Benchmarking Multi-Modal Graph-based Social Media Popularity Prediction

Utkarsh Sahu[1] **Zhisheng Qi**[2] **Li Zhu**[2] **Yizhao Yang**[1] **Jun Li**[1]
Ryan Rossi[3] **Yu Wang**[2]

[1]**University of Oregon** [2]**University of Georgia** [3]**Adobe Research**

{utkarsh, yizhao, lijun}@uoregon.edu
{zq03788, ryanlizhu, Yu.Wang6}@uga.edu
ryrossi@adobe.com

Abstract

Social media popularity prediction aims to forecast the future reach or influence of online content from early-stage observations. Accurate prediction enables key downstream applications, such as advertising optimization and strategic content planning by users, creators, and platforms. Despite substantial progress, existing popularity prediction works often fail to jointly consider multimodal content and temporal social interaction signals. Moreover, the literature remains highly fragmented across datasets, modalities, observation windows, prediction targets, and evaluation protocols. This fragmentation prevents fair comparison and obscures a systematic understanding of how textual, visual, temporal, and interaction-based signals jointly shape popularity dynamics. To address these challenges, we introduce `MMG-Pop`, a **M**ulti-**m**odal **G**raph-based **Pop**ularity Prediction benchmark, which unifies datasets, modalities, temporal interaction signals, and representative baselines under a standardized evaluation protocol. Furthermore, we propose `MMG-PopNet`, a unified multi-modal graph-based network that jointly models the aforementioned multi-modal signals and graph-structured social interactions. Extensive experiments on MMG-Pop, comprising four datasets across Bluesky and Reddit platforms, demonstrate the superior performance of MMG-PopNet and yield new insights into cross-platform training generalization, multi-task prediction benefits, multi-modality contributions, and LLM prediction limitation. These findings establish a unified foundation for future research on social dynamics modeling and intervention under heterogeneous modalities and socially-aware agentic ecosystem paradigms. The MMG-Pop benchmark and MMG-PopNet code are available at this Link.

1 Introduction

Social dynamics refers to the patterns of interactions and relationships among individuals within a society [1, 2, 3], emerging across diverse real-world contexts such as public health behavior change, collective responses during crises, and political mobilization [4, 5, 6, 7]. Accurately modeling social dynamics provides critical insights for analyzing, anticipating, and potentially intervening in collective social behaviors (e.g., early detection of toxic information cascades and timely intervention strategies on online platforms) [8, 9, 10, 11, 12] In this work, we focus on one of the most important social dynamics modeling problems, social media popularity prediction, which aims to leverage early observations of social content to predict its future popularity/influence (e.g., number of likes, reposts) across diverse social contexts and modalities [13, 14, 15]. Effectively predicting popularity on social media has important implications for both platforms and users. For platforms, it supports content recommendation, trend forecasting, advertising, and efficient allocation of moderation resources [16, 17, 18, 19] by estimating which content is likely to attract substantial future engagement. For users, creators, and organizations, it helps estimate future reach, optimize posting content and promotion strategies [20, 21], and plan social or marketing campaigns more effectively [22, 23].

Prior social media popularity prediction can be categorized into three lines. The first predicts social media popularity based on the initial media content without considering its subsequent spread [15, 38, 39, 26, 40, 41]. However, they are content-centric and fail to account for how early social interaction

Existing Work	Social Modality Signal					Popularity Metric	Popularity Prediction Method
	Graph	Text	Image	Time	Video		
@Username [24]	✔	✘	✘	✔	✘	Speed/Width/Depth	Cox PH + Log-linear Regression
Resubmissions [25]	✘	✔	✘	✔	✘	Reddit Karma	Supervised LDA + Linear Regression
Szabo-Huberman [15]	✘	✘	✘	✔	✘	View Count/ Digg votes	Log-linear Regression
Flickr-SVR [26]	✘	✔	✔	✘	✘	View Count	Support Vector Regression
SEISMIC [27]	✘	✘	✘	✔	✘	Cascade size	Self-exciting point process
Galton–Watson [28]	✔	✘	✘	✔	✘	Cascade size	Branching Hawkes process
HIP [25]	✘	✔	✘	✔	✘	View Count	Hawkes Intensity Processes
DeepHawkes [29]	✔	✘	✘	✔	✘	Cascade size	GRU + Time Decay
DeepCas [30]	✔	✘	✘	✘	✘	Cascade size	Bi-directional GRU
CasSeqGCN [31]	✔	✘	✘	✔	✘	Cascade size	GCN + LSTM
TSGNN [32]	✔	✘	✘	✘	✘	Cascade size	GAT + GLU
GraphLSTM [33]	✔	✔	✘	✔	✘	Reddit Karma	Graph-structured LSTM
UHAN [34]	✘	✔	✔	✘	✘	View Count	Multi-modal Attention
HMMVED [35]	✘	✔	✔	✘	✔	Comment/Repost/Likes/View	Hierarchical Multimodal VAE
MMRA [36]	✘	✔	✔	✘	✔	View Count	Multi-modal Attention + Retrieval
MASSL [37]	✘	✔	✔	✘	✔	View Count	Multimodal VAE

Table 1: Prior social media popularity prediction methods, organized by modality signal, prediction metric, modeling approach, and method category: feature engineering, statistical, and deep learning.

dynamics, such as high-impact comments and influencers' reshares [42], impact eventual popularity. The second leverages observed interaction patterns (e.g., reshares, replies, and user engagements) to model how a post propagates through the social interaction network. These interaction histories are represented as cascades and modeled using point processes or geometric deep learning (e.g., RNNs/GNNs) to capture the temporal/structural dynamics of information diffusion [27, 28, 29, 30, 31]. However, they under-utilize the semantic content of posts and responses, such as textual meaning and visual signals that shape engagement. A third line of work jointly leverages both post content and interaction structure [43, 33, 44, 45]. However, these studies primarily focus on tasks such as toxicity detection, sentiment analysis, rumor propagation, or conversation modeling, rather than directly addressing popularity prediction. More recently, emerging generative models and agentic AI have been applied to social dynamics modeling, either through LLM-based multi-agent simulations [46, 47] or by directly repurposing LLMs as autoregressive cascade predictors or reasoning-augmented regressors for popularity forecasting [48, 49]. However, none of these approaches has been systematically benchmarked against prior non-LLM baselines.

In addition to the above limitations, existing social media popularity prediction studies are constructed under heterogeneous yet inconsistent experimental settings as in Table 1. These differences span dataset versions (e.g., X [29, 30, 31] versus (vs.) Reddit [28, 33]), modality signals (e.g., text, image and video [37, 36, 35] vs. graph topology [30, 29, 32]), prediction targets (e.g., cascade size [32, 27, 31] vs. content view count [15, 34]). This fragmentation prevents meaningful comparison of existing methods, thereby hindering the derivation of key insights, such as which modalities contribute most and whether cross-platform transferability exists. Furthermore, existing popularity prediction benchmarks [50, 51] mainly focus on initial media content, overlooking the evolving cascade of user interactions over time. This motivates us to develop both a unified benchmark `MMG-Pop` and a popularity prediction model `MMG-Pop-Net` that jointly capture multi-modal content and temporal social interactions. The key contributions are summarized as follows:

- **Unified Popularity Prediction Benchmark.** We introduce the `MMG-Pop` benchmark, standardizing datasets, social modality signals, observation windows, prediction horizons, and popularity measures to enable consistent evaluation across in-domain forecasting, future-horizon prediction, and cross-platform transfer, with representative baselines.
- **Unified Multi-Modal Model.** We propose `MMG-Pop-Net`, the first unified architecture to jointly model multimodal content, graph-structured interaction dynamics, and temporal signals through bidirectional graph message passing, supporting multi-objective popularity prediction.
- **Comprehensive Experiments and Novel Insights.** We conduct extensive experiments to demonstrate the advantages of MMG-Pop-Net and the insights enabled by the MMG-Pop benchmark, highlighting the importance of jointly modeling multimodal content and cascade structure, the generalization gains from cross-community training, the benefits of multi-task training for engagement prediction, and the limited ability of LLMs to predict social popularity.

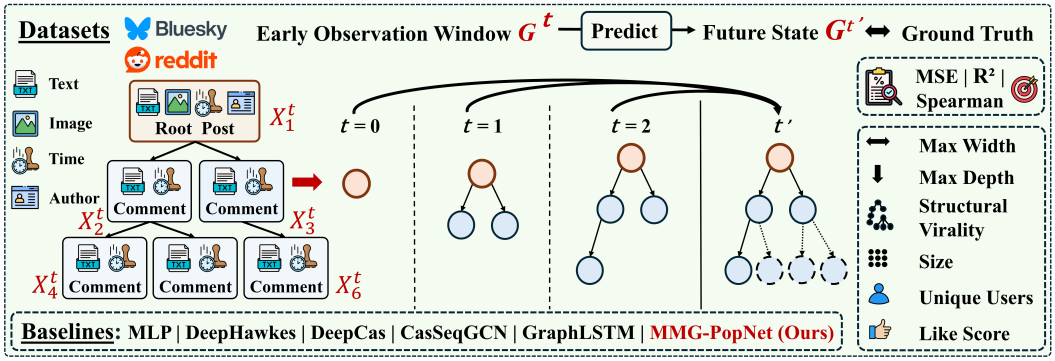


Figure 1: **Overview of MMG-Pop Benchmark.** Social cascades from Bluesky and Reddit are represented as tree-structured graphs, where each node carries multi-modal attributes. Given only an early observed prefix G^t, the task is to predict six complementary popularity dimensions characterizing the future cascade state $G^{t'}$. The benchmark evaluates baselines alongside our proposed `MMG-PopNet` across multiple observation windows.

2 Design Space of MMG-Pop Popularity Prediction Benchmark

This section outlines the design space of our proposed `MMG-Pop` benchmark for social media popularity prediction, encompassing the notation and problem formulation, popularity measurement, training and evaluation, dataset curation, and existing baselines.

2.1 Notation and Problem Formulation

Notation of Social Media Popularity Prediction. We represent an online social cascade as a directed tree-structured graph $G = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ denotes the set of nodes and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ denotes the set of directed edges. Each node $v \in \mathcal{V}$ corresponds to a content item (e.g., a post, comment, or reply), and a directed edge $(u, v) \in \mathcal{E}$ indicates that node v is generated in response to node u, capturing the information propagation (e.g., reply-to, repost/reshare, or quote relationships). The graph is rooted at a unique node $v_{\text{root}} \in \mathcal{V}$ representing the initial item of the cascade (e.g., a starter post on Bluesky or a submission on Reddit). Each node $v \in \mathcal{V}$ is associated with multimodal attributes $X_v = (X_v^{\text{Text}}, X_v^{\text{Visual}}, X_v^{\text{Graph}}, X_v^{\text{Social}}, X_v^{\text{Time}})$ where X_v^{Text} denotes textual content features (e.g., post/comment text, hashtags, or semantic embeddings [34, 25]), X_v^{Visual} denotes visual features when media is present (e.g., images, videos, or visual descriptors [26, 37]), X_v^{Graph} denotes local structural or network context (e.g., neighborhood subgraph statistics or position within the cascade [33]), X_v^{Social} denotes author-level social context features (e.g., user profile attributes [52] and interaction graph-derived influence proxies, including centrality and PageRank scores [53, 54]), and X_v^{Time} denotes temporal features (e.g., global timestamp or relative time to parent nodes within the cascade [33]). In addition, each cascade may be associated with thread-level contextual metadata X^{Thread}, capturing properties of the root post v_{root} (e.g., topic, presence of visual media, or root-author follower count).

Formulation of Social Media Popularity Prediction. The core objective is to predict the future evolution of a social cascade given only its early-stage observations. Let $t \geq 0$ denote the elapsed time since the root post. Given a cascade graph $G = (\mathcal{V}, \mathcal{E})$, we define the observed prefix at time t as $G^t = (\mathcal{V}^t, \mathcal{E}^t)$, where $\mathcal{V}^t = \{v \in \mathcal{V} \mid t_v \leq t\}$ and $\mathcal{E}^t$ contains all edges among nodes in $\mathcal{V}^t$. This prefix captures the historical context of the cascade, including time-truncated multimodal node attributes $\{X_v^t\}_{v \in \mathcal{V}^t}$ and thread-level context $X^{\text{Thread},t}$ observable up to time t. The prediction target is the future state of the cascade, $\mathbf{Y}_G \in \mathbb{R}^K$, consisting of K popularity measures defined in Section 2.2. We aim to learn a parametric mapping $\mathcal{F}_{\boldsymbol{\Theta}} : \left(G^t, \{X_v^t\}_{v \in \mathcal{V}^t}, X^{\text{Thread},t}\right) \mapsto \mathbf{Y}_G$.

2.2 Popularity Measurement

Social popularity can be quantified in multiple ways. Our benchmark, MMG-Pop, considers six distinct dimensions to comprehensively capture popularity dynamics, following prior literature [55, 56, 57, 15]. We categorize these into **STRUCTURAL**, **PARTICIPATION**, and **ENGAGEMENT** tasks:

- **Max Width:** It measures the largest breadth of the cascade [55, 57]. It is quantified as the maximum number of nodes appearing at any depth level in the cascade graph G.
- **Max Depth:** It measures the length of the longest reply chain in the cascade [57, 15]. It is quantified as the maximum distance from v_{root} to any node in the cascade graph G.

- **Structural Virality:** It quantifies whether cascade diffusion is dominated by shallow broadcast spread or deeper multi-hop propagation [56]. It is measured as the average shortest-path distance between all pairs of distinct nodes in the cascade graph.
- **Size:** It measures how many content items are generated in the cascade [57, 30]. It is quantified as the number of nodes in the cascade graph G, i.e., $|\mathcal{V}|$.
- **Unique Users:** It measures the number of unique users who participate in the cascade [57].
- **Like Score:** It captures platform-visible engagement received by the root post content v_{root} through platform-specific metrics such as likes or up-votes (e.g., Reddit Karma score) [15].

2.3 Training and Evaluation

Training. Let $\mathcal{G}$ denote the set of cascades, partitioned into $\mathcal{G}^{\text{Train}} \cup \mathcal{G}^{\text{Val}} \cup \mathcal{G}^{\text{Test}}$. For each cascade $G \in \mathcal{G}^{\text{Train}}$, we construct its observed prefix G^t with truncated node/thread-level features $\{X_v^t\}$ and $X^{\text{Thread},t}$. The objective is to predict future outcomes at a later time $t' > t$. Given heavy-tailed nature of social popularity signals [58, 59], we define popularity targets in the log-transformed space: $\widetilde{\mathbf{Y}}_G^{t'} = \log(\mathbf{I} + \mathbf{Y}_G^{t'})$. The model is trained to predict $\widetilde{\mathbf{Y}}_G^{t'}$ by optimizing: $\mathbf{\Theta}^* = \arg\min_{\mathbf{\Theta}} \sum_{G \in \mathcal{G}^{\text{Train}}} \mathcal{L}(\mathcal{F}_{\mathbf{\Theta}}(G^t, \{X_v^t\}_{v \in \mathcal{V}^t}, X^{\text{Thread},t}), \widetilde{\mathbf{Y}}_G^{t'})$ where $\mathcal{L}$ is the mean squared error (MSE) loss: $\mathcal{L}(\widehat{\mathbf{Y}}, \widetilde{\mathbf{Y}}) = \frac{1}{d}||\widehat{\mathbf{Y}} - \widetilde{\mathbf{Y}}||_2^2$, with d denoting the number of prediction targets.

Evaluation. During evaluation, the learned model $\mathcal{F}_{\mathbf{\Theta}}^*$ is applied to unseen cascades $G \in \mathcal{G}^{\text{Test}}$, using only their observed prefixes G^t and corresponding features. We evaluate performance by comparing the predicted future cascade dynamics at time t' with the corresponding ground-truth values. We report MSE, R^2, and Spearman correlation, all computed in the log-transformed space.

2.4 Dataset Curation

We curate social cascades from Bluesky and Reddit, two platforms with distinct platform dynamics. Each discussion thread is represented as a tree-structured social cascade following the formulation in Section 2.1, where the root post is v_{root}, posts or comments are nodes in $\mathcal{V}$, and parent–reply relations define directed edges $(u, v) \in \mathcal{E}$. Node attributes X_v and thread-level context X^{Thread} are instantiated from the available platform metadata.

Bluesky. We curate our Bluesky subset from the large-scale collection of [60], which originally contains approximately 235 million posts from 4 million users between February 2023 and March 2024. We construct cascades from reply interactions, which provide explicit conversational content for modeling discussion dynamics.

Reddit. We use Pushshift data dumps [61] from three communities: r/AMA, r/Gaming, and r/Futurology. These subreddits capture complementary discussion styles: centralized Q&A, media-rich entertainment discussion, and speculative scientific discourse. We construct cascades from submissions and their comment reply trees. The datasets cover July 2021–December 2024 for r/AMA, January 2023–August 2024 for r/Gaming, and August 2019–December 2024 for r/Futurology. Detailed dataset information is provided in Appendix A.

2.5 Representative Baselines

We evaluate representative baselines spanning structure-agnostic, temporal, sequence-based, graph-based, and content-aware cascade modeling baselines. **MLP** is a structure-agnostic baseline that represents each cascade using root-post features, aggregated reply features, and global thread metadata. **DeepHawkes** [29] captures temporal diffusion dynamics through user embeddings, diffusion-path encoding, and time-decay modeling. **DeepCas** [30] models cascades as sampled diffusion paths and learns sequence representations with attention. **CasSeqGCN** [31] represents temporal graph evolution by encoding graph snapshots with GCN to model time progression. **GraphLSTM** [33] is a content-aware graph sequence baseline that models reply-tree structure with textual, user, temporal, and structural features. Additional details are provided in the Appendix B.

3 Foundational Multi-modal Graph-based Popularity Prediction Network

This section proposes a unified framework that integrates multimodal content and graph-structured social interactions for social dynamics prediction, as shown in Figure 2. Our architecture first encodes heterogeneous multimodal cascade content (text semantics, visual media, and global context) into a unified representation, and then applies graph message passing to incorporate cascade social interaction structure, yielding a shared representation for multi-task popularity prediction.

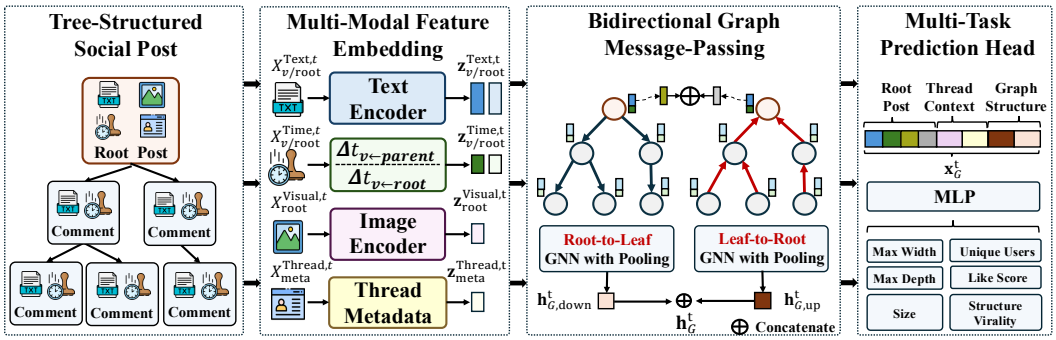


Figure 2: **Overview of MMG-PopNet Model:** The model embeds node-level text and temporal signals for bidirectional graph message passing over the cascade, and root visual content and thread metadata are encoded as separate contextual features. The learned root, graph, visual, and metadata representations are fused at the prediction stage to support multi-task popularity forecasting.

