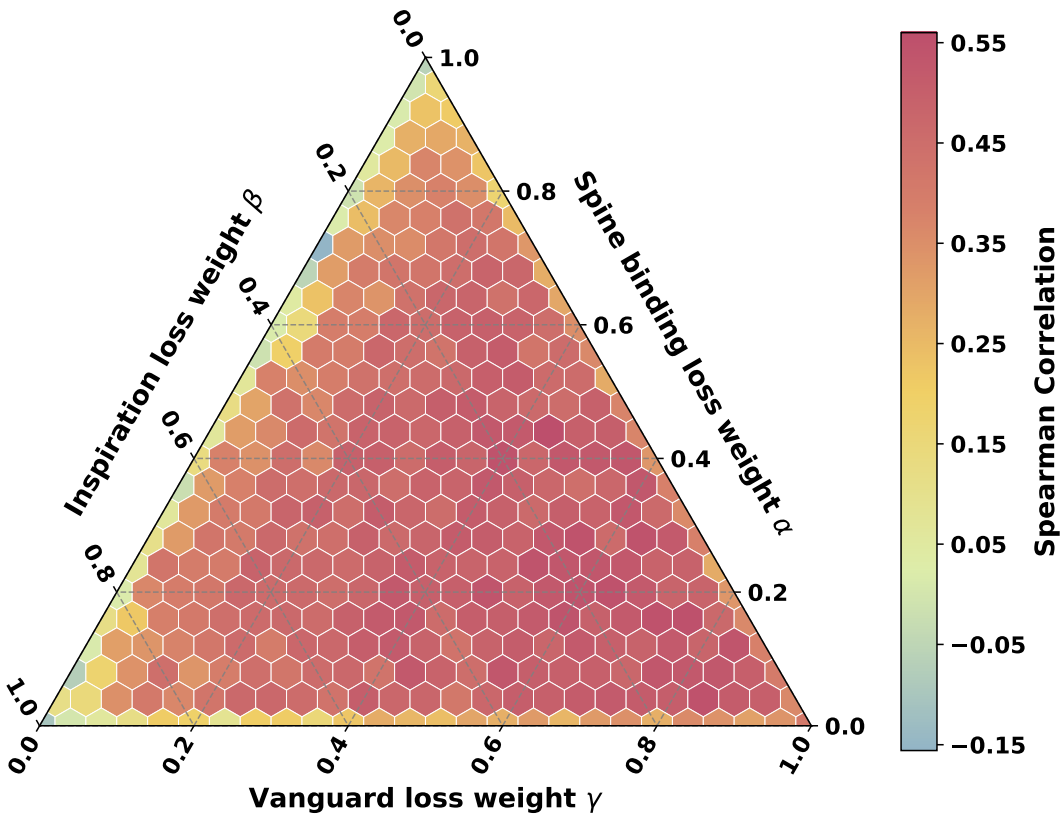


Figure 7: Ternary heatmap illustrating the sensitivity of the FAME framework's predictive performance to variations in the geometric loss weights. The color intensity corresponds to the Spearman rank correlation achieved across different combinatorial assignments of the spine binding loss weight (α), inspiration loss weight (β), and vanguard loss weight (γ). The visualization indicates the optimal operating region where all three structural constraints are synergistically balanced to structure the spatiotemporal manifold.

G Sensitivity Analysis of Topic Cluster Number

We further evaluate the sensitivity of FAME to the number of topic clusters K under the same sliding-window protocol. The main experiments use $K = 8$, and we vary the number of clusters over $K \in \{1, 2, 4, 8, 16, 24\}$. As shown in Table 8, performance mainly depends on avoiding severe under-clustering rather than selecting a narrowly tuned value of K. The setting $K = 4$ is already competitive, and performance effectively plateaus at $K = 8$.

When K is too small, papers from distinct subfields share a topic cluster and spine, pulling their manifold embeddings together and obscuring different evolutionary directions. Larger values have little adverse effect because similar semantic features and cross-cluster inspiration relationships keep related embeddings and topic spines close in manifold space. The mean correlations across the three periods are 0.527, 0.527, and 0.528 for $K \in \{8, 16, 24\}$, respectively, demonstrating that FAME is stable for moderately large values of K.