Multi-Modal Feature Embedding. To encode heterogeneous cascade attributes, we design modality-specific encoders tailored to the semantics of each signal. For each node $v \in \mathcal{V}^t$, we encode the available node-level textual and temporal attributes. For textual content X_v^{Text}, we use Transformer to obtain $\mathbf{z}_v^{\text{Text},t} = F^{\text{Text}}_{\mathbf{\Theta}^{\text{Text}}_{\text{Enc}}}(X_v^{\text{Text},t})$. For temporal information, we encode relative timing signals via $\Delta_v^{\text{Time}} = [\Delta t_{v\leftarrow\text{parent}}, \Delta t_{v\leftarrow\text{root}}]$, where $\Delta t_{v\leftarrow\text{parent}}$ and $\Delta t_{v\leftarrow\text{root}}$ denote the elapsed time since the parent node and the root post, reflecting the immediacy of user engagement and overall temporal stage. These timing signals are log-transformed and z-score normalized to obtain fixed temporal features $\mathbf{z}_v^{\text{Time},t}$. We also encode thread-level context as $X^{\text{Thread},t} = (X_{\text{root}}^{\text{Visual},t}, X_{\text{meta}}^{\text{Thread},t})$, denoting root post visual content and non-visual thread context related to the root post. The root image is encoded with a CLIP [62], $\mathbf{z}_{\text{root}}^{\text{Visual},t} = F^{\text{Visual}}_{\mathbf{\Theta}^{\text{Visual}}_{\text{Enc}}}(X_{\text{root}}^{\text{Visual},t})$. Finally, non-visual thread context is modeled with initial user's influence and posting time, i.e., $X_{\text{meta}}^{\text{Thread},t} = [\text{Followers}(v_{\text{root}}), \phi(t_{\text{post}})]$, where the follower count is log-transformed and standardized, and $\phi(\cdot)$ is a cyclic encoding of time-of-day as content posting time during a day also influences interactions. The resulting root visual $\mathbf{z}_{\text{root}}^{\text{Visual},t}$ and non-visual $X_{\text{meta}}^{\text{Thread},t}$ representations together concatenate into the thread-level contextual representation $\mathbf{z}^{\text{Thread},t}$.

Bidirectional Graph Message-Passing. While node-level textual and temporal encodings capture rich local signals at each node, modeling the structural context of the cascade is essential to understand how information propagates and evolves over time. In particular, early-stage cascade dynamics, such as branching patterns and response depth, provide strong indicators of future popularity. To encode such structural dependencies, we perform graph message-passing over the cascade. We first initialize each node $v \in \mathcal{V}^t$ representation by concatenating semantic and temporal features: $\mathbf{h}_v^{(0)} = \text{Concat}\big(\mathbf{z}_v^{\text{Text}}, \mathbf{z}_v^{\text{Time}}\big)$, capturing both content and temporal context. We then apply bidirectional message-passing to simultaneously model root-to-leaf and leaf-to-root propagation. The node embeddings are initialized in both directions $\mathbf{h}_{v,\text{down}}^{(0)} = \mathbf{h}_{v,\text{up}}^{(0)} = \mathbf{h}_v^{(0)}$ are updated at layer ℓ:

$$\mathbf{h}_{v,\text{down}}^{(\ell)} = \text{SAGE}_{\text{down}}^{(\ell)}(\mathbf{h}_{v,\text{down}}^{(\ell-1)}, \{\mathbf{h}_{u,\text{down}}^{(\ell-1)} : u \in \mathcal{N}_{\text{down}}(v)\}), \mathbf{h}_{v,\text{up}}^{(\ell)} = \text{SAGE}_{\text{up}}^{(\ell)}(\mathbf{h}_{v,\text{up}}^{(\ell-1)}, \{\mathbf{h}_{u,\text{up}}^{(\ell-1)} : u \in \mathcal{N}_{\text{up}}(v)\}), \quad (1)$$

where $\mathcal{N}_{\text{down}}(v)$ and $\mathcal{N}_{\text{up}}(v)$ denote the parent/child-side neighbors of node v. After L layers, we aggregate node representations via mean pooling to obtain direction-specific summaries $\mathbf{h}_{G,\text{down}}^t, \mathbf{h}_{G,\text{up}}^t$, which are then concatenated to form the final graph representation: $\mathbf{h}_G^t = \text{Concat}(\mathbf{h}_{G,\text{down}}^t, \mathbf{h}_{G,\text{up}}^t)$.

Multi-Modal Feature Fusion and Multi-Task Prediction. After encoding node/graph/thread-level signals, we fuse them into a unified cascade representation for prediction. Specifically, we aggregate (1) the raw root node representation $\mathbf{h}_{\text{root}}^0$ before message passing, (2) the final bidirectional root representation, $\mathbf{h}_{\text{root}}^{(L)} = \text{Concat}(\mathbf{h}_{\text{root,down}}^{(L)}, \mathbf{h}_{\text{root,up}}^{(L)})$, (3) the graph structural representation $\mathbf{h}_G^t$, and (4) the thread-level contextual representation, into a single vector: $\mathbf{x}_G^t = \mathbf{h}_{\text{root}}^{(0)} \parallel \mathbf{h}_{\text{root}}^{(L)} \parallel \mathbf{g}_G^t \parallel \mathbf{z}^{\text{Thread},t}$. We then apply a shared multi-layer perceptron (MLP) to map the fused representation into a multi-task output space: $\widehat{\mathbf{Y}}_G^{t'} = \text{MLP}_{\mathbf{\Theta}_{\text{Pred}}}(\mathbf{x}_G^t)$, where $\widehat{\mathbf{Y}}_G \in \mathbb{R}^K$ contains log-space predictions for K targets (e.g., final cascade size, unique users, or structural properties). The model is trained using a multi-task objective $\mathcal{L} = \frac{1}{K}\sum_{k=1}^{K} \mathcal{L}_k(\widehat{Y}_G^{(k)}, Y_G^{(k)})$, where all parameters (including the text encoder $\mathbf{\Theta}^{\text{Text}}_{\text{Enc}}$, image encoder $\mathbf{\Theta}^{\text{Image}}_{\text{Enc}}$, two GNNs, and prediction heads $\mathbf{\Theta}_{\text{Pred}}$) are jointly optimized end-to-end.

4 Related Work

Social Dynamics Modeling. Social dynamics modeling studies how local interactions among individuals give rise to collective outcomes such as consensus, segregation, polarization, and information diffusion [63, 64]. Classical models explain these phenomena through simple but expressive mechanisms, where [65] showed how individual preferences can produce macro-level segregation, while threshold and cascade models describe how behaviors spread once social reinforcement exceeds adoption barriers [66, 67]. Opinion dynamics and social-influence models further categorize how network structure, homophily, and repeated exposure shape agreement, diversity, and polarization [68, 69, 64]. With online platforms, this perspective has expanded to large-scale information diffusion, misinformation spread, and intervention analysis, where temporal interactions and network topology jointly determine collective trajectories [70, 8, 7]. Recent agent-based and LLM-driven simulations enrich this line by modeling adaptive, language-mediated agents to simulate broad social dynamics [47, 71].

Social Media Popularity Prediction. Social media popularity prediction models and forecasts social dynamics by predicting the future influence, engagement, or diffusion of online content [15, 6]. It has broad applications in trend forecasting, advertisement targeting, and public opinion analysis. Existing studies can be categorized based on social signals and modeling paradigms. From the signal perspective, prior work includes content-based methods leveraging textual or visual information [15, 38, 39, 26, 40, 41], structure-based methods modeling diffusion topology and user interactions [27, 28, 29, 30, 31], and hybrid approaches combining both [43, 33, 44, 45]. From the modeling perspective, earlier studies mainly relied on handcrafted features with classical machine learning [27, 28, 25], followed by sequential and geometric deep learning methods, including LSTMs and GNNs, to capture temporal and structural dynamics [29, 30, 31, 32, 33, 34, 35, 36, 37]. More recently, agentic social simulation approaches [46, 47, 48, 49] have been explored to model contextual user behaviors and social interactions. However, they remain fragmented across modalities, platforms, prediction targets, and evaluation protocols, motivating the need for a unified benchmark.

5 Experiment

We conduct extensive experiments with the MMG-Pop benchmark, designed to systematically address the following research questions:

$\mathbf{Q}_1$: *How do different baselines and MMG-Pop-Net perform in predicting cascade popularity across varying future horizons and early observation windows under our MMG-Pop benchmark?*

$\mathbf{Q}_2$: *Does unified training across communities and platforms improve social popularity prediction?*

$\mathbf{Q}_3$: *Can MM-LLMs serve as competitive predictors for multimodal social popularity forecasting?*

$\mathbf{Q}_4$: *Does multi-objective prediction benefit from jointly predicting popularity targets?*

$\mathbf{Q}_5$: *How do different modalities contribute to popularity prediction?*

Detailed experiment settings are described in Appendix C.

5.1 $\mathbf{Q}_1$: Popularity Prediction across Varying Future Horizons/Early Observation Windows.

Popularity prediction of final cascade state under different early observation. Given an early observation window t, we aim to predict the final popularity of a social cascade based on the cascade state observed up to the data collection time. To assess the role of early information, we consider a root-only setting and three early-observation windows for each dataset. The root-only setting, denoted as window 0, includes only the root post and its thread-level context. The remaining windows capture increasingly mature cascade prefixes: 2, 10, and 20 minutes for Bluesky; 15, 30, and 60 minutes for r/AMA; 20, 50, and 90 minutes for r/Gaming; and 30, 90, and 180 minutes for r/Futurology. Window lengths vary by dataset because cascades unfold at different speeds across platforms and communities. Table 2 reports MSE in the log-transformed target space across all datasets, observation windows, and prediction targets. MMG-PopNet achieves the best overall average MSE for every target category, with 4.6% to 17.0% reductions over the strongest non-MMG-PopNet baselines. Among the baselines, Graph-LSTM is the strongest competitor on most structural targets, reflecting the value of reply-tree structure, temporal ordering, and textual content. CasSeqGCN also performs competitively, likely by modeling evolving cascade as snapshots, which capture propagation topology. MLP is particularly strong for LIKE SCORE prediction because this target measures engagement received by the root post, whose features are directly combined with the mean representation of early observed nodes. However, without message passing, MLP cannot fully model cascade-level dependencies, whereas MMG-PopNet integrates these early signals through bidirectional graph propagation and achieves the lowest average MSE across all targets.

Table 2: MSE results for final cascade-state prediction under different early observation windows, covering **STRUCTURAL PREDICTION TASKS** (max width, max depth, structural virality, and cascade size), **UNIQUE-USER PREDICTION**, and **LIKE-SCORE PREDICTION**. Lower values indicate better performance. The best results are highlighted in **bold**, while the second-best results are underlined. Statistical significance analyses show that improvements are significant across settings in Table 11.

Task	Model	Bluesky					r/AMA					r/Gaming					r/Futurology					Avg
		0	2	10	20	Avg	0	15	30	60	Avg	0	20	50	90	Avg	0	30	90	180	Avg	
MAX WIDTH	MLP	0.458	0.427	0.371	0.350	0.402	0.664	0.528	0.493	0.415	0.525	1.755	1.182	0.877	0.708	1.128	1.810	1.341	0.697	0.457	1.076	0.783
	DeepHawkes	0.684	0.617	0.403	0.338	0.510	0.713	0.554	0.491	0.365	0.531	1.888	1.268	1.147	0.839	1.288	1.876	1.643	1.184	0.783	1.371	0.925
	DeepCas	0.633	0.675	0.676	0.676	0.666	0.841	0.757	0.767	0.709	0.769	2.361	1.895	1.833	1.690	1.945	1.738	1.643	1.190	1.101	1.443	1.199
	CasSeqGCN	0.676	0.538	0.357	0.286	0.465	0.713	0.534	0.446	0.313	0.502	1.864	1.140	0.841	0.587	1.108	1.820	1.258	0.679	0.366	1.031	0.776
	Graph-LSTM	0.657	0.526	0.345	0.281	0.452	0.686	0.510	0.430	0.305	0.483	1.771	1.155	0.811	**0.518**	1.064	1.749	1.149	**0.631**	**0.345**	0.969	0.742
	MMG-PopNet	**0.457**	**0.369**	**0.284**	**0.234**	**0.336**	**0.625**	**0.491**	**0.429**	**0.276**	**0.455**	**1.557**	**1.096**	**0.714**	0.562	**0.982**	**1.581**	**1.132**	0.665	0.372	**0.938**	**0.678**
MAX DEPTH	MLP	**0.356**	0.346	0.296	0.270	0.318	0.318	0.267	0.225	0.198	0.252	0.344	0.279	0.214	0.186	0.256	0.593	0.484	0.289	0.221	0.397	0.305
	DeepHawkes	0.367	0.358	0.344	0.300	0.342	0.329	0.280	0.253	0.240	0.276	0.346	0.290	0.269	0.229	0.284	0.602	0.567	0.429	0.325	0.482	0.345
	DeepCas	0.361	0.364	0.364	0.364	0.363	0.432	0.375	0.341	0.333	0.370	0.577	0.409	0.387	0.340	0.427	0.596	0.554	0.410	0.374	0.483	0.411
	CasSeqGCN	0.364	0.348	0.294	0.258	0.316	0.328	0.281	0.227	0.193	0.257	0.347	0.296	0.247	0.210	0.275	0.586	0.477	0.317	0.230	0.403	0.313
	Graph-LSTM	0.364	0.340	0.285	0.243	0.308	0.323	0.264	**0.215**	0.176	0.245	0.353	0.288	0.209	0.161	0.253	0.570	0.453	0.298	0.232	0.388	0.298
	MMG-PopNet	0.363	**0.325**	**0.273**	**0.238**	**0.300**	**0.314**	**0.256**	0.219	**0.172**	**0.240**	**0.335**	**0.275**	**0.200**	**0.160**	**0.243**	**0.517**	**0.418**	**0.282**	**0.205**	**0.356**	**0.284**
STRUCTURAL VIRALITY	MLP	**0.147**	0.143	0.123	0.114	0.132	**0.129**	0.105	0.089	0.074	0.099	0.096	**0.078**	0.059	0.049	0.071	0.209	0.164	0.102	0.077	0.138	0.110
	DeepHawkes	0.158	0.154	0.144	0.119	0.144	0.134	0.111	0.101	0.094	0.110	**0.095**	0.085	0.083	0.069	0.082	0.200	0.195	0.157	0.127	0.170	0.127
	DeepCas	0.155	0.157	0.157	0.157	0.157	0.218	0.168	0.146	0.141	0.167	0.265	0.154	0.140	0.117	0.169	0.213	0.200	0.165	0.145	0.181	0.169
	CasSeqGCN	0.157	0.148	0.124	0.109	0.135	0.134	0.111	0.089	0.070	0.102	**0.095**	0.086	0.074	0.066	0.080	0.196	0.162	0.116	0.081	0.139	0.114
	Graph-LSTM	0.157	0.145	0.122	0.104	0.132	0.131	0.106	**0.085**	0.065	0.097	0.103	0.083	0.060	**0.044**	0.073	0.195	0.157	0.106	0.084	0.136	0.109
	MMG-PopNet	0.155	**0.135**	**0.114**	**0.101**	**0.126**	0.130	**0.100**	0.087	**0.064**	**0.095**	0.098	0.081	**0.057**	0.045	**0.070**	0.182	**0.146**	**0.101**	**0.073**	**0.126**	**0.104**
SIZE	MLP	**0.695**	0.652	0.556	0.515	0.605	0.918	0.711	0.626	0.519	0.689	2.018	1.399	0.993	0.801	1.303	2.747	2.094	1.066	0.637	1.637	1.059
	DeepHawkes	0.899	0.817	0.618	0.499	0.709	0.982	0.738	0.635	0.478	0.709	2.137	1.472	1.314	0.939	1.466	2.903	2.556	1.904	1.159	2.130	1.253
	DeepCas	0.832	0.879	0.880	0.880	0.868	1.262	1.093	1.053	1.002	1.103	2.952	2.228	2.152	1.918	2.313	2.673	2.517	1.767	1.605	2.016	1.606
	CasSeqGCN	0.881	0.751	0.551	0.459	0.660	0.979	0.735	0.599	0.421	0.684	2.101	1.400	1.021	0.715	1.309	2.760	2.015	1.124	0.583	1.621	1.068
	Graph-LSTM	0.859	0.733	0.531	0.446	0.642	0.945	0.700	**0.572**	0.405	0.656	2.031	1.418	0.988	**0.641**	1.270	2.640	1.852	1.052	0.563	1.527	1.024
	MMG-PopNet	0.705	**0.587**	**0.470**	**0.397**	**0.540**	**0.863**	**0.661**	0.573	**0.366**	**0.616**	**1.829**	**1.317**	**0.842**	0.653	**1.160**	**2.356**	**1.743**	**0.997**	**0.559**	**1.414**	**0.932**
UNIQUE USERS	MLP	**0.465**	0.432	0.374	0.351	0.405	0.657	0.531	0.502	0.418	0.527	1.883	1.306	0.949	0.756	1.224	2.139	1.642	0.847	0.512	1.290	0.860
	DeepHawkes	0.734	0.660	0.439	0.371	0.551	0.707	0.559	0.505	0.370	0.533	2.009	1.392	1.257	0.917	1.394	2.228	1.996	1.439	0.896	1.639	1.030
	DeepCas	0.666	0.721	0.723	0.722	0.706	0.869	0.767	0.771	0.716	0.779	2.629	2.019	2.002	1.796	2.112	2.073	1.976	1.381	1.269	1.675	1.319
	CasSeqGCN	0.723	0.584	0.395	0.318	0.505	0.706	0.543	0.468	0.328	0.511	1.979	1.294	0.966	0.671	1.228	2.157	1.583	0.872	0.449	1.265	0.877
	Graph-LSTM	0.700	0.560	0.368	0.296	0.481	0.679	0.514	0.448	0.320	0.490	1.898	1.307	0.928	**0.599**	1.183	2.074	1.441	0.816	0.456	1.197	0.838
	MMG-PopNet	0.467	**0.383**	**0.302**	**0.255**	**0.352**	**0.622**	**0.494**	**0.447**	**0.290**	**0.463**	**1.693**	**1.218**	**0.792**	0.613	**1.079**	**1.859**	**1.374**	**0.785**	**0.439**	**1.114**	**0.752**
LIKE SCORE	MLP	1.311	1.298	1.254	1.232	1.274	1.351	1.284	1.292	1.193	1.280	5.633	5.593	4.991	4.660	5.219	6.462	6.104	4.061	3.746	5.093	3.217
	DeepHawkes	2.532	2.400	2.010	1.855	2.199	1.438	1.411	1.406	1.325	1.395	6.240	6.122	5.973	6.130	6.116	6.885	6.880	6.046	5.304	6.031	3.997
	DeepCas	2.356	2.502	2.507	2.505	2.493	1.444	1.428	1.483	1.435	1.450	6.359	6.099	6.048	6.268	6.194	5.725	5.785	4.576	4.905	5.273	3.839
	CasSeqGCN	2.508	2.266	1.901	1.749	2.104	1.436	1.406	1.400	1.307	1.387	6.206	6.127	6.043	6.004	6.095	6.778	6.604	5.428	5.000	5.452	3.885
	Graph-LSTM	2.467	2.186	1.791	1.607	2.013	1.360	1.311	1.337	1.237	1.311	5.707	5.576	5.302	4.866	5.363	6.239	5.395	4.343	4.393	5.092	3.445
	MMG-PopNet	**1.260**	**1.087**	**1.026**	**0.950**	**1.081**	**1.311**	**1.212**	**1.166**	**1.028**	**1.179**	**5.067**	**4.654**	**4.307**	**4.073**	**4.525**	**4.999**	**4.596**	**3.214**	**2.750**	**3.890**	**2.669**

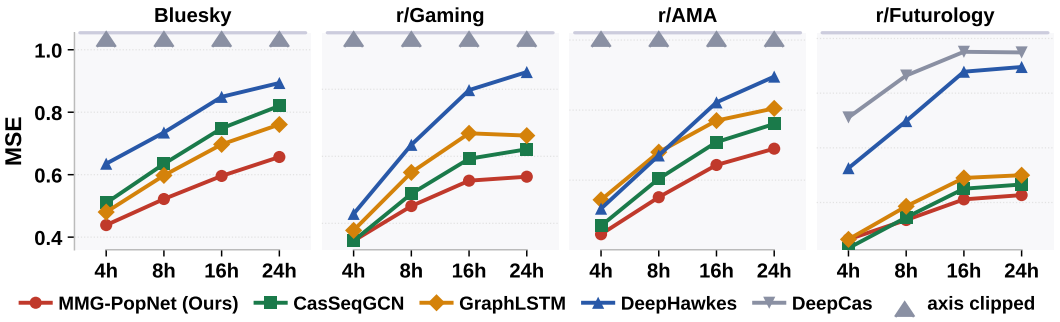


Figure 3: MSE-Loss trajectories across datasets, comparing different models for target SIZE. **Lower is better**. MMG-PopNet achieves the lowest MSE, with the strongest gains at later horizons.

Popularity Prediction of Cascade States at Future Horizons. Beyond prediction at the terminal state, we evaluate social media popularity prediction at intermediate future horizons of cascades. Given an early observed prefix G^t (as described in previous section), each method predicts future cascade outcomes at later times $t' > t$, using targets computed from the cascade state at horizons $\{4\text{h}, 8\text{h}, 16\text{h}, 24\text{h}\}$. This setting tests whether a model can forecast not only terminal popularity, but also the trajectory of cascade growth over time. Figure 3 reports MSE trajectories for the SIZE target across datasets. Prediction error generally increases with the forecasting horizon, reflecting the greater uncertainty of longer-range cascade growth. Despite this increased difficulty, MMG-PopNet consistently achieves the lowest error across datasets and horizons. Graph-LSTM and CasSeqGCN are the closest baselines, with comparable performance at 4h and 8h, but they fall behind at 16h and 24h as forecasting uncertainty increases. In contrast, DeepCas has substantially higher error in several cases, with clipped values indicating that its trajectory predictions fall outside the plotted range. Similar trends are observed for other popularity targets in Appendix D.

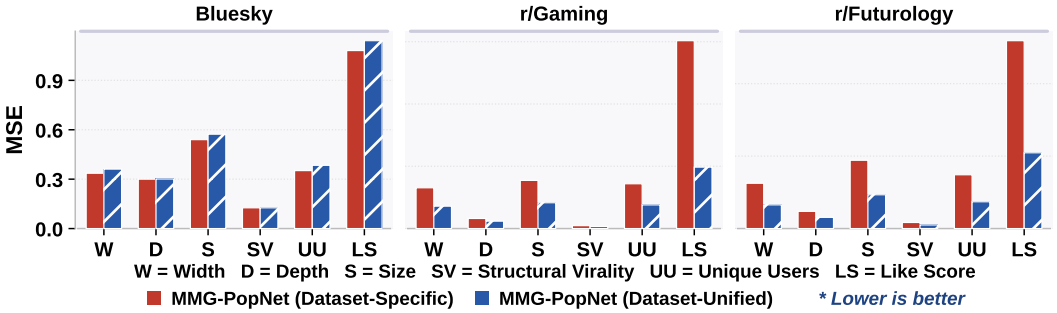


Figure 4: **Dataset-Specific vs. Unified Training.** Avg MSE of MMG-PopNet under dataset-specific and unified training. Lower is better. Unified training greatly improves performance on Reddit communities and remains competitive on Bluesky.

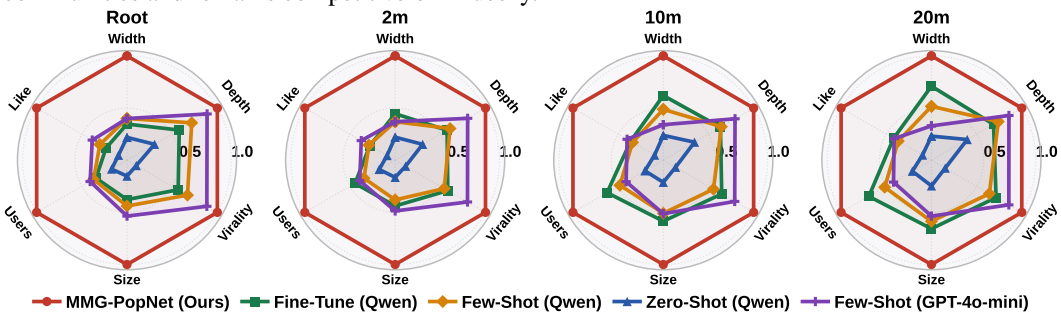


Figure 5: **Normalized LLM Performance on Bluesky.** Scores are normalized with MMG-PopNet as the reference baseline, fixed at 1.0 on all axes, where smaller areas indicate worse performance. MMG-PopNet outperforms LLM baselines across all settings. Among LLMs, retrieval-augmented few-shot prompting performs better in sparse early windows, while fine-tuning becomes stronger as longer cascade prefixes provide richer temporal and structural training signals.

5.2 Q_2: Unified Training Across Communities and Platforms

We investigate whether popularity prediction benefits from unified training across datasets from multiple platforms. Instead of training isolated MMG-PopNet models per dataset and observation window, we train a single model on the combined cascades from all datasets. This evaluates whether joint supervision over heterogeneous cascades improves generalization compared to dataset-specific training. Figure 4 demonstrates that unified training yields substantial performance gains on Reddit while maintaining comparable accuracy on Bluesky. On Reddit, the unified model reduces average MSE across all popularity targets, achieving dramatic error reductions for LIKE SCORE, SIZE, and UNIQUE USERS. Conversely, dataset-specific training retains a marginal edge on Bluesky. This pattern suggests that the unified model benefits most when cross-community training shares a platform-level interaction structure, while Bluesky introduces distinct dynamics less represented in the combined training distribution. Detailed results and analysis can be found in Appendix E.

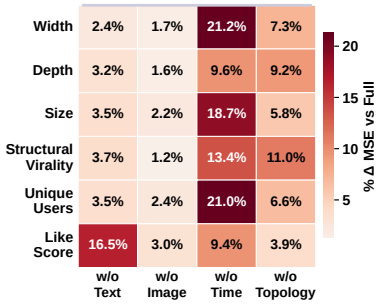
5.3 Q_3: Comparison with LLM-Based Approaches

We compare MMG-PopNet with multimodal LLM models under three settings: zero-shot prompting, retrieval-augmented few-shot prompting, and supervised fine-tuning. Zero-shot setting serialized the observed cascade prefix as a structured JSON input prompt for LLM-based prediction. The retrieval-augmented few-shot setting includes four training examples with similar root posts. Fine-tuning setting trains LLM with early-observation cascade inputs. Figure 5 shows the normalized comparison on Bluesky. Scores are computed as $\mathrm{MSE}_{\mathrm{model}}/\mathrm{MSE}_{\mathrm{MMG-PopNet}}$, so MMG-PopNet forms the reference score of 1.0 on every axis, and smaller polygons indicate worse performance. MMG-PopNet uniformly outperforms all LLM baselines across all targets and observation windows. Zero-shot prompting yields the highest error, proving that direct prompting lacks the numerical calibration of LLMs for popularity prediction despite structured inputs. Few-shot prompting rivals or exceeds fine-tuning given root-only or early windows, where historical examples provide crucial context for sparse cascades. Fine-tuning overtakes prompting as the window expands, likely because it better exploits the richer reply structure, temporal progression, and participation signals in longer observation windows. Overall, MMG-PopNet performs better likely because it models topology, timing features, and multimodal context for popularity prediction more directly, rather than relying on prompt-driven inference over serialized cascade inputs. Detailed setting in Appendix F.

Table 3: Comparison of multi-task versus single-task MSE results across datasets using the following windows: Bluesky @ 10min, r/AMA @ 30min, r/Gaming @ 50min, and r/Futurology @ 90min. **Best in bold.**

Task	Bluesky		r/AMA		r/Gaming		r/Futurology	
	Single	Multi	Single	Multi	Single	Multi	Single	Multi
MAX WIDTH	**0.280**	0.284	**0.411**	0.429	**0.663**	0.714	**0.583**	0.665
MAX DEPTH	**0.272**	0.273	**0.207**	0.219	**0.182**	0.200	**0.272**	0.282
STRUCTURAL VIRALITY	0.115	**0.114**	**0.082**	0.087	**0.050**	0.057	**0.094**	0.101
SIZE	0.494	**0.470**	0.578	**0.573**	0.843	**0.842**	**0.987**	0.997
UNIQUE USERS	**0.295**	0.302	**0.431**	0.447	**0.766**	0.792	**0.774**	0.785
LIKE SCORE	1.575	**1.026**	1.284	**1.166**	4.482	**4.307**	3.708	**3.214**

Figure 6: **Modality Ablation**: Avg. MSE **increase** per excluded modality relative to full MMG-PopNet.



5.4 Q_5: Single-Task vs. Multi-Task Training

We examine whether multi-objective prediction benefits from jointly modeling complementary popularity targets. Here, we compare the multi-task MMG-PopNet with task-specific variants trained independently for each target. Table 3 shows that the benefit of joint training depends on the target. Single-task training is stronger for topology-driven objectives. It achieves lower MSE for MAX WIDTH and MAX DEPTH across all datasets, indicating that these structural properties benefit from dedicated training. Similar trends appear for STRUCTURAL VIRALITY and UNIQUE USERS, although the gaps are smaller. Multi-task training remains competitive for SIZE, matching or slightly improving over single-task models on three datasets. Joint modeling is most beneficial for engagement. Multi-task MMG-PopNet lowers LIKE SCORE MSE on every dataset, with large gains on Bluesky and r/Futurology. This suggests that engagement prediction can benefit from shared signals of cascade structure and user participation. Overall, joint modeling does not improve every target. However, it offers a useful deployment trade-off by remaining competitive on most outcomes while consistently improving LIKE SCORE prediction.

5.5 Q_6: Modality Ablation Analysis

We evaluate how each modality contributes by removing one input source from MMG-PopNet at a time and measuring the relative MSE increase over the full model. The ablations remove textual semantics X^{Text}, root visual content X^{Visual}, node temporal features X^{Time}, or reply-tree topology X^{Graph}.Figure 6 reports results on r/Gaming and r/Futurology, where all four modalities are available. All removals increase error, showing that each modality adds useful information. Temporal features have the largest effect on cascade growth and participation. They encode each node's time since its parent reply and since the root post. Removing these features sharply hurts MAX WIDTH (21.2%), UNIQUE USERS (21.0%), and SIZE (18.7%). These features capture the pace of early discussion. Fast replies signal bursty growth relevant to width and size, while slower temporal medians can indicate longer-lived discussions with more distinct users. Text is most important for engagement. Removing textual semantics increases LIKE SCORE error by 16.5%, while only mildly affecting structural targets. This suggests that audience approval depends strongly on what is said, not only how the cascade grows. Reply-tree topology mainly supports structural prediction, especially STRUCTURAL VIRALITY and MAX DEPTH. Root visual content has the smallest effect, but its consistent gains indicate a modest complementary role. Detailed results in Appendix G.

6 Conclusion and Future Work

In this paper, we introduced MMG-Pop, a unified benchmark for multi-modal social media popularity prediction, and MMG-PopNet, a unified model that captures content, temporal dynamics, and reply structure to forecast multiple forms of popularity. Our experiments show that MMG-Pop enables systematic evaluation across datasets, communities, observation windows, prediction horizons, and engagement targets. Results show that MMG-PopNet improves prediction by jointly modeling multimodal content and cascade structure, with different modalities offering complementary signals. In addition, cross-community training improves generalization, while multi-task training captures shared engagement patterns. In contrast, LLMs remain limited in predicting social popularity, suggesting that language understanding alone is insufficient for modeling social dynamics. Together, these findings establish MMG-Pop as a useful benchmark and MMG-PopNet as an effective model for integrating the signals that shape social media popularity. Furthermore, we have conducted real-world case studies with MMG-PopNet in Appendix H. Future work will explore agentic social simulation to predict popularity. Limitations of this work and additional discussion are in Appendix I.

References

[1] UNESCO. Social dynamics. https://www.unesco.org/en/tags/social-dynamics-0.

[2] Thomas W Farmer, Betsy Talbott, Molly Dawes, Heartley B Huber, Debbie S Brooks, and Emily E Powers. Social dynamics management: What is it and why is it important for intervention? *Journal of Emotional and Behavioral Disorders*, 2018.

[3] William A Brock and Steven N Durlauf. Discrete choice with social interactions. *The Review of Economic Studies*, 2001.

[4] Damon Centola. The spread of behavior in an online social network experiment. *science*, 2010.

[5] Mark EJ Newman. The structure of scientific collaboration networks. *Proceedings of the national academy of sciences*, 2001.

[6] Fan Zhou, Xovee Xu, Goce Trajcevski, and Kunpeng Zhang. A survey of information cascade analysis: Models, predictions, and recent advances. *ACM Computing Surveys (CSUR)*, 2021.

[7] Christopher A Bail, Lisa P Argyle, Taylor W Brown, John P Bumpus, Haohan Chen, MB Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, 2018.

[8] Joseph B Bak-Coleman, Ian Kennedy, Morgan Wack, Andrew Beers, Joseph S Schafer, Emma S Spiro, Kate Starbird, and Jevin D West. Combining interventions to reduce the spread of viral misinformation. *Nature Human Behaviour*, 2022.

[9] Zhe Zhao, Paul Resnick, and Qiaozhu Mei. Enquiring minds: Early detection of rumors in social media from enquiry posts. In *Proceedings of the 24th international conference on world wide web*, 2015.

[10] Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. The spread of low-credibility content by social bots. *Nature communications*, 2018.

[11] Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. Anyone can become a troll: Causes of trolling behavior in online discussions. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, 2017.

[12] Ken-Yu Lin, Roy Ka-Wei Lee, Wei Gao, and Wen-Chih Peng. Early prediction of hate speech propagation. In *2021 International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2021.

[13] Kristina Lerman and Tad Hogg. Using a model of social dynamics to predict popularity of news. In *Proceedings of the 19th international conference on World wide web*, 2010.

[14] Mayank Meghawat, Satyendra Yadav, Debanjan Mahata, Yifang Yin, Rajiv Ratn Shah, and Roger Zimmermann. A multimodal approach to predict social media popularity. In *2018 IEEE conference on multimedia information processing and retrieval (MIPR)*, 2018.

[15] Gabor Szabo and Bernardo A Huberman. Predicting the popularity of online content. *Communications of the ACM*, 2010.

[16] Henrique Pinto, Jussara M Almeida, and Marcos A Gonçalves. Using early view patterns to predict the popularity of youtube videos. In *Proceedings of the sixth ACM international conference on Web search and data mining*, 2013.

[17] Jennifer Cobbe. Algorithmic censorship by social platforms: Power and resistance. 2021.

[18] Linpeng Tang, Qi Huang, Amit Puntambekar, Ymir Vigfusson, Wyatt Lloyd, and Kai Li. Popularity prediction of facebook videos for higher quality streaming. In *2017 USENIX Annual Technical Conference (USENIX ATC 17)*, 2017.

[19] Amin Javari and Mahdi Jalili. Accurate and novel recommendations: an algorithm based on popularity forecasting. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2014.

[20] Masoud Mazloom, Robert Rietveld, Stevan Rudinac, Marcel Worring, and Willemijn Van Dolen. Multimodal popularity prediction of brand-related social media posts. In *Proceedings of the 24th ACM international conference on Multimedia*, 2016.

[21] Zhongping Zhang, Tianlang Chen, Zheng Zhou, Jiaxin Li, and Jiebo Luo. How to become instagram famous: Post popularity prediction with dual-attention. In *2018 IEEE international conference on big data (big data)*. IEEE, 2018.

[22] Bei Yu, Miao Chen, and Linchi Kwok. Toward predicting popularity of social marketing messages. In *International conference on social computing, behavioral-cultural modeling, and prediction*. Springer, 2011.

[23] Peter Van Aelst, Patrick Van Erkel, Evelien D'heer, and Raymond A Harder. Who is leading the campaign charts? comparing individual popularity on old and new media. *Information, communication & society*, 2017.

[24] Jiang Yang and Scott Counts. Predicting the speed, scale, and range of information diffusion in twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, 2010.

[25] Himabindu Lakkaraju, Julian McAuley, and Jure Leskovec. What's in a name? understanding the interplay between titles, content, and communities in social media. In *Proceedings of the international AAAI conference on web and social media*, 2013.

[26] Aditya Khosla, Atish Das Sarma, and Raffay Hamid. What makes an image popular? In *Proceedings of the 23rd international conference on World wide web*, pages 867–876, 2014.

[27] Qingyuan Zhao, Murat A Erdogdu, Hera Y He, Anand Rajaraman, and Jure Leskovec. Seismic: A self-exciting point process model for predicting tweet popularity. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 2015.

[28] Alexey N Medvedev, Jean-Charles Delvenne, and Renaud Lambiotte. Modelling structure and predicting dynamics of discussion threads in online boards. *Journal of Complex Networks*, 2019.

[29] Qi Cao, Huawei Shen, Keting Cen, Wentao Ouyang, and Xueqi Cheng. Deephawkes: Bridging the gap between prediction and understanding of information cascades. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 2017.

[30] Cheng Li, Jiaqi Ma, Xiaoxiao Guo, and Qiaozhu Mei. Deepcas: An end-to-end predictor of information cascades. In *Proceedings of the 26th international conference on World Wide Web*, pages 577–586, 2017.

[31] Yansong Wang, Xiaomeng Wang, Yijun Ran, Radosław Michalski, and Tao Jia. Casseqgcn: Combining network structure and temporal sequence to predict information cascades. *Expert Systems with Applications*, 2022.

[32] Yujia Liu, Kang Zeng, Haiyang Wang, Xin Song, and Bin Zhou. Content matters: A gnn-based model combined with text semantics for social network cascade prediction. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2021.

[33] Victoria Zayats and Mari Ostendorf. Conversation modeling on reddit using a graph-structured lstm. *Transactions of the Association for Computational Linguistics*, 6:121–132, 2018.

[34] Wei Zhang, Wen Wang, Jun Wang, and Hongyuan Zha. User-guided hierarchical attention network for multi-modal social image popularity prediction. In *Proceedings of the 2018 world wide web conference*, 2018.

[35] Jiayi Xie, Yaochen Zhu, and Zhenzhong Chen. Micro-video popularity prediction via multi-modal variational information bottleneck. *IEEE Transactions on Multimedia*, 2021.

[36] Ting Zhong, Jian Lang, Yifan Zhang, Zhangtao Cheng, Kunpeng Zhang, and Fan Zhou. Predicting micro-video popularity via multi-modal retrieval augmentation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024.

[37] Zhuoran Zhang, Shibiao Xu, Li Guo, and Wenke Lian. Multi-modal variational auto-encoder model for micro-video popularity prediction. In *Proceedings of the 8th International Conference on Communication and Information Processing*, 2022.

[38] Roja Bandari, Sitaram Asur, and Bernardo Huberman. The pulse of news in social media: Forecasting popularity. In *Proceedings of the International AAAI Conference on Web and Social Media*, 2012.