Table 8: Sensitivity of FAME to the number of topic clusters K. Performance is measured by the Spearman rank correlation ρ_s between predicted and ground-truth composite impact weights. Each entry reports the mean and standard deviation across the corresponding six monthly evaluation windows.

K	2024.06–2024.11	2024.12–2025.05	2025.06–2025.11
1	0.484 ± 0.054	0.460 ± 0.052	0.475 ± 0.113
2	0.469 ± 0.058	0.483 ± 0.060	0.509 ± 0.097
4	0.519 ± 0.070	0.491 ± 0.100	0.493 ± 0.123
8	0.535 ± 0.052	0.513 ± 0.064	0.532 ± 0.112
16	0.544 ± 0.045	0.522 ± 0.078	0.516 ± 0.159
24	0.552 ± 0.064	0.516 ± 0.076	0.516 ± 0.119

H Evaluation on Lattice QCD

To evaluate FAME beyond computer science, we collect 600 Lattice QCD papers from theoretical and computational physics and apply the same prospective sliding-window protocol. Because software adoption is not a representative signal in this field, the impact target is constructed using citation counts and Altmetric scores, excluding GitHub stars and influential citation counts from the original four-signal composite.

As shown in Table 9, FAME achieves the highest average correlation in all three six-month periods: 0.268, 0.237, and 0.467, compared with the strongest competing averages of 0.167, 0.103, and 0.126, respectively. This result extends the evidence beyond the original machine-learning subfields and four-signal impact target. Nevertheless, evaluation on one additional field does not establish universal applicability, and our conclusions remain limited to the evaluated machine-learning subfields and Lattice QCD setting.

Table 9: Prospective forecasting performance on 600 Lattice QCD papers across an 18-month sliding window evaluation from June 2024 to November 2025. Performance is measured by the Spearman rank correlation ρ_s between predicted and ground-truth impact weights constructed from citation counts and Altmetric scores. The experiments are carried out three times, and the standard deviation of the performance is reported. Best results are shown in bold.

Model	2024.06	2024.07	2024.08	2024.09	2024.10	2024.11	**Avg**
MLP (raw text embedding)	0.066 ± 0.158	0.173 ± 0.054	0.211 ± 0.059	0.000 ± 0.126	0.223 ± 0.050	0.105 ± 0.056	0.130
MLP + temporal embedding	0.168 ± 0.054	-0.036 ± 0.063	-0.143 ± 0.066	-0.110 ± 0.014	0.171 ± 0.045	0.107 ± 0.046	0.026
GPT-4.1-mini	0.238 ± 0.000	0.216 ± 0.003	0.168 ± 0.008	**0.099 ± 0.045**	0.186 ± 0.000	0.076 ± 0.018	0.164
GPT-4.1	0.310 ± 0.000	**0.265 ± 0.000**	0.200 ± 0.027	0.081 ± 0.029	0.098 ± 0.031	0.050 ± 0.010	0.167
FAME (Ours)	**0.449 ± 0.090**	0.237 ± 0.017	**0.218 ± 0.049**	0.012 ± 0.012	**0.509 ± 0.025**	**0.182 ± 0.012**	**0.268**

Model	2024.12	2025.01	2025.02	2025.03	2025.04	2025.05	**Avg**
MLP (raw text embedding)	-0.023 ± 0.020	0.017 ± 0.069	-0.062 ± 0.031	-0.086 ± 0.077	0.026 ± 0.065	0.146 ± 0.030	0.003
MLP + temporal embedding	0.044 ± 0.046	0.038 ± 0.032	-0.080 ± 0.053	-0.000 ± 0.022	0.088 ± 0.042	0.098 ± 0.019	0.031
GPT-4.1-mini	0.049 ± 0.011	0.102 ± 0.013	0.084 ± 0.011	0.079 ± 0.034	0.093 ± 0.035	**0.162 ± 0.039**	0.095
GPT-4.1	0.052 ± 0.009	0.097 ± 0.005	0.105 ± 0.007	0.020 ± 0.008	-0.009 ± 0.041	0.031 ± 0.008	0.049
Gemini-3-Flash-Preview	-	0.028 ± 0.005	0.108 ± 0.005	0.064 ± 0.065	0.013 ± 0.002	-0.049 ± 0.015	0.033
Gemini-3.1-Pro-Preview	-	0.051 ± 0.023	0.069 ± 0.024	0.027 ± 0.011	-0.135 ± 0.032	-0.103 ± 0.008	-0.018
DeepSeek-V3.2	-	0.133 ± 0.045	0.173 ± 0.016	0.156 ± 0.032	0.044 ± 0.034	0.010 ± 0.003	0.103
FAME (Ours)	**0.176 ± 0.024**	**0.227 ± 0.035**	**0.243 ± 0.063**	**0.360 ± 0.025**	**0.270 ± 0.006**	0.148 ± 0.009	**0.237**