[39] Oren Tsur and Ari Rappoport. What's in a hashtag? content based prediction of the spread of ideas in microblogging communities. In *Proceedings of the fifth ACM international conference on Web search and data mining*, 2012.

[40] Francesco Gelli, Tiberio Uricchio, Marco Bertini, Alberto Del Bimbo, and Shih-Fu Chang. Image popularity prediction in social media using sentiment and context features. In *Proceedings of the 23rd ACM international conference on Multimedia*, 2015.

[41] Keyan Ding, Ronggang Wang, and Shiqi Wang. Social media popularity prediction: A multiple feature fusion approach with deep neural networks. In *Proceedings of the 27th ACM International Conference on Multimedia*, 2019.

[42] David Garcia, Pavlin Mavrodiev, Daniele Casati, and Frank Schweitzer. Understanding popularity, reputation, and social influence in the twitter society. *Policy & Internet*, 9(3):343–364, 2017.

[43] Pablo Aragón, Vicenç Gómez, David García, and Andreas Kaltenbrunner. Generative models of online discussion threads: state of the art and research challenges. *Journal of Internet Services and Applications*, 2017.

[44] Justine Zhang, Jonathan Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Dario Taraborelli, and Nithum Thain. Conversations gone awry: Detecting early signs of conversational failure. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018.

[45] Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PloS one*, 2016.

[46] Yijun Liu, Wu Liu, Xiaoyan Gu, Allen He, Weiping Wang, and Yongdong Zhang. Popsim: Social network simulation for social media popularity prediction. *arXiv preprint arXiv:2512.02533*, 2025.

[47] Ziyi Yang, Zaibin Zhang, Zirui Zheng, Yuxian Jiang, Ziyue Gan, Zhiyu Wang, Zijian Ling, Jinsong Chen, Martz Ma, Bowen Dong, et al. Oasis: Open agent social interaction simulations with one million agents. *arXiv preprint arXiv:2411.11581*, 2024.

[48] Yuhao Zheng, Chenghua Gong, Rui Sun, Juyuan Zhang, Liming Pan, and Linyuan Lv. Autocas: Autoregressive cascade predictor in social networks via large language models. *arXiv preprint arXiv:2502.18040*, 2025.

[49] Yifei Xu, Jiaying Wu, Herun Wan, Yang Li, Zhen Hou, and Min-Yen Kan. Forecasting the buzz: Enriching hashtag popularity prediction with llm reasoning. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, 2025.

[50] Yijie Xu, Bolun Zheng, Wei Zhu, Hangjia Pan, Yuchen Yao, Ning Xu, Anan Liu, Quan Zhang, and Chenggang Yan. Smtpd: A new benchmark for temporal prediction of social media popularity. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025.

[51] Bo Wu, Peiye Liu, Wen-Huang Cheng, Bei Liu, Zhaoyang Zeng, Jia Wang, Qiushi Huang, and Jiebo Luo. Smp challenge: An overview and analysis of social media prediction challenge. In *Proceedings of the 31st ACM International Conference on Multimedia*, 2023.

[52] Yaser Keneshloo, Shuguang Wang, Eui-Hong Han, and Naren Ramakrishnan. Predicting the popularity of news articles. In *Proceedings of the 2016 SIAM international conference on data mining*. SIAM, 2016.

[53] Ruocheng Guo and Paulo Shakarian. A comparison of methods for cascade prediction. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 2016.

[54] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 1998.

[55] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *science*, 2018.

[56] Sharad Goel, Ashton Anderson, Jake Hofman, and Duncan J Watts. The structural virality of online diffusion. *Management science*, 2016.

[57] Yafei Zhang, Lin Wang, Jonathan JH Zhu, and Xiaofan Wang. Conspiracy vs science: A large-scale analysis of online discussion cascades. *World wide web*, 2021.

[58] Meeyoung Cha, Alan Mislove, and Krishna P Gummadi. A measurement-driven analysis of information propagation in the flickr social network. In *Proceedings of the 18th international conference on World wide web*, 2009.

[59] Alexandru Tatar, Marcelo Dias De Amorim, Serge Fdida, and Panayotis Antoniadis. A survey on predicting the popularity of web content. *Journal of Internet Services and Applications*, 2014.

[60] Andrea Failla and Giulio Rossetti. "i'm in the bluesky tonight": insights from a year worth of social data. *PloS one*, 2024.

[61] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 830–839, 2020.

[62] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[63] Claudio Castellano, Santo Fortunato, and Vittorio Loreto. Statistical physics of social dynamics. *Reviews of modern physics*, 2009.

[64] Andreas Flache, Michael Mäs, Thomas Feliciani, Edmund Chattoe-Brown, Guillaume Deffuant, Sylvie Huet, and Jan Lorenz. Models of social influence: Towards the next frontiers. *Journal of Artificial Societies and Social Simulation*, 2017.

[65] Thomas C Schelling. Dynamic models of segregation. *Journal of mathematical sociology*, 1971.

[66] Mark Granovetter. Threshold models of collective behavior. *American journal of sociology*, 1978.

[67] Duncan J Watts. A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences*, 99, 2002.

[68] Morris H DeGroot. Reaching a consensus. *Journal of the American Statistical association*, 69(345):118–121, 1974.

[69] Noah E Friedkin and Eugene C Johnsen. Social influence and opinions. *Journal of mathematical sociology*, 15(3-4):193–206, 1990.

[70] Adrien Guille, Hakim Hacid, Cecile Favre, and Djamel A Zighed. Information diffusion in online social networks: A survey. *ACM Sigmod Record*, 42(2):17–28, 2013.

[71] Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*, 2024.

[72] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European conference on computer vision (ECCV)*, 2018.

[73] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 2017.

[74] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. Chapman and Hall/CRC, 1994.

[75] Oren Barkan, Edan Hauon, Avi Caciularu, Ori Katz, Itzik Malkiel, Omri Armstrong, and Noam Koenigstein. Grad-sam: Explaining transformers via gradient self-attention maps. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 2882–2887, 2021.

[76] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: visual explanations from deep networks via gradient-based localization. *International journal of computer vision*, 128, 2020.

[77] Jacob Gildenblat and contributors. Pytorch library for cam methods. `https://github.com/jacobgil/pytorch-grad-cam`, 2021.

Appendix

Table of Contents

A Dataset Details

Here, we present the detailed information about the curation of the datasets and their statistics.

A.1 Curation Details

A.1.1 Cascade construction.

For Bluesky, the metadata supports multiple interaction networks, including reply, repost, and quote networks. We use the reply network because replies capture explicit conversational interactions and provide richer content signals than diffusion-oriented actions such as reposts. Posts are first grouped by their discussion thread identifier, and parent references are then used to add reply edges within each thread. For Reddit, each submission defines a thread, and each comment provides a parent identifier indicating whether it replies to the root submission or to another comment.

A.1.2 Node attributes and thread-level context.

For both Bluesky and Reddit, each post is associated with textual content and timestamp information, which instantiate X_v^{Text} and X_v^{Time}, respectively. Platform-specific engagement signals are also available at the post level. Bluesky provides a like count for each post, whereas Reddit provides a karma score, computed as upvotes minus downvotes. These engagement signals are available for individual posts, but our Like Score popularity target is defined only on the root post v_{root} and is evaluated using its final engagement count. For Bluesky, this translates to the number of likes that the social cascade initiating post receives, while on Reddit, it means the Karma score, which is defined as upvotes minus downvotes for the initial submission post. Thread-level context X^{Thread} is derived from metadata associated with the root post, including its timestamp. For Bluesky, this additionally includes the follower count of the root post's author, which provides a proxy for the initiating user's social influence.

A.1.3 Modalities.

The r/Gaming and r/Futurology data subsets include image content for root posts, although not every root post contains an image. Specifically, 37.9% of r/Gaming cascades and 68.5% of r/Futurology cascades contain root-post images. In contrast, Bluesky and r/AMA are text-based in our benchmark, so X_v^{Image} is empty for these datasets. Since images are only available at the root-post level in the Reddit multimodal subsets, visual features are included as part of the corresponding thread-level context when present.

A.1.4 Missing and removed content.

We discard posts whose parent post is missing. If the root post of a cascade is missing but associated replies are present, we discard the entire cascade, since v_{root} is required to define the discussion tree. For Reddit, some posts or comments may no longer contain their original textual content because they were deleted by users or removed by moderators at the time of collection and are instead represented by markers such as `[deleted]` or `[removed]`. We retain such cascades when the remaining thread structure and metadata are intact, since these posts still correspond to observed participation in the cascade. Although the original text is no longer accessible, the presence of deletion or removal markers may still provide information to the model about moderation or deletion patterns in the dataset communities. Dataset sizes are computed after these filtering steps.

A.1.5 Dataset Sampling and Imbalance Mitigation.

In the raw Bluesky dataset, cascade sizes are heavily skewed, with small conversation trees containing 3 to 10 posts comprising approximately 87% of the filtered data. Training directly on this distribution would bias the model toward shallow dynamics and obscure structural patterns present in more complex cascades. To address this imbalance, we employ square-root sampling [72]. Specifically, each size bin is sampled proportionally to the square root of its empirical frequency, yielding a more balanced subset of roughly 64,000 Bluesky discussion trees for robust model training and evaluation.

Table 4: Dataset statistics across observation windows. Final columns describe complete cascades retained for each window, while early columns describe the corresponding observed prefixes G^t.

Dataset	Window	Cascades	Final			Early			Early % of Final
			Nodes	Avg.	Med.	Nodes	Avg.	Med.	
Bluesky	2min	63,904	1,198,634	18.76	7.0	107,920	1.62	1.0	9.0%
	10min	60,321	1,183,956	19.63	7.0	218,619	2.93	2.0	18.4%
	20min	52,972	1,144,617	21.61	8.0	305,955	5.76	3.0	26.7%
r/AMA	15min	38,982	1,500,181	38.5	15.0	156,139	4.0	3.0	10.4%
	30min	37,682	1,491,365	39.6	16.0	259,185	6.9	5.0	17.4%
	60min	35,677	1,471,684	41.3	17.0	392,444	11.0	7.0	26.7%
r/Gaming	20min	13,398	1,739,720	129.8	22.0	70,939	5.3	4.0	4.1%
	50min	12,882	1,735,771	134.7	24.0	143,475	11.1	7.0	8.3%
	90min	12,413	1,730,833	139.4	25.0	237,263	19.1	9.0	13.7%
r/Futurology	30min	9,780	1,220,687	124.8	14.0	35,077	3.6	2.0	2.9%
	90min	8,765	1,215,134	138.6	18.0	95,953	10.9	4.0	7.9%
	180min	8,349	1,211,576	145.1	19.0	197,398	23.6	7.0	16.3%

A.2 Dataset Statistics.

Table 4 summarizes the dataset statistics for each benchmark subset after preprocessing and observation-window filtering. Each row corresponds to one dataset under one early observation window. The observation window specifies how much of each cascade is visible to the model before prediction. For example, a 2min window for Bluesky means that only the first 2 minutes of each cascade are observed, while a 15min window for r/AMA means that only the first 15 minutes are observed.

The **Dataset** column indicates the benchmark subset, including Bluesky and the Reddit communities r/AMA, r/Gaming, and r/Futurology. The **Window** column gives the length of the early observation period. The **Cascades** column reports the number of cascades retained for that dataset and window. Since cascades that have already completed before a given observation window are excluded, larger windows may retain fewer cascades. Therefore, statistics are reported separately for each observation window.

The **Final** columns describe the complete cascades after they have fully unfolded, up to the date of data collection. These values represent the final cascade states used as prediction targets. Under **Final**, **Nodes** reports the total number of nodes across all retained complete cascades, **Avg.** reports the average number of final nodes per cascade, and **Med.** reports the median number of final nodes per cascade.

The **Early** columns describe the observed cascade prefixes within the specified observation window. These values represent the information available to the model at prediction time. Under **Early**, **Nodes** reports the total number of nodes observed during the early window across all retained cascades, **Avg.** reports the average number of observed nodes per cascade, and **Med.** reports the median number of observed nodes per cascade.

Finally, the **Early % of Final** column reports the fraction of the complete cascade that is visible within the observation window. It is computed by comparing the total number of early observed nodes with the total number of final nodes for the same retained cascade set. Larger values indicate that a greater portion of the final cascade is available to the model before prediction. Overall, the table shows that longer observation windows provide more early cascade information, while often reducing the number of retained cascades because completed cascades are filtered out.

B Baseline Implementation Details

B.1 MLP.

MLP is a structure-agnostic baseline that represents each social cascade as a fixed vector rather than a reply tree. For each observed cascade prefix, the model builds a cascade-level representation by concatenating the initial root post's feature vector, the mean feature vector over all observed posts, and the thread-level context vector. Each post feature vector contains a projected text embedding and two temporal features: time since the root post and time since the parent post. The thread-level context contains thread-level metadata, including posting-time and follower count. The model passes the resulting cascade representation through a shared multilayer perceptron where it predicts all 6 popularity targets. All targets are predicted in log-transformed space.

Hyperparameter Settings: The precomputed text embedding dimension is 384, and the projected text embedding dimension is 32. The temporal features are log-transformed and standardized using training-set statistics. The MLP has 2 layers of hidden dimensions 128, dropout is 0.3, and the output dimension is 6, corresponding to the six popularity targets. Training uses Adam with learning rate 10^{-3}, batch size 256, a maximum of 200 epochs and with early stopping patience 10.

B.2 DeepHawkes.

We implemented a DeepHawkes-style baseline [29] for cascade prediction. DeepHawkes represents an information cascade as a set of diffusion paths and learns path-level representations that capture user influence, self-excitation, and temporal decay in an end-to-end neural architecture. In our setting, each MMG-Pop training instance corresponds to one social cascade, represented by the observed root-to-node diffusion paths within the observation window.

Following the DeepHawkes formulation, each path is encoded as a sequence of user embeddings and summarized with a recurrent encoder. To incorporate temporal effects, we assign each path to one of 10 recency bins according to the time of its final post relative to the end of the observation window. The model learns a positive scalar weight for each recency bin, and the cascade representation is obtained by a weighted aggregation of path representations. This preserves the main DeepHawkes design while adapting it to the MMG-Pop cascade format.

Hyperparameter Settings: We use 50-dimensional user embeddings and a GRU hidden size of 32. The prediction MLP has hidden dimensions 32 and 16 with ReLU activations and dropout rate 0.5. The output dimension is 6, corresponding to the six prediction targets. Models are trained with Adam using batch size 32, up to 200 epochs, early stopping patience 10, and weight decay 10^{-4}. User embeddings use learning rate 5×10^{-4}, while the remaining parameters use learning rate 5×10^{-3}.

B.3 DeepCas.

We implement DeepCas [30] as a path-based neural cascade encoder. Each cascade is represented by random walks sampled from the observed early-window conversation tree, using only nodes and edges available within the observation window. Each walk starts at the initial post of the social cascade. During sampling, the walker either moves to a child node or jumps to another observed node and where, both child transitions and jump targets are sampled using degree-based weights. If the current node is a leaf, the walker performs a jump using the same weighting rule. The sampled user sequences are mapped to a pretrained user-embedding vocabulary. The embeddings are trained separately on a train-split global interaction graph constructed from reply links across all cascades in the training set. In this graph, each directed weighted edge connects a replying user to the author being replied to, and repeated reply interactions are aggregated as edge weights. This graph is used to capture user-user interactions across the dataset, rather than within a single cascade. The pretrained embeddings are then loaded into DeepCas and kept fixed during supervised training. Here, the model is adjusted to predict all 6 popularity targets as a 6 dimension vector.

Hyperparameter Settings: We use K = 200 sampled walks, T = 10 steps per walk, pretrained interaction-graph embeddings with dimension 128, GRU hidden dimension 128 per direction, random-walk group size 5, two MLP hidden layers of size 128 with ReLU and dropout 0.4, and a 6-dimensional output layer. Training uses Adam over trainable parameters only, learning rate 10^{-3}, weight decay

5×10^{-4} batch size 256, maximum 200 epochs, early-stopping patience 10, and gradient clipping with maximum norm 1.0.

B.4 CasSeqGCN.

CasSeqGCN [31] represents each cascade as a temporal sequence of graph snapshots constructed from the observed early-window conversation tree. The original model combines structural and temporal cascade information by first encoding each snapshot with graph convolution, then aggregating node embeddings into a snapshot representation with dynamic routing, and finally processing the snapshot sequence with an LSTM before prediction. Given this ordered sequence, snapshots are formed by adding posts in fixed increments of $Q = 5$. We keep at most $K_{\max} = 15$ snapshots and force the final snapshot to include the last observed post if it is not already included by the fixed stride. Thus, each cascade is represented as a sequence of up to 15 partial graph snapshots. Here, the model is adjusted to predict all 6 popularity targets as a 6 dimension vector.

Hyperparameter Settings: Node embedding dimension is 32, the snapshot embedding dimension is 32, the GCN hidden dimension is 32, the number of GCN layers is 2, the LSTM hidden dimension is 32, the number of LSTM layers is 2, the number of dynamic routing iterations is 3, and dropout is 0.5. For model selection, we search over learning rates $\{0.005, 0.01, 0.03, 0.05\}$ separately for each dataset/window. Each candidate is trained for at most 20 epochs with patience 5, and the candidate with the lowest validation loss is selected for full training. Full training uses Adam with the selected learning rate, weight decay 5×10^{-5} batch size 256, at most 200 epochs, early-stopping patience 10, gradient clipping with maximum norm 1.0.

B.5 GraphLSTM.

GraphLSTM [33] represents each cascade as a conversation tree. Each post is a node, and each reply forms an edge from the parent post to the reply post. For every node, the model builds an input vector by concatenating structural features with a mean-pooled text embedding. The structural features describe the node's timing and position in the observed tree. The text embedding summarizes the post text using learned token embeddings. The model applies two graph-LSTM passes over the tree. The forward pass uses information from the parent node and the previous sibling node. This gives each node a representation of the conversation context that came before it. The backward pass uses information from the first child node and the next sibling node. This gives each node a representation of the response context that follows it. The forward and backward states are then concatenated to obtain a context-aware representation for each node. To obtain a cascade-level representation, we mean-pool the concatenated node states over all nodes in the tree. Here, the model is adjusted to predict all 6 popularity targets as a 6 dimension vector.

Hyperparameter Settings: The token embedding dimension is 100, the graph-LSTM hidden dimension is 128, the final MLP hidden dimension is 128, dropout is 0.3, the maximum post length is 100 tokens, the vocabulary minimum frequency is 10. Training uses AdamW with learning rate 10^{-3}, embedding learning rate 10^{-4}, weight decay 10^{-5}, batch size 256, a maximum of 200 epochs, early stopping patience 10, and gradient clipping with maximum norm 1.0.

C Experimental Setup

C.1 MMG-PopNet Settings.

MMG-PopNet consists of three modality-specific components and a prediction head. We use `all-MiniLM-L6-v2` as the text encoder, `CLIP ViT-B/32` as the vision encoder, and GraphSAGE [73] as the graph message-passing backbone.

Hyperparameter Settings: The text encoder `all-MiniLM-L6-v2` produces a 384-dimensional representation. This representation is mapped to a 32-dimensional text embedding using a two-layer projection network with hidden dimension 64 and dropout rate 0.15. All layers of `all-MiniLM-L6-v2` are updated during training. The vision encoder `CLIP ViT-B/32` produces a 512-dimensional representation. This representation is mapped to a 64-dimensional image embedding using a linear projection layer with dropout rate 0.3. During training, only the final layer of the CLIP vision encoder is updated. The graph component uses a 3-layer GraphSAGE network with mean aggregation, hidden

dimension 128, and dropout rate 0.3. The fused representation is passed to a two-layer MLP prediction head with hidden dimension 128. The prediction head outputs a 6-dimensional vector, with one output corresponding to each popularity target. We train MMG-PopNet using Adam with separate learning rates for different parameter groups. The learning rate is 5×10^{-6} for the text encoder, 10^{-6} for the trainable CLIP vision parameters, 10^{-3} for the image projection parameters, and 10^{-3} for the GNN parameters. The image projection parameters use weight decay 10^{-2}. Training uses variable batch sizes with gradient accumulation to obtain an effective batch size of 256. The maximum number of training epochs is 100, and early stopping is applied with patience 10. We use mixed-precision training and clip gradients to a maximum norm of 0.5. The learning-rate schedule consists of linear warmup followed by cosine decay, with warmup ratio 0.05 and minimum learning-rate factor 0.0.

C.2 Completed Cascade Exclusion.

Cascades that complete before the observation window are excluded from that setting to avoid trivial prediction cases. For example, a cascade that ends after 8 minutes is excluded from the 10-minute window because its observed prefix would already equal the final cascade.

C.3 Training Split and Compute Resources.

Dataset was divided into 80/10/10 splits of train, validation and test respectively. For computation, 4 $\times$ Nvidia L40s GPUs were used to train the MMG-PopNet to allow for handling the Out-of-Memory issue due to large text content associated with long social cascades. To ensure fair comparison, all models had an effective batch size of 256. Other representative baselines were trained on 1 L40s GPU.

D Popularity Prediction Across Future Horizons

D.1 Popularity prediction of final cascade state under different early observation.

In addition to the MSE results reported in Table 2, we report R^2 and Spearman rank correlation results for the same final cascade state prediction setting. These two metrics offer complementary perspectives on model performance, where R^2 evaluates the exact predictive fit of the model's estimates against true values, while Spearman correlation assesses the rank-order agreement between predicted and actual outcomes.

The R^2 metric (coefficient of determination) measures the proportion of variance in the final cascade states that can be explained by the early observation signals. A higher R^2 score indicates that a model's numerical predictions tightly fit the actual popularity distributions. As shown in Table 5, MMG-PopNet achieves the strongest overall performance, obtaining the highest average score for every target metric. The gains are consistent across the Bluesky and Reddit datasets, indicating that the model successfully explains social cascade variance across platforms with diverse dynamics. Furthermore, the advantage is maintained across all early observation windows. In contrast, models relying on narrower temporal or structural signals, such as DeepHawkes and DeepCas, show unstable results with R^2 values frequently approaching zero or dropping negative, meaning they fail to capture the variance better than simply predicting the mean.

Conversely, the Spearman rank correlation evaluates how well a model preserves the relative ordering of cascades, independent of absolute numerical errors. This is particularly important for downstream applications where the goal is to identify which threads will become the most viral or structurally complex. Table 6 demonstrates that MMG-PopNet consistently achieves the highest overall average for every prediction task. While baselines like Graph-LSTM and MLP perform reasonably well at ranking tasks compared to their R^2 fit, they still fall short of MMG-PopNet. The strong Spearman results confirm that MMG-PopNet's multimodal design not only minimizes numerical error but reliably orders final cascade outcomes, making it highly effective for trend identification across heterogeneous platforms.

Table 5: R^2 results for final cascade state prediction under different early observation windows, including **STRUCTURAL TASKS** (max width, max depth, structural virality, size), **UNIQUE USERS**, and **LIKE SCORE**. Higher is better. Best values are **bolded**, and second-best values are underlined.

Task	Model	Bluesky					r/AMA					r/Gaming					r/Futurology					Avg
		0	2	10	20	Avg	0	15	30	60	Avg	0	20	50	90	Avg	0	30	90	180	Avg	
MAX WIDTH	MLP	0.322	0.368	0.452	0.482	0.406	0.069	0.259	0.341	0.398	0.267	0.058	0.366	0.530	0.617	0.393	-0.005	0.255	0.570	0.694	0.378	0.361
	DeepHawkes	-0.012	0.087	0.404	0.500	0.245	-0.001	0.223	0.344	0.470	0.259	-0.013	0.320	0.385	0.547	0.310	-0.042	0.087	0.269	0.475	0.197	0.253
	DeepCas	0.063	0.002	0.000	0.000	0.016	-0.180	-0.062	-0.025	-0.030	-0.074	-0.267	-0.017	0.018	0.087	-0.045	0.035	0.087	0.265	0.262	0.162	0.015
	CasSeqGCN	0.000	0.204	0.473	0.577	0.314	0.000	0.250	0.404	0.545	0.300	0.000	0.388	0.549	0.683	0.405	-0.011	0.301	0.581	0.755	0.406	0.356
	Graph-LSTM	0.029	0.222	0.490	0.584	0.331	0.038	0.285	0.425	0.557	0.326	0.050	0.380	0.565	**0.720**	0.429	0.028	0.361	**0.611**	**0.769**	0.442	0.382
	MMG-PopNet (Ours)	**0.325**	**0.454**	**0.580**	**0.654**	**0.503**	**0.124**	**0.312**	**0.427**	**0.598**	**0.365**	**0.164**	**0.412**	**0.618**	0.696	**0.472**	**0.121**	**0.371**	0.590	0.751	**0.458**	**0.450**
MAX DEPTH	MLP	**0.022**	0.049	0.188	0.258	0.129	0.031	0.186	0.299	0.386	0.226	0.008	0.194	0.387	0.444	0.258	-0.017	0.170	0.434	0.537	0.281	0.224
	DeepHawkes	-0.008	0.015	0.055	0.177	0.060	-0.003	0.146	0.214	0.256	0.153	0.002	0.163	0.228	0.315	0.177	-0.034	0.027	0.159	0.319	0.118	0.127
	DeepCas	0.009	0.000	-0.000	0.001	0.002	-0.316	-0.143	-0.061	-0.032	-0.138	-0.664	-0.180	-0.109	-0.017	-0.243	-0.023	0.050	0.197	0.215	0.110	-0.067
	CasSeqGCN	0.000	0.045	0.193	0.293	0.133	0.000	0.143	0.295	0.402	0.210	-0.001	0.146	0.294	0.373	0.203	-0.006	0.182	0.380	0.517	0.268	0.203
	Graph-LSTM	0.001	0.067	0.217	0.333	0.155	0.015	0.194	**0.330**	0.455	0.248	-0.020	0.168	0.401	0.518	0.267	0.021	0.223	0.417	0.514	0.294	0.241
	MMG-PopNet (Ours)	0.003	**0.106**	**0.251**	**0.345**	**0.176**	**0.042**	**0.219**	0.318	**0.466**	**0.261**	**0.033**	**0.205**	**0.427**	**0.523**	**0.297**	**0.112**	**0.282**	**0.447**	**0.571**	**0.353**	**0.272**
STRUCTURAL VIRALITY	MLP	**0.063**	0.092	0.215	0.274	0.161	**0.037**	0.211	0.328	0.430	0.252	-0.007	**0.184**	0.370	0.468	0.254	-0.070	0.160	0.430	0.539	0.265	0.233
	DeepHawkes	-0.007	0.021	0.087	0.241	0.085	-0.007	0.170	0.236	0.283	0.170	-0.001	0.111	0.112	0.252	0.118	-0.027	0.002	0.126	0.244	0.086	0.115
	DeepCas	0.016	-0.000	-0.000	0.001	0.004	-0.632	-0.257	-0.099	-0.078	-0.267	-1.780	-0.613	-0.490	-0.260	-0.786	-0.091	-0.025	0.079	0.139	0.026	-0.256
	CasSeqGCN	0.000	0.056	0.214	0.309	0.145	-0.001	0.170	0.329	0.461	0.240	**0.000**	0.098	0.211	0.288	0.149	-0.004	0.171	0.350	0.515	0.258	0.198
	Graph-LSTM	0.001	0.075	0.224	0.340	0.160	0.016	0.209	**0.359**	0.505	0.272	-0.079	0.127	0.361	**0.521**	0.232	0.001	0.195	0.407	0.498	0.275	0.235
	MMG-PopNet (Ours)	0.015	**0.142**	**0.273**	**0.360**	**0.198**	0.023	**0.251**	0.343	**0.508**	**0.281**	-0.025	0.151	**0.389**	0.513	**0.257**	**0.067**	**0.251**	**0.437**	**0.564**	**0.330**	**0.266**
SIZE	MLP	**0.211**	0.260	0.369	0.415	0.314	0.063	0.274	0.380	0.455	0.293	0.039	0.334	0.529	0.616	0.380	-0.009	0.231	0.560	0.717	0.375	0.340
	DeepHawkes	-0.021	0.072	0.298	0.433	0.196	-0.003	0.247	0.371	0.498	0.278	-0.018	0.299	0.377	0.550	0.302	-0.066	0.061	0.214	0.486	0.174	0.237
	DeepCas	0.055	0.002	0.000	0.001	0.015	-0.288	-0.116	-0.044	-0.053	-0.125	-0.406	-0.061	-0.021	0.081	-0.102	0.018	0.076	0.270	0.288	0.163	-0.012
	CasSeqGCN	0.000	0.147	0.374	0.478	0.250	0.000	0.250	0.407	0.558	0.304	-0.001	0.333	0.516	0.657	0.376	-0.013	0.260	0.536	0.742	0.381	0.328
	Graph-LSTM	0.024	0.168	0.397	0.494	0.271	0.035	0.285	**0.433**	0.575	0.332	0.033	0.325	0.531	**0.693**	0.395	0.031	0.320	0.566	0.750	0.417	0.354
	MMG-PopNet (Ours)	0.199	**0.334**	**0.466**	**0.549**	**0.387**	**0.118**	**0.325**	**0.433**	**0.615**	**0.373**	**0.129**	**0.373**	**0.601**	0.687	**0.448**	**0.135**	**0.360**	**0.588**	**0.752**	**0.459**	**0.416**
UNIQUE USERS	MLP	**0.357**	0.402	0.482	0.515	0.439	0.070	0.249	0.327	0.389	0.259	0.048	0.340	0.523	0.617	0.382	-0.003	0.230	0.551	0.710	0.372	0.363
	DeepHawkes	-0.016	0.087	0.392	0.487	0.237	-0.001	0.209	0.322	0.459	0.247	-0.015	0.296	0.368	0.536	0.296	-0.044	0.064	0.238	0.492	0.188	0.242
	DeepCas	0.078	0.002	0.000	0.001	0.020	-0.231	-0.086	-0.034	-0.048	-0.100	-0.329	-0.021	-0.006	0.090	-0.067	0.028	0.074	0.268	0.280	0.163	0.004
	CasSeqGCN	0.000	0.191	0.454	0.559	0.301	0.000	0.231	0.373	0.521	0.281	0.000	0.346	0.515	0.660	0.380	-0.011	0.258	0.538	0.745	0.383	0.336
	Graph-LSTM	0.031	0.225	0.491	0.590	0.334	0.039	0.272	0.399	0.532	0.310	0.040	0.339	0.534	**0.697**	0.402	0.028	0.325	0.568	0.741	0.415	0.366
	MMG-PopNet (Ours)	0.354	**0.470**	**0.582**	**0.648**	**0.513**	**0.119**	**0.300**	**0.400**	**0.576**	**0.349**	**0.144**	**0.384**	**0.602**	0.690	**0.455**	**0.129**	**0.356**	**0.584**	**0.751**	**0.455**	**0.443**
LIKE SCORE	MLP	0.477	0.482	0.500	0.508	0.492	0.059	0.105	0.134	0.165	0.116	0.092	0.099	0.204	0.306	0.175	0.041	0.094	0.332	0.406	0.218	0.250
	DeepHawkes	-0.010	0.043	0.198	0.260	0.123	-0.002	0.017	0.057	0.073	0.036	-0.006	0.013	0.048	0.087	0.035	-0.022	-0.021	0.006	0.158	0.030	0.056
	DeepCas	0.060	0.002	0.000	0.001	0.016	-0.006	0.005	0.006	-0.005	0.000	-0.025	0.017	0.036	0.066	0.024	0.150	0.141	0.247	0.222	0.190	0.057
	CasSeqGCN	0.000	0.096	0.242	0.303	0.160	-0.001	0.020	0.061	0.085	0.041	0.000	0.013	0.036	0.106	0.039	-0.006	0.020	0.107	0.206	0.082	0.081
	Graph-LSTM	0.016	0.128	0.286	0.359	0.197	0.052	0.087	0.104	0.134	0.094	0.080	0.101	0.155	0.275	0.153	0.074	0.199	0.286	0.303	0.215	0.165
	MMG-PopNet (Ours)	**0.497**	**0.566**	**0.591**	**0.621**	**0.569**	**0.087**	**0.156**	**0.218**	**0.281**	**0.185**	**0.183**	**0.250**	**0.313**	**0.393**	**0.285**	**0.258**	**0.318**	**0.471**	**0.564**	**0.403**	**0.360**

Description: The R^2 metric indicates the proportion of variance in the final popularity outcomes that the models successfully explain. MMG-PopNet consistently outperforms all baselines, achieving the highest average scores across structural, user, and engagement prediction targets. Notably, traditional baselines like DeepCas and DeepHawkes struggle significantly, often yielding near-zero or negative values, while MMG-PopNet maintains a robust predictive fit regardless of the observation horizon.

D.2 Popularity Prediction of Cascade States at Future Horizon

We provide the full future-horizon prediction results in Tables 7, 8, 9, and 10. These tables extend the main results by reporting all target variables across all datasets, observation windows, and prediction horizons. In addition to the intermediate horizons $\{4\text{h}, 8\text{h}, 16\text{h}, 24\text{h}\}$, we also report prediction error for the final cascade state. Thus, in this setting, each observed prefix G^t can produce multiple supervised instances, where each instance pairs the same prefix with a different target state $\mathbf{Y}_G^{t'}$. To specify which future state is being predicted, we append a one-hot encoding of the target horizon to the input representation. This training setup differs from terminal-state-only prediction because each model receives supervision from both intermediate cascade states and the final state. The learned representation is therefore shaped by multiple stages of cascade evolution rather than only by the terminal outcome.

If a cascade has already terminated before a given horizon, the corresponding intermediate target is unavailable and is excluded from that horizon-specific training and evaluation set. The final-state target differs from the fixed intermediate horizons because it is defined for every completed cascade. For this reason, the final-state MSE can be lower than the error at longer horizons such as 16h or 24h. Some cascades terminate before those later horizons, so their final state occurs earlier than the fixed horizon and is easier to infer from the observed prefix. This effect should be considered when comparing the final-state column with the 16h and 24h columns.

Table 6: Spearman results for final cascade state prediction under different early observation windows, including **STRUCTURAL TASKS** (max width, max depth, structural virality, size), **UNIQUE USERS**, and **LIKE SCORE**. Higher is better. Best values are **bolded**, and second-best values are underlined.

Task	Model	Bluesky					r/AMA					r/Gaming					r/Futurology					Avg
		0	2	10	20	Avg	0	15	30	60	Avg	0	20	50	90	Avg	0	30	90	180	Avg	
MAX WIDTH	MLP	0.523	0.538	0.585	0.610	0.564	0.274	0.517	0.611	0.651	0.513	0.244	0.597	0.710	0.770	0.580	0.153	0.519	0.729	0.798	0.550	0.552
	DeepHawkes	0.030	0.238	0.524	0.553	0.336	-0.006	0.507	0.636	0.728	0.466	0.004	0.552	0.690	0.773	0.505	-0.035	0.409	0.615	0.731	0.430	0.434
	DeepCas	0.083	0.105	0.087	0.084	0.090	0.023	0.126	0.140	0.140	0.107	0.158	0.244	0.246	0.306	0.238	0.220	0.359	0.523	0.569	0.418	0.213
	CasSeqGCN	0.000	0.319	0.546	0.653	0.380	0.000	0.515	0.670	0.760	0.486	0.000	0.609	0.749	0.819	0.544	0.000	0.573	0.755	**0.865**	0.548	0.490
	Graph-LSTM	0.163	0.335	0.553	0.646	0.424	0.197	0.536	0.677	0.771	0.545	0.221	0.605	0.781	**0.846**	0.613	0.192	0.598	**0.774**	0.854	0.605	0.547
	MMG-PopNet (Ours)	**0.534**	**0.589**	**0.667**	**0.714**	**0.626**	**0.366**	**0.578**	**0.681**	**0.779**	**0.601**	**0.430**	**0.652**	**0.783**	0.838	**0.676**	**0.410**	**0.634**	0.771	0.855	**0.667**	**0.643**
MAX DEPTH	MLP	0.130	0.180	0.367	0.467	0.286	0.172	0.428	0.553	0.622	0.444	0.121	0.449	0.623	0.662	0.464	0.122	0.413	0.648	0.718	0.475	0.417
	DeepHawkes	0.007	0.106	0.217	0.360	0.172	-0.025	0.398	0.503	0.556	0.358	0.012	0.418	0.582	0.637	0.412	-0.057	0.330	0.530	0.688	0.373	0.329
	DeepCas	0.038	0.037	0.044	0.059	0.044	0.057	0.126	0.135	0.139	0.114	0.081	0.161	0.164	0.230	0.159	0.196	0.304	0.464	0.491	0.364	0.170
	CasSeqGCN	0.000	0.176	0.353	0.492	0.255	0.000	0.388	0.542	0.626	0.389	0.000	0.386	0.556	0.611	0.388	0.000	0.402	0.613	0.705	0.430	0.366
	Graph-LSTM	0.114	0.207	0.423	0.531	0.319	0.134	0.425	0.571	0.670	0.450	0.115	0.417	0.637	0.719	0.472	0.164	0.469	0.656	0.706	0.499	0.435
	MMG-PopNet (Ours)	**0.194**	**0.279**	**0.452**	**0.546**	**0.368**	**0.255**	**0.469**	**0.583**	**0.681**	**0.497**	**0.257**	**0.473**	**0.667**	**0.724**	**0.530**	**0.395**	**0.543**	**0.663**	**0.745**	**0.587**	**0.495**
STRUCTURAL VIRALITY	MLP	0.278	0.314	0.449	0.507	0.387	0.190	0.465	0.592	0.666	0.478	0.131	0.440	0.616	0.675	0.465	0.128	0.422	0.664	0.737	0.488	0.455
	DeepHawkes	0.013	0.163	0.346	0.478	0.250	-0.026	0.439	0.556	0.621	0.398	0.001	0.395	0.550	0.590	0.384	-0.051	0.346	0.535	0.683	0.378	0.352
	DeepCas	0.049	0.060	0.062	0.076	0.062	0.052	0.128	0.147	0.137	0.116	0.081	0.154	0.151	0.215	0.150	0.207	0.308	0.459	0.486	0.365	0.173
	CasSeqGCN	0.000	0.225	0.413	0.533	0.293	0.000	0.431	0.590	0.682	0.426	0.000	0.359	0.520	0.569	0.362	0.000	0.424	0.629	0.729	0.446	0.382
	Graph-LSTM	0.128	0.236	0.445	0.539	0.337	0.149	0.456	0.609	0.712	0.481	0.088	0.389	0.626	0.726	0.457	0.178	0.482	0.674	0.715	0.512	0.447
	MMG-PopNet (Ours)	**0.303**	**0.372**	**0.496**	**0.574**	**0.436**	**0.276**	**0.508**	**0.618**	**0.721**	**0.531**	**0.239**	**0.476**	**0.668**	**0.733**	**0.529**	**0.405**	**0.538**	**0.684**	**0.761**	**0.597**	**0.523**
SIZE	MLP	0.393	0.421	0.518	0.566	0.475	0.253	0.527	0.641	0.698	0.530	0.207	0.584	0.721	0.775	0.572	0.137	0.495	0.732	0.814	0.544	0.530
	DeepHawkes	0.024	0.213	0.454	0.550	0.310	-0.013	0.522	0.658	0.758	0.481	0.009	0.550	0.712	0.788	0.515	-0.042	0.394	0.619	0.760	0.433	0.435
	DeepCas	0.071	0.084	0.079	0.088	0.080	0.041	0.141	0.157	0.151	0.122	0.151	0.237	0.232	0.307	0.232	0.222	0.344	0.524	0.573	0.416	0.213
	CasSeqGCN	0.000	0.280	0.477	0.588	0.336	0.000	0.512	0.663	0.763	0.485	0.000	0.575	0.741	0.805	0.530	0.000	0.525	0.731	**0.847**	0.526	0.469
	Graph-LSTM	0.146	0.297	0.500	0.590	0.383	0.197	0.534	0.676	0.778	0.546	0.197	0.567	0.768	**0.833**	0.591	0.190	0.566	0.752	0.832	0.585	0.526
	MMG-PopNet (Ours)	**0.415**	**0.473**	**0.575**	**0.638**	**0.525**	**0.359**	**0.581**	**0.681**	**0.788**	**0.602**	**0.401**	**0.624**	**0.779**	**0.833**	**0.659**	**0.418**	**0.618**	**0.762**	0.845	**0.661**	**0.612**
UNIQUE USERS	MLP	0.562	0.574	0.615	0.638	0.597	0.278	0.515	0.611	0.652	0.514	0.229	0.586	0.712	0.774	0.575	0.156	0.495	0.718	0.805	0.543	0.557
	DeepHawkes	0.042	0.222	0.499	0.517	0.320	-0.009	0.510	0.635	0.728	0.466	0.006	0.546	0.692	0.773	0.504	-0.033	0.390	0.602	0.751	0.427	0.429
	DeepCas	0.075	0.104	0.085	0.084	0.087	0.028	0.132	0.145	0.141	0.111	0.169	0.258	0.253	0.326	0.252	0.218	0.342	0.524	0.576	0.415	0.216
	CasSeqGCN	0.000	0.289	0.514	0.612	0.354	0.000	0.512	0.659	0.750	0.480	0.000	0.589	0.741	0.809	0.535	0.000	0.524	0.722	**0.850**	0.524	0.473
	Graph-LSTM	0.172	0.344	0.560	0.646	0.430	0.199	0.530	0.668	0.762	0.540	0.209	0.584	0.774	**0.840**	0.602	0.189	0.566	0.751	0.824	0.582	0.539
	MMG-PopNet (Ours)	**0.569**	**0.616**	**0.675**	**0.707**	**0.642**	**0.370**	**0.574**	**0.669**	**0.767**	**0.595**	**0.422**	**0.641**	**0.780**	0.836	**0.670**	**0.422**	**0.619**	**0.758**	0.843	**0.660**	**0.642**
LIKE SCORE	MLP	0.680	0.682	0.690	0.694	0.686	0.232	0.266	0.262	0.270	0.258	0.249	0.258	0.366	0.444	0.329	0.219	0.298	0.524	0.548	0.397	0.418
	DeepHawkes	0.057	0.165	0.376	0.372	0.242	-0.014	0.050	0.095	0.110	0.060	-0.020	0.010	0.112	0.184	0.072	-0.002	0.151	0.213	0.317	0.170	0.136
	DeepCas	0.091	0.112	0.093	0.089	0.096	0.039	0.068	0.133	0.101	0.085	0.086	0.156	0.186	0.255	0.171	0.411	0.418	0.459	0.456	0.436	0.197
	CasSeqGCN	0.000	0.205	0.382	0.451	0.260	0.000	0.038	0.083	0.103	0.056	0.000	0.015	0.088	0.177	0.070	0.000	0.140	0.258	0.369	0.192	0.144
	Graph-LSTM	0.127	0.275	0.441	0.520	0.341	0.189	0.207	0.208	0.246	0.212	0.298	0.292	0.343	0.433	0.342	0.280	0.462	0.543	0.445	0.433	0.332
	MMG-PopNet (Ours)	**0.711**	**0.750**	**0.765**	**0.776**	**0.750**	**0.251**	**0.308**	**0.347**	**0.356**	**0.316**	**0.397**	**0.432**	**0.489**	**0.525**	**0.461**	**0.532**	**0.575**	**0.689**	**0.726**	**0.630**	**0.539**