Model	2025.06	2025.07	2025.08	2025.09	2025.10	2025.11	**Avg**
MLP (raw text embedding)	0.218 ± 0.026	0.106 ± 0.095	-0.121 ± 0.146	0.077 ± 0.028	0.277 ± 0.057	0.197 ± 0.071	0.126
MLP + temporal embedding	0.302 ± 0.063	0.118 ± 0.043	-0.160 ± 0.093	0.020 ± 0.179	0.180 ± 0.080	0.257 ± 0.046	0.119
GPT-4.1-mini	-0.146 ± 0.058	0.038 ± 0.026	0.146 ± 0.053	-0.024 ± 0.027	-0.028 ± 0.000	0.275 ± 0.019	0.043
GPT-4.1	0.056 ± 0.048	-0.036 ± 0.007	0.073 ± 0.038	0.020 ± 0.039	0.097 ± 0.032	0.171 ± 0.040	0.064
Gemini-3-Flash-Preview	-0.086 ± 0.067	-0.038 ± 0.005	0.079 ± 0.029	-0.083 ± 0.032	0.033 ± 0.036	0.055 ± 0.072	-0.007
Gemini-3.1-Pro-Preview	0.003 ± 0.016	-0.035 ± 0.020	0.133 ± 0.019	-0.117 ± 0.018	0.127 ± 0.027	0.107 ± 0.029	0.036
DeepSeek-V3.2	-0.046 ± 0.050	-0.006 ± 0.012	0.169 ± 0.013	0.000 ± 0.027	0.184 ± 0.034	0.125 ± 0.006	0.071
GPT-5.2	-	-	0.162 ± 0.004	-0.052 ± 0.024	0.135 ± 0.000	0.092 ± 0.018	0.084
FAME (Ours)	**0.584 ± 0.000**	**0.551 ± 0.036**	**0.263 ± 0.027**	**0.411 ± 0.077**	**0.397 ± 0.041**	**0.595 ± 0.044**	**0.467**

I More Case On Text Embeddings vs. Manifold Embeddings

We provide more cases on comparing using the original text embeddings of the papers produced by the text embedding model and the manifold embeddings produced by FAME.

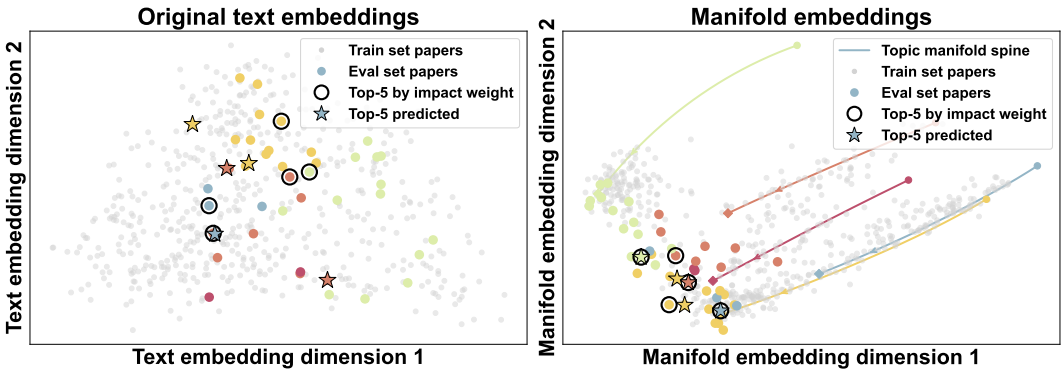


Figure 8: Additional examples comparing original text embedding distributions against FAME manifold embeddings.