Description: The Spearman rank correlation evaluates the models' ability to accurately preserve the relative ordinal ranking of cascades based on their final states. MMG-PopNet achieves the highest average correlation across all tasks and platforms, demonstrating superior capability in ranking future popularity trends. While sequence-based models like Graph-LSTM provide competitive baseline rankings, MMG-PopNet's multimodal approach captures complementary signals that result in more reliable and consistent cascade orderings.

For engagement target LIKE SCORE, ground truth is available only at the final cascade state. Consequently, those targets are evaluated only in the Final column and are marked as unavailable at intermediate horizons.

Across the four datasets, MMG-PopNet achieves the lowest MSE in most settings across the popularity targets. It consistently outperforms the structure-only baselines such as DeepCas and DeepHawkes, and this underscores the necessity of multimodal representations for capturing both the semantic and structural dynamics that influence a social cascade. A notable pattern is that MLP is often competitive for intermediate horizons and sometimes obtains the best result. This is particularly visible under root-only inputs and shorter forecasting horizons, where a fixed cascade-level representation can capture much of the available predictive signal.

The MLP baseline uses the root-post features, the mean representation of observed posts, and thread-level metadata, so its strong performance indicates that early content, aggregate response features, posting context, and author-level context are highly informative for near-term cascade growth. This result also shows that complex structural encoders are not always necessary when the target horizon is close to the observation window or when little cascade topology is available.

Graph-LSTM and CasSeqGCN form the strongest structure-aware baselines in many settings. They are especially competitive when longer observation windows expose more of the reply tree, and their strongest results occur for MAX DEPTH and STRUCTURAL VIRALITY.

Overall, the appendix results support the trends reported in the main section while adding a more detailed view across targets.

Table 7: Bluesky future-horizon prediction MSE across observation windows and popularity targets. **Lower is better**. Best values are **bolded** and second-best values are underlined. Column groups correspond to Root Only, 2 min, 10 min, and 20 min observation windows, and columns within each group report prediction error at $\{4\text{h}, 8\text{h}, 16\text{h}, 24\text{h}\}$ and at the final cascade state. LIKE SCORE is reported only for the final state because its intermediate-horizon labels are unavailable.

Task	Model	Root Only					2 min					10 min					20 min				
		4h	8h	16h	24h	Final	4h	8h	16h	24h	Final	4h	8h	16h	24h	Final	4h	8h	16h	24h	Final
MAX WIDTH	MLP	**0.472**	0.570	0.694	0.800	0.458	0.410	0.499	0.620	0.721	0.429	0.318	0.398	0.497	0.584	0.373	0.291	0.364	0.454	0.532	0.348
	DeepHawkes	0.809	0.999	1.262	1.444	0.728	0.683	0.856	1.087	1.270	0.634	0.369	0.494	0.661	0.786	0.436	0.261	0.347	0.462	0.549	0.330
	DeepCas	0.707	0.864	1.051	1.198	0.628	0.787	0.968	1.175	1.320	0.675	0.790	0.967	1.174	1.323	0.676	0.790	0.968	1.173	1.320	0.677
	CasSeqGCN	0.789	0.969	1.173	1.317	0.676	0.538	0.666	0.831	0.974	0.542	0.289	0.390	0.508	0.619	0.360	0.211	0.296	0.398	0.493	0.292
	Graph-LSTM	0.751	0.905	1.083	1.217	0.660	0.517	0.632	0.778	0.886	0.523	0.284	0.382	0.498	0.599	0.354	0.212	0.293	0.386	0.474	0.288
	MMG-PopNet	**0.472**	**0.559**	**0.667**	**0.762**	**0.448**	**0.383**	**0.447**	**0.532**	**0.612**	**0.386**	**0.257**	**0.320**	**0.404**	**0.478**	**0.295**	**0.199**	**0.256**	**0.327**	**0.394**	**0.243**
MAX DEPTH	MLP	0.498	0.504	0.558	0.565	**0.356**	0.450	0.459	0.510	0.527	0.347	0.348	0.358	0.410	0.428	0.295	0.301	0.323	0.380	0.400	0.268
	DeepHawkes	0.596	0.612	0.679	0.689	0.403	0.560	0.556	0.615	0.626	0.384	0.428	0.472	0.501	0.562	0.340	0.404	0.414	0.467	0.483	0.334
	DeepCas	0.558	0.575	0.638	0.646	0.362	0.575	0.595	0.662	0.671	0.371	0.578	0.594	0.663	0.673	0.364	0.576	0.594	0.663	0.673	0.364
	CasSeqGCN	0.576	0.594	0.663	0.672	0.365	0.491	0.504	0.566	0.583	0.351	0.355	0.368	0.424	0.444	0.300	0.300	0.317	0.373	0.391	0.261
	Graph-LSTM	0.554	0.565	0.622	0.627	0.364	0.470	0.478	0.532	0.545	0.341	0.337	0.350	0.400	0.418	**0.279**	0.287	0.306	0.359	0.375	0.246
	MMG-PopNet	**0.485**	**0.493**	**0.541**	**0.542**	0.379	**0.421**	**0.421**	**0.460**	**0.468**	**0.329**	**0.328**	**0.334**	**0.374**	**0.386**	**0.279**	**0.278**	**0.291**	**0.336**	**0.346**	**0.240**
STRUCTURAL VIRALITY	MLP	0.259	0.257	0.282	0.284	**0.147**	0.232	0.232	0.256	0.264	0.143	0.181	0.184	0.210	0.219	0.123	0.157	0.166	0.193	0.206	0.113
	DeepHawkes	0.324	0.324	0.355	0.357	0.174	0.302	0.296	0.326	0.330	0.170	0.224	0.244	0.257	0.291	0.141	0.210	0.211	0.238	0.247	0.135
	DeepCas	0.307	0.308	0.337	0.338	0.155	0.317	0.320	0.349	0.349	0.159	0.318	0.319	0.349	0.350	0.157	0.318	0.319	0.349	0.350	0.157
	CasSeqGCN	0.317	0.319	0.349	0.349	0.157	0.265	0.266	0.295	0.303	0.150	0.189	0.195	0.223	0.236	0.126	0.160	0.168	0.195	0.208	0.111
	Graph-LSTM	0.306	0.304	0.328	0.327	0.158	0.256	0.256	0.280	0.287	0.146	0.182	0.187	0.212	0.222	0.118	0.152	0.163	0.189	0.199	0.105
	MMG-PopNet	**0.257**	**0.255**	**0.276**	**0.274**	0.161	**0.220**	**0.216**	**0.234**	**0.240**	**0.137**	**0.172**	**0.174**	**0.193**	**0.201**	**0.116**	**0.147**	**0.153**	**0.175**	**0.183**	**0.102**
SIZE	MLP	0.842	0.957	1.123	1.238	**0.694**	0.726	0.833	0.992	1.113	0.655	0.542	0.639	0.778	0.885	0.557	0.475	0.577	0.711	0.817	0.512
	DeepHawkes	1.262	1.490	1.816	2.007	0.967	1.084	1.261	1.545	1.740	0.852	0.630	0.802	0.991	1.169	0.626	0.520	0.619	0.782	0.894	0.518
	DeepCas	1.129	1.315	1.561	1.718	0.830	1.232	1.447	1.717	1.870	0.878	1.236	1.448	1.717	1.876	0.881	1.236	1.448	1.717	1.874	0.881
	CasSeqGCN	1.237	1.448	1.717	1.874	0.881	0.890	1.041	1.253	1.420	0.752	0.529	0.652	0.819	0.958	0.552	0.408	0.519	0.668	0.800	0.463
	Graph-LSTM	1.173	1.348	1.568	1.705	0.865	0.857	0.990	1.175	1.302	0.731	0.509	0.629	0.784	0.910	0.532	0.404	0.512	0.648	0.761	0.454
	MMG-PopNet	**0.829**	**0.936**	**1.075**	**1.169**	0.704	**0.690**	**0.762**	**0.865**	**0.957**	**0.605**	**0.482**	**0.558**	**0.669**	**0.760**	**0.483**	**0.387**	**0.464**	**0.567**	**0.657**	**0.408**
UNIQUE USERS	MLP	0.511	0.606	0.735	0.843	0.464	0.445	0.531	0.660	0.762	0.435	0.345	0.425	0.528	0.621	0.376	0.310	0.384	0.477	0.562	0.348
	DeepHawkes	0.907	1.119	1.417	1.620	0.782	0.761	0.944	1.203	1.400	0.678	0.424	0.560	0.746	0.887	0.467	0.312	0.405	0.537	0.637	0.361
	DeepCas	0.780	0.948	1.156	1.320	0.660	0.886	1.082	1.319	1.484	0.721	0.890	1.082	1.319	1.489	0.723	0.892	1.084	1.318	1.486	0.724
	CasSeqGCN	0.890	1.085	1.319	1.486	0.723	0.615	0.754	0.941	1.097	0.591	0.339	0.446	0.579	0.699	0.397	0.252	0.341	0.455	0.564	0.324
	Graph-LSTM	0.845	1.008	1.208	1.355	0.704	0.586	0.709	0.874	0.991	0.558	0.326	0.430	0.557	0.668	0.375	0.245	0.330	0.434	0.532	0.305
	MMG-PopNet	**0.510**	**0.594**	**0.703**	**0.798**	**0.454**	**0.432**	**0.495**	**0.583**	**0.665**	**0.397**	**0.301**	**0.366**	**0.455**	**0.537**	**0.313**	**0.235**	**0.293**	**0.367**	**0.444**	**0.258**
LIKE SCORE	MLP	–	–	–	–	1.312	–	–	–	–	1.299	–	–	–	–	1.260	–	–	–	–	1.225
	DeepHawkes	–	–	–	–	2.533	–	–	–	–	2.398	–	–	–	–	2.114	–	–	–	–	1.898
	DeepCas	–	–	–	–	2.347	–	–	–	–	2.502	–	–	–	–	2.510	–	–	–	–	2.514
	CasSeqGCN	–	–	–	–	2.509	–	–	–	–	2.280	–	–	–	–	1.904	–	–	–	–	1.756
	Graph-LSTM	–	–	–	–	2.469	–	–	–	–	2.205	–	–	–	–	1.839	–	–	–	–	1.649
	MMG-PopNet	–	–	–	–	**1.294**	–	–	–	–	**1.140**	–	–	–	–	**1.069**	–	–	–	–	**1.006**

Description: The results demonstrate that MMG-PopNet consistently achieves the lowest prediction error across all targets. The data reveals a steep drop in MSE as the observation window expands from 0 minutes ("Root Only") to 20 minutes, highlighting the value of early engagement data. Notably, while baseline models like DeepHawkes and DeepCas struggle significantly where they frequently exceed an MSE of 1.0 on taregts SIZE and UNIQUE USERS. Here, MMG-PopNet maintains a substantial advantage over these targets.

D.3 Statistical Significance.

To rigorously assess whether MMG-PopNet's improvements over baseline models reflect genuine performance gains rather than sampling variability, we conduct a formal statistical significance analysis on the target SIZE for the final cascade state. For each dataset, we use the largest available observation window as the input context. This corresponds to 20 min for Bluesky, 90 min for Gaming, 60 min for AMA, and 180 min for Futurology. The largest observation window provides each model with the richest possible input signal, making it the most demanding and representative setting in which to compare methods, as any advantage held by MMG-PopNet cannot be attributed to information asymmetry. We apply a paired bootstrap significance test [74] (5,000 resamples) in log1p-MSE space. The test is paired because all models predict the same set of cascades, matched by cascade identifier, and this pairing removes between-cascade variance. Errors are computed in log1p space similar to other experiments. For each test cascade $G \in \mathcal{G}^{\text{Test}}$, the ground-truth target

Table 8: r/AMA future-horizon prediction MSE across observation windows and popularity targets. **Lower is better**. Best values are **bolded** and second-best values are underlined. Column groups correspond to Root Only, 15 min, 30 min, and 60 min observation windows, and columns within each group report prediction error at $\{4\text{h}, 8\text{h}, 16\text{h}, 24\text{h}\}$ and at the final cascade state. LIKE SCORE is reported only for the final state because its intermediate-horizon labels are unavailable.

Task	Model	Root Only					15 min					30 min					60 min				
		4h	8h	16h	24h	Final	4h	8h	16h	24h	Final	4h	8h	16h	24h	Final	4h	8h	16h	24h	Final
MAX WIDTH	MLP	0.545	0.647	0.807	0.960	0.670	0.383	0.478	0.622	0.743	0.533	0.324	0.425	0.565	0.687	0.493	0.264	0.343	0.463	0.552	0.407
	DeepHawkes	0.609	0.719	0.908	1.079	0.717	0.419	0.511	0.666	0.803	0.546	0.366	0.464	0.631	0.777	0.514	0.207	0.295	0.428	0.553	0.361
	DeepCas	0.790	0.900	1.066	1.228	0.898	0.627	0.724	0.887	1.068	0.744	0.660	0.784	1.007	1.139	0.796	0.581	0.692	0.862	1.001	0.707
	CasSeqGCN	0.606	0.714	0.891	1.057	0.714	0.393	0.492	0.655	0.798	0.528	0.273	0.386	0.533	0.652	0.447	0.155	0.236	0.346	0.436	0.307
	Graph-LSTM	0.574	0.671	0.836	0.984	0.681	0.359	0.467	0.626	0.756	0.507	0.260	0.375	0.524	0.657	0.438	0.173	0.259	0.373	0.470	0.316
	MMG-PopNet	**0.536**	**0.630**	**0.787**	**0.947**	**0.663**	**0.348**	**0.440**	**0.579**	**0.694**	**0.491**	**0.259**	**0.359**	**0.489**	**0.586**	**0.426**	**0.153**	**0.223**	**0.318**	**0.392**	**0.285**
MAX DEPTH	MLP	**0.371**	**0.364**	0.364	0.361	**0.321**	0.280	0.286	0.289	0.294	0.268	0.223	0.224	0.225	0.234	0.228	0.161	0.179	0.187	0.198	0.199
	DeepHawkes	0.394	0.388	0.388	0.390	0.334	0.326	0.321	0.316	0.321	0.287	0.304	0.282	0.275	0.293	0.274	0.212	0.223	0.227	0.242	0.231
	DeepCas	0.556	0.535	0.523	0.511	0.514	0.420	0.401	0.389	0.392	0.365	0.436	0.410	0.409	0.400	0.372	0.378	0.375	0.361	0.366	0.339
	CasSeqGCN	0.394	0.388	0.384	0.384	0.328	0.299	0.298	0.299	0.309	0.273	0.224	0.226	0.230	0.243	0.224	0.151	0.168	0.175	0.188	0.187
	Graph-LSTM	0.381	0.376	0.368	0.366	0.323	0.278	0.285	0.289	0.295	0.263	0.221	0.225	0.228	0.250	0.223	0.148	0.166	0.175	0.192	0.186
	MMG-PopNet	0.378	0.369	**0.355**	**0.351**	0.330	**0.268**	**0.275**	**0.275**	**0.284**	**0.262**	**0.207**	**0.207**	**0.210**	**0.218**	**0.215**	**0.133**	**0.149**	**0.156**	**0.170**	**0.177**
STRUCTURAL VIRALITY	MLP	**0.171**	**0.158**	0.151	0.141	**0.130**	0.125	0.121	0.118	0.114	0.106	0.101	0.095	0.090	0.093	0.090	0.071	0.075	0.073	0.074	0.075
	DeepHawkes	0.184	0.170	0.162	0.154	0.137	0.150	0.139	0.130	0.125	0.114	0.142	0.121	0.114	0.120	0.111	0.091	0.092	0.089	0.094	0.089
	DeepCas	0.291	0.269	0.248	0.229	0.247	0.215	0.190	0.170	0.162	0.164	0.237	0.208	0.198	0.180	0.181	0.191	0.174	0.156	0.150	0.146
	CasSeqGCN	0.184	0.170	0.160	0.150	0.134	0.137	0.128	0.122	0.119	0.108	0.101	0.094	0.092	0.098	0.088	0.065	0.069	0.067	0.072	0.069
	Graph-LSTM	0.177	0.164	0.153	0.144	0.132	0.126	0.122	0.120	0.115	**0.105**	0.104	0.098	0.094	0.104	0.091	0.063	0.068	0.066	0.072	0.068
	MMG-PopNet	0.179	0.165	**0.147**	**0.138**	0.139	**0.121**	**0.117**	**0.112**	**0.110**	**0.105**	**0.094**	**0.087**	**0.085**	**0.088**	**0.087**	**0.060**	**0.063**	**0.061**	**0.065**	**0.066**
SIZE	MLP	0.838	0.934	1.110	1.265	**0.923**	0.553	0.661	0.827	0.960	0.714	0.443	0.551	0.687	0.828	0.631	0.331	0.433	0.559	0.661	0.511
	DeepHawkes	0.940	1.051	1.254	1.434	0.987	0.639	0.732	0.896	1.049	0.739	0.553	0.627	0.785	0.963	0.670	0.273	0.384	0.534	0.696	0.456
	DeepCas	1.398	1.472	1.634	1.792	1.460	0.996	1.073	1.236	1.435	1.057	1.089	1.189	1.411	1.538	1.139	0.934	1.051	1.218	1.374	1.011
	CasSeqGCN	0.937	1.045	1.235	1.407	0.980	0.607	0.711	0.888	1.049	0.726	0.418	0.540	0.691	0.835	0.596	0.226	0.331	0.447	0.560	0.413
	Graph-LSTM	0.887	0.985	1.149	1.300	0.942	0.544	0.666	0.842	0.981	0.693	0.409	0.539	0.686	0.860	0.595	0.261	0.367	0.487	0.605	0.436
	MMG-PopNet	**0.832**	**0.919**	**1.075**	**1.242**	0.925	**0.517**	**0.617**	**0.768**	**0.897**	**0.668**	**0.383**	**0.487**	**0.618**	**0.732**	**0.565**	**0.209**	**0.298**	**0.398**	**0.490**	**0.378**
UNIQUE USERS	MLP	0.529	0.661	0.857	1.037	**0.663**	0.372	0.495	0.669	0.817	0.532	0.317	0.440	0.612	0.767	0.502	0.251	0.351	0.497	0.600	0.410
	DeepHawkes	0.590	0.737	0.966	1.165	0.712	0.411	0.531	0.718	0.887	0.546	0.358	0.484	0.687	0.870	0.529	0.198	0.303	0.462	0.603	0.370
	DeepCas	0.800	0.937	1.129	1.306	0.937	0.612	0.739	0.938	1.140	0.747	0.658	0.811	1.072	1.238	0.818	0.554	0.694	0.896	1.057	0.714
	CasSeqGCN	0.586	0.731	0.951	1.145	0.706	0.396	0.527	0.723	0.897	0.537	0.290	0.425	0.606	0.765	0.470	0.161	0.257	0.389	0.499	0.321
	Graph-LSTM	0.556	0.687	0.886	1.059	0.677	0.352	0.488	0.677	0.834	0.510	0.273	0.410	0.589	0.759	0.460	0.184	0.286	0.426	0.539	0.338
	MMG-PopNet	**0.523**	**0.650**	**0.839**	**1.027**	0.665	**0.344**	**0.459**	**0.624**	**0.765**	**0.494**	**0.267**	**0.384**	**0.541**	**0.671**	**0.441**	**0.160**	**0.241**	**0.356**	**0.445**	**0.298**
LIKE SCORE	MLP	–	–	–	–	1.367	–	–	–	–	1.316	–	–	–	–	1.283	–	–	–	–	1.204
	DeepHawkes	–	–	–	–	1.441	–	–	–	–	1.406	–	–	–	–	1.430	–	–	–	–	1.338
	DeepCas	–	–	–	–	1.468	–	–	–	–	1.422	–	–	–	–	1.497	–	–	–	–	1.419
	CasSeqGCN	–	–	–	–	1.436	–	–	–	–	1.405	–	–	–	–	1.403	–	–	–	–	1.305
	Graph-LSTM	–	–	–	–	**1.359**	–	–	–	–	1.350	–	–	–	–	1.389	–	–	–	–	1.281
	MMG-PopNet	–	–	–	–	1.403	–	–	–	–	**1.281**	–	–	–	–	**1.239**	–	–	–	–	**1.086**

Description: While MMG-PopNet establishes dominance in long horizon predictions and larger observation windows, the data reveals a unique trend in the "Root Only" setting where the simpler MLP model frequently matches or outperforms complex graph models in predicting MAX DEPTH and STRUCTURAL VIRALITY.

is $\widetilde{\mathbf{Y}}_G^{t'} = \log(1 + \mathbf{Y}_G^{t'})$, and each model predicts $\widehat{\mathbf{Y}}_G^{t'}$. We compute the per-cascade squared error of model m as $e_G^{(m)} = (\widehat{\mathbf{Y}}_G^{t',(m)} - \widetilde{\mathbf{Y}}_G^{t'})^2$. The reported effect size compares a baseline model b against MMG-PopNet: $\Delta_b = \frac{1}{|\mathcal{G}^{\text{Test}}|} \sum_{G \in \mathcal{G}^{\text{Test}}} (e_G^{(b)} - e_G^{(\text{MMG})})$. Thus, $\Delta_b > 0$ indicates that MMG-PopNet has lower squared error than baseline b, while $\Delta_b < 0$ would indicate that the baseline performs better. The 95% confidence interval is obtained using the bootstrap percentile method. Since we conduct simultaneous multiple comparisons, we apply Benjamini–Hochberg FDR correction to control the rate of false discoveries across all tests jointly.

Results. Table 11 reports the full results. MMG-PopNet significantly outperforms all four baselines on all four datasets (**16/16 comparisons**, $p_{\text{BH}} < 0.05$). All confidence intervals are strictly positive, with lower bounds above zero. The largest improvements are against DeepCas, where Δ_b exceeds $+1.0$ in log1p-MSE on every dataset, including Gaming ($\Delta_b = +1.573$) and Bluesky ($\Delta_b = +1.522$). CasSeqGCN is the closest competitor. The smallest significant improvement occurs on Futurology ($\Delta_b = +0.102$, 95% CI $[+0.029, +0.183]$, $p_{\text{BH}} = 0.004$). These results show that MMG-PopNet's gains are consistent, statistically reliable, and not driven by any single dataset or baseline.

Table 9: r/Gaming future-horizon prediction MSE across observation windows and popularity targets. **Lower is better**. Best values are **bolded** and second-best values are underlined. Column groups correspond to Root Only, 20 min, 50 min, and 90 min observation windows, and columns within each group report prediction error at $\{4\text{h}, 8\text{h}, 16\text{h}, 24\text{h}\}$ and at the final cascade state. LIKE SCORE is reported only for the final state because its intermediate-horizon labels are unavailable.

Task	Model	Root Only					20 min					50 min					90 min				
		4h	8h	16h	24h	Final	4h	8h	16h	24h	Final	4h	8h	16h	24h	Final	4h	8h	16h	24h	Final
MAX WIDTH	MLP	1.301	1.623	1.882	2.067	1.778	0.701	0.952	1.176	1.300	1.174	0.506	0.715	0.868	0.970	0.879	0.406	0.567	0.693	0.783	0.715
	DeepHawkes	1.401	1.757	2.117	2.393	1.919	0.933	1.111	1.368	1.552	1.295	0.594	0.974	1.279	1.580	1.256	0.354	0.573	0.841	1.073	0.866
	DeepCas	1.996	2.317	2.722	2.898	2.353	1.367	1.669	1.975	2.147	1.815	1.336	1.652	1.900	1.940	1.786	1.250	1.538	1.757	1.913	1.700
	CasSeqGCN	1.393	1.713	1.992	2.164	1.868	0.664	0.931	1.186	1.345	1.135	0.388	0.649	0.870	0.971	0.855	0.225	0.393	0.572	0.665	0.590
	Graph-LSTM	1.345	1.651	1.941	2.101	1.815	**0.622**	**0.893**	1.137	1.285	1.104	**0.334**	0.589	0.785	0.897	0.760	**0.211**	0.411	0.619	0.738	0.587
	MMG-PopNet	**1.160**	**1.442**	**1.687**	**1.818**	**1.554**	0.718	0.940	**1.116**	**1.213**	**1.103**	0.343	**0.511**	**0.659**	**0.733**	**0.662**	0.214	**0.348**	**0.494**	**0.575**	**0.490**
MAX DEPTH	MLP	**0.295**	0.320	0.342	0.366	0.342	**0.215**	**0.238**	0.273	0.296	0.280	0.157	0.175	0.188	0.192	0.208	0.126	0.150	0.168	0.173	0.176
	DeepHawkes	0.317	0.332	0.366	0.409	0.354	0.275	0.282	0.315	0.346	0.300	0.198	0.232	0.264	0.305	0.279	0.199	0.210	0.228	0.243	0.234
	DeepCas	0.577	0.564	0.605	0.600	0.589	0.337	0.349	0.380	0.397	0.371	0.338	0.346	0.361	0.353	0.375	0.313	0.323	0.334	0.326	0.323
	CasSeqGCN	0.309	0.329	0.355	0.381	0.347	0.237	0.262	0.297	0.324	0.295	0.187	0.212	0.234	0.248	0.239	0.172	0.192	0.209	0.210	0.208
	Graph-LSTM	0.307	0.328	0.357	0.384	0.351	0.217	0.243	0.282	0.310	0.281	0.152	0.181	0.201	0.213	0.216	**0.101**	0.135	0.158	0.172	0.166
	MMG-PopNet	0.296	**0.310**	**0.329**	**0.346**	**0.336**	0.225	0.242	**0.261**	**0.278**	**0.273**	**0.142**	**0.163**	**0.179**	**0.189**	**0.198**	0.112	**0.130**	**0.146**	**0.151**	**0.155**
STRUCTURAL VIRALITY	MLP	**0.118**	**0.113**	**0.106**	**0.109**	**0.094**	**0.084**	**0.085**	**0.088**	0.093	**0.078**	**0.054**	**0.054**	0.059	0.056	**0.057**	0.048	0.049	0.050	0.051	0.047
	DeepHawkes	0.127	0.118	0.113	0.119	0.098	0.113	0.106	0.106	0.114	0.090	0.078	0.078	0.084	0.087	0.080	0.092	0.081	0.079	0.084	0.071
	DeepCas	0.351	0.307	0.280	0.248	0.282	0.157	0.137	0.128	0.129	0.118	0.146	0.124	0.120	0.106	0.120	0.145	0.121	0.109	0.105	0.099
	CasSeqGCN	0.124	0.117	0.110	0.114	0.095	0.095	0.096	0.099	0.106	0.086	0.066	0.069	0.075	0.075	0.071	0.072	0.071	0.074	0.078	0.066
	Graph-LSTM	0.125	0.119	0.116	0.120	0.100	0.092	0.091	0.093	0.100	0.082	0.063	0.060	0.062	0.061	0.062	**0.043**	**0.048**	0.050	0.053	0.047
	MMG-PopNet	0.126	0.119	0.109	0.110	0.100	0.098	0.093	**0.088**	**0.090**	0.079	0.061	0.058	**0.058**	**0.055**	0.059	0.058	0.050	**0.046**	**0.045**	**0.044**
SIZE	MLP	1.494	1.894	2.174	2.384	2.034	0.825	1.143	1.417	1.570	1.395	0.545	0.792	0.990	1.109	0.990	0.422	0.631	0.791	0.890	0.799
	DeepHawkes	1.608	2.101	2.488	2.793	2.246	1.063	1.338	1.668	1.901	1.511	0.675	1.137	1.519	1.890	1.457	0.417	0.696	1.023	1.303	1.015
	DeepCas	2.579	2.949	3.376	3.575	2.958	1.585	1.949	2.284	2.477	2.070	1.592	1.944	2.226	2.246	2.064	1.436	1.774	2.019	2.159	1.906
	CasSeqGCN	1.589	1.975	2.274	2.476	2.103	0.840	1.186	1.491	1.684	1.393	0.472	0.793	1.069	1.195	1.010	0.275	0.500	0.727	0.842	0.717
	Graph-LSTM	1.546	1.922	2.238	2.421	2.067	**0.784**	**1.136**	1.430	1.618	1.358	0.472	0.792	1.015	1.162	0.972	0.265	0.524	0.779	0.924	0.725
	MMG-PopNet	**1.380**	**1.715**	**1.970**	**2.126**	**1.828**	0.873	1.148	**1.336**	**1.450**	**1.322**	**0.405**	**0.609**	**0.793**	**0.887**	**0.791**	**0.250**	**0.415**	**0.590**	**0.678**	**0.580**
UNIQUE USERS	MLP	1.397	1.796	2.103	2.319	1.905	0.776	1.092	1.364	1.513	1.296	0.531	0.780	0.977	1.100	0.949	0.408	0.613	0.774	0.874	0.760
	DeepHawkes	1.494	1.927	2.334	2.654	2.048	1.039	1.282	1.593	1.809	1.428	0.665	1.126	1.501	1.856	1.413	0.397	0.678	1.003	1.269	0.974
	DeepCas	2.246	2.633	3.074	3.265	2.623	1.455	1.826	2.182	2.381	1.926	1.451	1.825	2.122	2.179	1.928	1.323	1.666	1.932	2.088	1.793
	CasSeqGCN	1.480	1.876	2.201	2.407	1.983	0.780	1.120	1.428	1.620	1.289	0.460	0.780	1.059	1.191	0.972	0.257	0.475	0.698	0.810	0.672
	Graph-LSTM	1.442	1.823	2.159	2.345	1.939	**0.725**	**1.069**	1.367	1.549	1.253	0.431	0.740	0.968	1.120	0.891	0.264	0.513	0.761	0.898	0.684
	MMG-PopNet	**1.271**	**1.608**	**1.893**	**2.049**	**1.690**	0.800	1.074	**1.279**	**1.391**	**1.211**	**0.381**	**0.582**	**0.766**	**0.857**	**0.737**	**0.232**	**0.394**	**0.567**	**0.646**	**0.539**
LIKE SCORE	MLP	–	–	–	–	5.856	–	–	–	–	5.410	–	–	–	–	4.883	–	–	–	–	4.614
	DeepHawkes	–	–	–	–	6.268	–	–	–	–	6.068	–	–	–	–	6.497	–	–	–	–	6.629
	DeepCas	–	–	–	–	6.466	–	–	–	–	6.291	–	–	–	–	6.253	–	–	–	–	6.673
	CasSeqGCN	–	–	–	–	6.208	–	–	–	–	6.124	–	–	–	–	5.992	–	–	–	–	6.048
	Graph-LSTM	–	–	–	–	5.796	–	–	–	–	6.032	–	–	–	–	5.647	–	–	–	–	5.505
	MMG-PopNet	–	–	–	–	**5.180**	–	–	–	–	**4.819**	–	–	–	–	**4.325**	–	–	–	–	**4.073**

Description: The results show that while MMG-PopNet remains the most robust model for predicting "Final" outcomes, Graph-LSTM and MLP are highly competitive and occasionally superior at forecasting short-term 4-hour horizons. The table also underscores the extreme difficulty of predicting the LIKE SCORE in gaming communities, with all models exhibiting massive error rates (MSE > 4.0). However, feeding the models 90 minutes of initial social cascade data reduces the error by over 60% compared to Root Only predictions.

Table 10: r/Futurology future-horizon prediction MSE across observation windows and popularity targets. **Lower is better**. Best values are **bolded** and second-best values are underlined. Column groups correspond to Root Only, 30 min, 90 min, and 180 min observation windows, and columns within each group report prediction error at $\{4\text{h}, 8\text{h}, 16\text{h}, 24\text{h}\}$ and at the final cascade state. LIKE SCORE is reported only for the final state because its intermediate-horizon labels are unavailable.