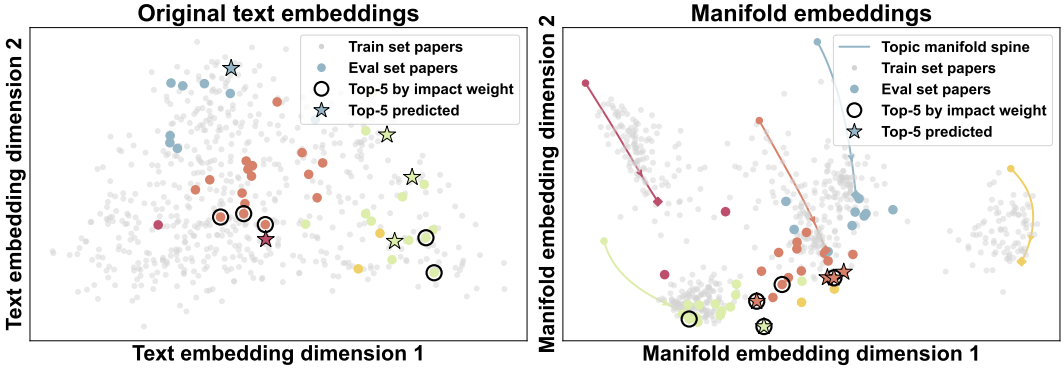


Figure 9: Additional examples comparing original text embedding distributions against FAME manifold embeddings.

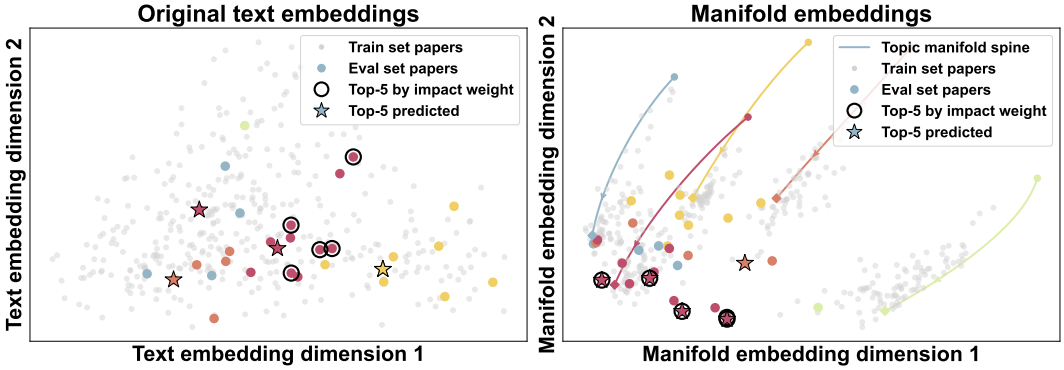


Figure 10: Additional examples comparing original text embedding distributions against FAME manifold embeddings.

J Correlation of Normalized Impact Score Components with Human Ratings

Detailed scatter plots demonstrate a strong alignment between our impact score and human ratings across three research domains of 300 papers. Plotting human ratings (x-axis) against metric scores (y-axis) alongside dashed linear regression lines and Pearson correlations (ρ), reveals that the composite impact score captures human consensus significantly better than any individual normalized component.

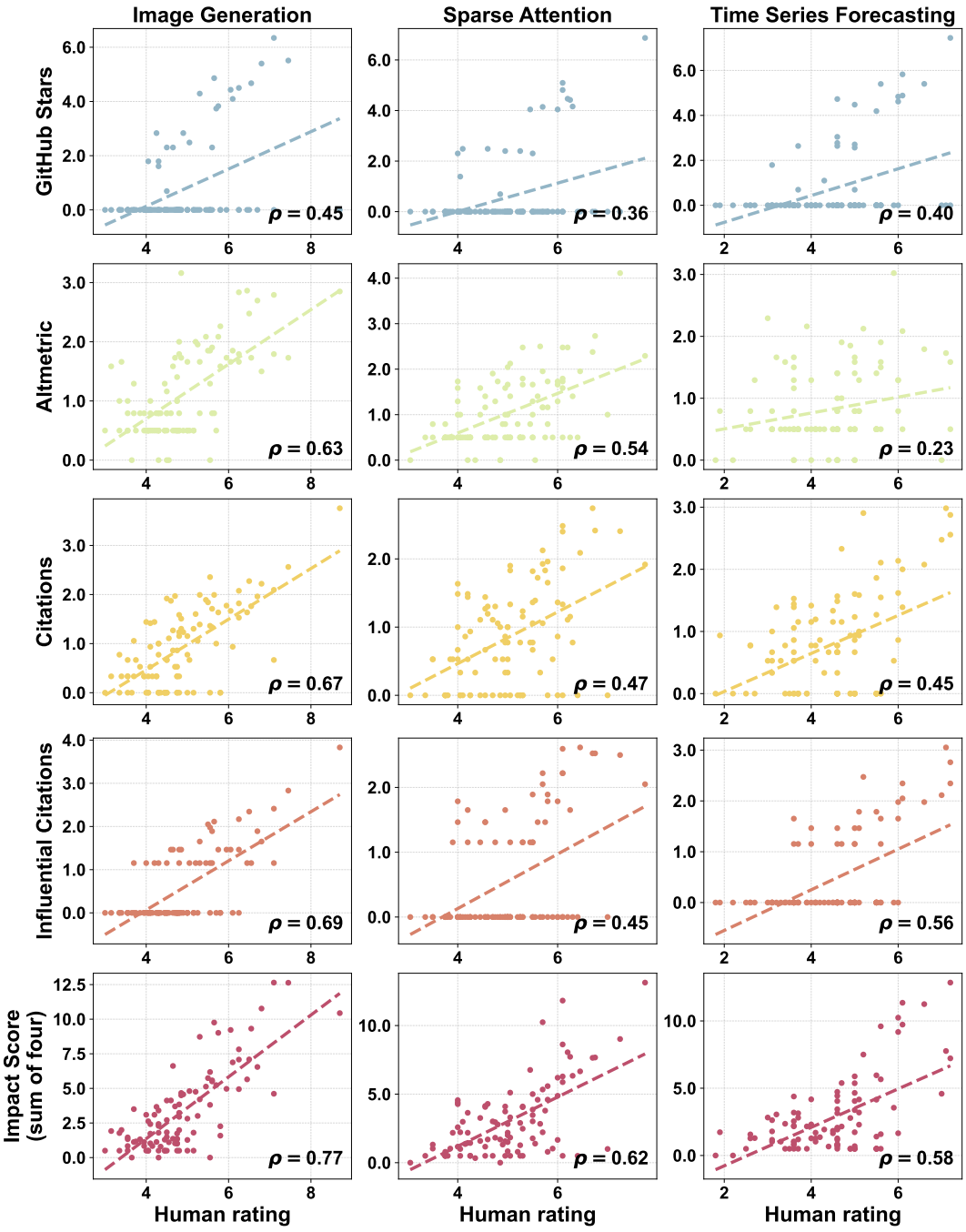


Figure 11: Correlation of normalized impact score components with human ratings. The composite impact score correlates more strongly with human consensus than any individual component.

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: The main claims made in the abstract and introduction reflect the paper's contributions.

   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 include a limitations section at Section 6 in the paper.

   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: [N/A]

Justification: The paper does not include theoretical results

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 detailed parameter setting in Appendix A and experiment results in Section 4.

   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 submit the code in the supplementary materials.

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: We provide detailed parameter setting of the experiments in Appendix A.

   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: We report standard deviation of the performance in the paper.

   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 provide CPU and GPU types used in the experiment and running time in Appendix A.

   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: The research conform with the NeurIPS Code of Ethics.

   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: [N/A]

    Justification: There is no societal impact of the work performed.

    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: The paper poses no such risks.

    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 used public APIs for the datasets.

    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 detailed README file alongside the code.

    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: The paper does not involve crowdsourcing nor research with 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: The paper does not involve crowdsourcing nor research with 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: [Yes]

Justification: We described how the LLM is integrated in our framework in detail.

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.