Metric	Model	Root Only					30 min					90 min					180 min				
		4h	8h	16h	24h	Final	4h	8h	16h	24h	Final	4h	8h	16h	24h	Final	4h	8h	16h	24h	Final
MAX WIDTH	MLP	1.100	1.522	1.879	1.968	1.761	0.640	1.029	1.404	1.445	1.312	0.301	0.506	0.749	0.812	0.706	0.246	0.329	0.461	0.526	0.454
	DeepHawkes	1.141	1.636	2.118	2.249	1.894	0.813	1.260	1.819	1.969	1.556	0.632	0.957	1.394	1.468	1.217	0.557	0.730	1.016	1.186	0.972
	DeepCas	1.158	1.556	1.955	2.014	1.732	0.686	1.460	1.841	1.961	1.636	0.802	1.095	1.363	1.454	1.228	0.768	1.004	1.229	1.341	1.122
	CasSeqGCN	1.129	1.576	1.954	2.002	1.814	0.585	0.967	1.378	1.454	1.243	0.226	0.458	0.728	0.807	0.658	**0.084**	**0.200**	0.359	0.427	0.347
	Graph-LSTM	1.090	1.508	1.889	1.951	1.721	**0.541**	**0.931**	1.326	1.419	1.182	**0.206**	**0.448**	**0.713**	**0.779**	0.647	0.091	0.219	0.382	0.465	0.365
	MMG-PopNet	**1.019**	**1.419**	**1.724**	**1.789**	**1.559**	0.613	0.963	**1.286**	**1.360**	**1.170**	0.309	0.483	0.715	0.780	**0.633**	0.123	0.210	**0.333**	**0.391**	**0.320**
MAX DEPTH	MLP	0.422	0.468	0.457	0.454	0.572	0.291	0.362	0.378	0.387	0.479	0.174	0.231	0.272	0.276	0.294	0.102	0.143	0.182	0.191	0.212
	DeepHawkes	0.443	0.508	0.529	0.540	0.621	0.343	0.425	0.491	0.518	0.555	0.307	0.364	0.446	0.455	0.444	0.297	0.327	0.378	0.423	0.425
	DeepCas	0.450	0.500	0.518	0.506	0.589	0.419	0.495	0.512	0.527	0.572	0.320	0.380	0.413	0.422	0.417	0.274	0.333	0.362	0.354	0.361
	CasSeqGCN	0.432	0.483	0.481	0.472	0.586	0.305	0.371	0.400	0.412	0.483	0.208	0.257	0.296	0.310	0.314	0.114	0.165	0.198	0.211	0.227
	Graph-LSTM	0.421	0.469	0.469	0.460	0.569	**0.279**	0.359	0.383	0.397	0.472	**0.159**	0.225	0.268	0.273	0.292	**0.068**	0.131	0.175	0.201	0.212
	MMG-PopNet	**0.397**	**0.453**	**0.440**	**0.436**	**0.522**	**0.279**	**0.355**	**0.360**	**0.371**	**0.447**	0.169	**0.216**	**0.251**	**0.262**	**0.276**	0.071	**0.118**	**0.156**	**0.173**	**0.198**
STRUCTURAL VIRALITY	MLP	0.202	0.198	**0.169**	**0.161**	0.193	0.143	0.156	0.148	0.147	0.165	0.075	0.094	0.111	0.113	0.104	0.038	0.053	0.066	0.071	0.072
	DeepHawkes	0.210	0.211	0.198	0.200	0.209	0.166	0.179	0.187	0.192	0.190	0.155	0.167	0.200	0.200	0.173	0.145	0.136	0.147	0.162	0.154
	DeepCas	0.222	0.219	0.203	0.193	0.210	0.210	0.222	0.208	0.208	0.209	0.151	0.160	0.171	0.170	0.155	0.123	0.134	0.139	0.133	0.132
	CasSeqGCN	0.207	0.205	0.178	0.176	0.198	0.152	0.158	0.154	0.155	0.165	0.094	0.108	0.125	0.129	0.114	0.045	0.062	0.074	0.082	0.081
	Graph-LSTM	0.201	0.199	0.176	0.165	0.192	**0.138**	**0.153**	0.148	0.148	0.161	**0.068**	**0.090**	0.107	0.107	**0.099**	**0.028**	0.050	0.065	0.073	0.070
	MMG-PopNet	**0.197**	**0.196**	0.171	**0.161**	**0.182**	0.145	0.156	**0.140**	**0.138**	**0.157**	0.084	0.091	**0.102**	**0.105**	**0.099**	0.033	**0.046**	**0.057**	**0.064**	**0.068**
SIZE	MLP	1.674	2.276	2.678	2.756	2.658	0.979	1.598	2.094	2.152	2.055	0.399	0.765	1.146	1.255	1.080	0.264	0.425	0.646	0.719	0.642
	DeepHawkes	1.743	2.528	3.136	3.253	2.943	1.247	1.948	2.733	2.954	2.437	1.062	1.633	2.348	2.503	2.004	0.845	1.093	1.513	1.739	1.455
	DeepCas	1.771	2.347	2.826	2.870	2.632	1.629	2.287	2.768	2.923	2.544	1.138	1.616	2.011	2.146	1.803	1.066	1.456	1.778	1.872	1.602
	CasSeqGCN	1.736	2.390	2.812	2.829	2.755	0.963	1.573	2.128	2.235	2.013	0.372	0.786	1.211	1.358	1.103	**0.085**	0.315	0.575	0.665	0.566
	Graph-LSTM	1.650	2.258	2.704	2.749	2.604	**0.890**	1.509	2.037	2.156	1.919	**0.350**	0.762	1.158	1.250	1.070	0.131	0.362	0.629	0.750	0.612
	MMG-PopNet	**1.559**	**2.129**	**2.480**	**2.516**	**2.353**	0.961	**1.500**	**1.907**	**1.992**	**1.823**	0.427	**0.731**	**1.074**	**1.169**	**0.962**	0.143	**0.302**	**0.498**	**0.568**	**0.506**
UNIQUE USERS	MLP	1.344	1.888	2.282	2.345	2.089	0.827	1.359	1.808	1.835	1.619	0.342	0.662	0.979	1.067	0.862	0.241	0.385	0.571	0.624	0.515
	DeepHawkes	1.408	2.095	2.650	2.736	2.326	1.024	1.636	2.324	2.481	1.928	0.805	1.282	1.854	1.966	1.516	0.676	0.925	1.292	1.473	1.153
	DeepCas	1.425	1.948	2.385	2.411	2.072	1.312	1.887	2.318	2.427	1.999	0.917	1.337	1.662	1.766	1.402	0.882	1.225	1.506	1.599	1.276
	CasSeqGCN	1.385	1.971	2.395	2.415	2.163	0.805	1.332	1.826	1.897	1.582	0.308	0.658	1.007	1.122	0.858	**0.092**	0.283	0.501	0.561	0.437
	Graph-LSTM	1.332	1.877	2.300	2.344	2.043	**0.724**	1.257	1.724	1.808	1.489	**0.304**	0.664	1.001	1.075	0.834	0.137	0.316	0.536	0.618	0.451
	MMG-PopNet	**1.250**	**1.772**	**2.122**	**2.151**	**1.849**	0.773	**1.247**	**1.624**	**1.691**	**1.425**	0.370	**0.633**	**0.917**	**0.997**	**0.764**	0.143	**0.275**	**0.438**	**0.480**	**0.389**
LIKE SCORE	MLP	–	–	–	–	6.469	–	–	–	–	6.309	–	–	–	–	4.524	–	–	–	–	3.709
	DeepHawkes	–	–	–	–	7.002	–	–	–	–	6.877	–	–	–	–	6.688	–	–	–	–	5.646
	DeepCas	–	–	–	–	5.749	–	–	–	–	6.013	–	–	–	–	4.828	–	–	–	–	4.925
	CasSeqGCN	–	–	–	–	6.800	–	–	–	–	6.639	–	–	–	–	5.402	–	–	–	–	5.093
	Graph-LSTM	–	–	–	–	6.447	–	–	–	–	5.822	–	–	–	–	4.674	–	–	–	–	4.012
	MMG-PopNet	–	–	–	–	**5.308**	–	–	–	–	**4.886**	–	–	–	–	**3.428**	–	–	–	–	**2.897**

Description: The results that for given 180 min of observation data, graph-based models like CasSeqGCN can achieve near-perfect accuracy (MSE < 0.1) for immediate 4-hour horizon predictions regarding cascade SIZE and MAX WIDTH. However, prediction accuracy decays steeply as the horizon extends to 24 hours. MMG-PopNet distinguishes itself most prominently in the LIKE SCORE category, leveraging multimodal data to achieve an MSE of 2.897 at the 180-minute mark, vastly outperforming the nearest baseline (MLP at 3.709).

Table 11: Statistical significance of **MMG-PopNet (Ours)** versus baselines for SIZE prediction one the final cascade state in log1p-MSE space. $\Delta_b > 0$ indicates that MMG-PopNet has lower squared error than baseline b. 95% confidence intervals and p_{raw} are computed using a paired bootstrap test with 5,000 resamples. p_{BH} denotes Benjamini–Hochberg correction across all 16 comparisons at $\alpha = 0.05$. **Yes** indicates that MMG-PopNet is significantly better.

Dataset	**Baseline**	n	$\boldsymbol{\Delta_b}$	**95% CI**	$\boldsymbol{p_{\mathrm{raw}}}$	$\boldsymbol{p_{\mathrm{BH}}}$	**Sig.**
Bluesky	CasSeqGCN	1510	$+0.2213$	$[+0.1729, +0.2717]$	< 0.001	< 0.001	**Yes**
	Graph-LSTM	1510	$+0.1375$	$[+0.0983, +0.1795]$	< 0.001	< 0.001	**Yes**
	DeepHawkes	1510	$+0.3601$	$[+0.3054, +0.4171]$	< 0.001	< 0.001	**Yes**
	DeepCas	1510	$+1.1522$	$[+1.0206, +1.2909]$	< 0.001	< 0.001	**Yes**
r/Gaming	CasSeqGCN	675	$+0.1328$	$[+0.0572, +0.2078]$	< 0.001	< 0.001	**Yes**
	Graph-LSTM	675	$+0.3519$	$[+0.2695, +0.4358]$	< 0.001	< 0.001	**Yes**
	DeepHawkes	675	$+0.6019$	$[+0.4750, +0.7306]$	< 0.001	< 0.001	**Yes**
	DeepCas	675	$+1.5730$	$[+1.3016, +1.8464]$	< 0.001	< 0.001	**Yes**
r/AMA	CasSeqGCN	1420	$+0.0986$	$[+0.0696, +0.1298]$	< 0.001	< 0.001	**Yes**
	Graph-LSTM	1420	$+0.2269$	$[+0.1923, +0.2626]$	< 0.001	< 0.001	**Yes**
	DeepHawkes	1420	$+0.2167$	$[+0.1776, +0.2578]$	< 0.001	< 0.001	**Yes**
	DeepCas	1420	$+1.0033$	$[+0.8893, +1.1251]$	< 0.001	< 0.001	**Yes**
r/Futurology	CasSeqGCN	527	$+0.1016$	$[+0.0293, +0.1828]$	0.004	0.004	**Yes**
	Graph-LSTM	527	$+0.2892$	$[+0.1987, +0.3867]$	< 0.001	< 0.001	**Yes**
	DeepHawkes	527	$+1.3799$	$[+1.1493, +1.6302]$	< 0.001	< 0.001	**Yes**
	DeepCas	527	$+1.3210$	$[+1.0222, +1.6373]$	< 0.001	< 0.001	**Yes**

E Unified Training Details and Full Results

Tables 12 and 13 report the full comparison between dataset-specific MMG-PopNet and unified-trained MMG-PopNet across datasets, observation windows, and prediction targets. The dataset-specific setting trains a separate model for each dataset and observation window. In contrast, MMG-PopNet (unified-dataset) is trained once on the combined training split and is evaluated separately on each dataset-window test split. For each training example, the unified model receives one-hot indicators for the dataset and observation window. These indicators allow the model to condition its predictions on both the community source and the amount of observed cascade history.

The full results show that the benefit of unified training is largest on the Reddit datasets. On r/AMA, r/Gaming, and r/Futurology, MMG-PopNet (unified-dataset) reduces the dataset-level average MSE for every prediction target. The reductions are most pronounced for LIKE SCORE, SIZE, and UNIQUE USERS. These targets have the largest dataset-level errors under dataset-specific training on r/Gaming and r/Futurology, and they also show the largest absolute reductions after unified training. For example, on r/Gaming, the dataset-level average MSE for LIKE SCORE decreases from 4.525 to 1.479. On r/Futurology, the corresponding average decreases from 3.890 to 1.569.

Bluesky follows a different pattern. On this dataset, the dataset-specific model obtains lower dataset-level average MSE for most targets. The unified model remains close in absolute MSE, but it does not improve over the dataset-specific model on Bluesky. This result indicates that the cross-dataset benefit is stronger when the evaluation dataset is closer to the Reddit training communities.

Averaged across all datasets and observation windows, MMG-PopNet (unified-dataset) obtains lower MSE for all six targets. The overall average decreases from 0.678 to 0.426 for MAX WIDTH, from 0.284 to 0.232 for MAX DEPTH, from 0.932 to 0.586 for SIZE, from 0.104 to 0.084 for STRUCTURAL VIRALITY, from 0.752 to 0.453 for UNIQUE USERS, and from 2.669 to 1.217 for LIKE SCORE. Thus, unified training improves benchmark-level performance across targets while maintaining comparable accuracy on the outlying Bluesky platform.

Table 12: **Comparison between dataset-specific and unified-dataset MMG-PopNet training, Part I.** Dataset-specific models are trained separately for each dataset and observation window, while the unified-dataset model is trained once on the combined training data across all datasets and windows. Results are reported as MSE, where **lower is better and marked in bold**. For readability, results are split by dataset. Each subtable reports observation-window MSEs, the dataset-level average, and the same final overall average across all datasets. This table reports results for Bluesky and r/AMA; Table 13 continues with r/Gaming and r/Futurology.

Bluesky

Task	Model	0	2	10	20	Dataset Avg	Overall Avg
MAX WIDTH	MMG-PopNet (specific)	0.457	**0.369**	**0.284**	**0.234**	**0.336**	0.678
	MMG-PopNet (unified)	**0.449**	0.401	0.316	0.276	0.361	**0.426**
MAX DEPTH	MMG-PopNet (specific)	0.363	**0.325**	**0.273**	**0.238**	**0.300**	0.284
	MMG-PopNet (unified)	**0.347**	0.332	0.280	0.247	0.301	**0.232**
SIZE	MMG-PopNet (specific)	0.705	0.587	**0.470**	**0.397**	**0.540**	0.932
	MMG-PopNet (unified)	**0.688**	**0.637**	0.517	0.450	0.573	**0.586**
STRUCTURAL VIRALITY	MMG-PopNet (specific)	0.155	**0.135**	**0.114**	**0.101**	**0.126**	0.104
	MMG-PopNet (unified)	**0.144**	0.138	0.117	0.104	**0.126**	**0.084**
UNIQUE USERS	MMG-PopNet (specific)	**0.467**	**0.383**	**0.302**	**0.255**	**0.352**	0.752
	MMG-PopNet (unified)	0.472	0.426	0.341	0.298	0.384	**0.453**
LIKE SCORE	MMG-PopNet (specific)	1.260	**1.087**	**1.026**	**0.950**	**1.081**	2.669
	MMG-PopNet (unified)	**1.238**	1.191	1.093	1.043	1.141	**1.217**

r/AMA

Task	Model	0	15	30	60	Dataset Avg	Overall Avg
MAX WIDTH	MMG-PopNet (specific)	0.625	0.491	0.429	0.276	0.455	0.678
	MMG-PopNet (unified)	**0.476**	**0.354**	**0.239**	**0.184**	**0.313**	**0.426**
MAX DEPTH	MMG-PopNet (specific)	0.314	0.256	0.219	0.172	0.240	0.284
	MMG-PopNet (unified)	**0.284**	**0.236**	**0.189**	**0.158**	**0.217**	**0.232**
SIZE	MMG-PopNet (specific)	0.863	0.661	0.573	0.366	0.616	0.932
	MMG-PopNet (unified)	**0.681**	**0.509**	**0.349**	**0.257**	**0.449**	**0.586**
STRUCTURAL VIRALITY	MMG-PopNet (specific)	0.130	0.100	0.087	0.064	0.095	0.104
	MMG-PopNet (unified)	**0.115**	**0.093**	**0.075**	**0.060**	**0.086**	**0.084**
UNIQUE USERS	MMG-PopNet (specific)	0.622	0.494	0.447	0.290	0.463	0.752
	MMG-PopNet (unified)	**0.454**	**0.343**	**0.235**	**0.180**	**0.303**	**0.453**
LIKE SCORE	MMG-PopNet (specific)	1.311	1.212	1.166	1.028	1.179	2.669
	MMG-PopNet (unified)	**0.810**	**0.737**	**0.590**	**0.582**	**0.680**	**1.217**

Table 13: **Comparison between dataset-specific and unified-dataset MMG-PopNet training, Part II.** Continuation of Table 12. Results are reported as MSE, where **lower is better and marked in bold**. For readability, results are split by dataset. Each subtable reports observation-window MSEs, the dataset-level average, and the same final overall average across all datasets. This table reports results for r/Gaming and r/Futurology.

r/Gaming							
Task	**Model**	**0**	**20**	**50**	**90**	**Dataset Avg**	**Overall Avg**
MAX WIDTH	MMG-PopNet (specific)	1.557	1.096	0.714	0.562	0.982	0.678
	MMG-PopNet (unified)	**0.962**	**0.631**	**0.299**	**0.266**	**0.539**	**0.426**
MAX DEPTH	MMG-PopNet (specific)	0.335	0.275	0.200	0.160	0.243	0.284
	MMG-PopNet (unified)	**0.248**	**0.204**	**0.140**	**0.121**	**0.178**	**0.232**
SIZE	MMG-PopNet (specific)	1.829	1.317	0.842	0.653	1.160	0.932
	MMG-PopNet (unified)	**1.092**	**0.754**	**0.348**	**0.294**	**0.622**	**0.586**
STRUCTURAL VIRALITY	MMG-PopNet (specific)	0.098	0.081	0.057	0.045	0.070	0.104
	MMG-PopNet (unified)	**0.071**	**0.057**	**0.038**	**0.032**	**0.050**	**0.084**
UNIQUE USERS	MMG-PopNet (specific)	1.693	1.218	0.792	0.613	1.079	0.752
	MMG-PopNet (unified)	**1.004**	**0.684**	**0.318**	**0.274**	**0.570**	**0.453**
LIKE SCORE	MMG-PopNet (specific)	5.067	4.654	4.307	4.073	4.525	2.669
	MMG-PopNet (unified)	**2.077**	**1.809**	**0.960**	**1.070**	**1.479**	**1.217**

r/Futurology							
Task	**Model**	**0**	**30**	**90**	**180**	**Dataset Avg**	**Overall Avg**
MAX WIDTH	MMG-PopNet (specific)	1.581	1.132	0.665	0.372	0.938	0.678
	MMG-PopNet (unified)	**0.886**	**0.634**	**0.242**	**0.202**	**0.491**	**0.426**
MAX DEPTH	MMG-PopNet (specific)	0.517	0.418	0.282	0.205	0.356	0.284
	MMG-PopNet (unified)	**0.336**	**0.282**	**0.164**	**0.147**	**0.232**	**0.232**
SIZE	MMG-PopNet (specific)	2.356	1.743	0.997	0.559	1.414	0.932
	MMG-PopNet (unified)	**1.253**	**0.932**	**0.348**	**0.267**	**0.700**	**0.586**
STRUCTURAL VIRALITY	MMG-PopNet (specific)	0.182	0.146	0.101	0.073	0.126	0.104
	MMG-PopNet (unified)	**0.110**	**0.090**	**0.053**	**0.047**	**0.075**	**0.084**
UNIQUE USERS	MMG-PopNet (specific)	1.859	1.374	0.785	0.439	1.114	0.752
	MMG-PopNet (unified)	**0.987**	**0.736**	**0.277**	**0.219**	**0.555**	**0.453**
LIKE SCORE	MMG-PopNet (specific)	4.999	4.596	3.214	2.750	3.890	2.669
	MMG-PopNet (unified)	**2.283**	**2.054**	**1.041**	**0.898**	**1.569**	**1.217**

F Detailed LLM-based comparison.

To evaluate whether LLMs can serve as competitive predictors for multimodal social popularity forecasting, we conduct experiments comparing the performance of various LLM-based approaches against MMG-PopNet. We evaluate three models, namely `Qwen3-VL-8B-Instruct`, `Gemma-3-12b-it`, and `GPT-4o-mini`. To manage computational costs for the Bluesky dataset, we randomly sample 15,000 cascades from the original data and evaluate all models including MMG-PopNet on this subset. Furthermore, large cascades can contain excessive content that exceeds standard context windows and prevents fair comparison. We address this by restricting the maximum number of observed nodes to 25 for all social cascades and apply this exact limit to MMG-PopNet. Given the early observation data statistics where the average node count ranges from 1.62 to 23.6, this setting provides sufficient context for the vast majority of cascades.

We evaluate three prompting and training settings. In the zero-shot setting, the observed cascade prefix is serialized as a structured JSON input containing the available content and interaction, along with an image, and the model predicts the popularity targets directly. In the retrieval-augmented few-shot setting, each test instance is paired with four training examples selected by root-post similarity, and their ground-truth targets are included in the prompt. In the fine-tuning setting, the model is trained on early-observation cascade inputs to predict the same targets as MMG-PopNet. Note that `GPT-4o-mini` is only evaluated under the zero-shot and few-shot settings. Prompts used for the experiment are provided here F.

Tables 14-19 report results across the LLM models and settings, and datasets. Across both datasets and all tested models, MMG-PopNet achieves the lowest MSE for every target and observation window. Zero-shot prompting consistently performs worst, confirming that structured input alone is insufficient for calibrated numerical prediction. Few-shot prompting substantially reduces error, especially under root-only and early observation windows where retrieved examples provide useful target-range calibration for sparse cascades. Fine-tuning becomes more effective as the observation window increases for `Qwen3-VL-8B-Instruct` and to some extent for `Gemma-3-12b-it`. This trend is most visible for targets such as MAX WIDTH, STRUCTURAL VIRALITY, SIZE, and UNIQUE USERS, where longer prefixes expose richer interaction and temporal signals that supervised adaptation can exploit. Here, `Qwen3-VL-8B-Instruct` exhibits a more consistent increase with fine-tuning whereas, on `Gemma-3-12b-it` few-shot setting is better across observation windows for most targets for the r/Futurology dataset and with mixed results for Bluesky dataset.

For `GPT-4o-mini`, the few shot setting results as the best setting when compared with the zero-shot. However, it still lags far behind the MMG-PopNet in performance. Improvement in few-shot setting with increased observation window is higher and consistent in r/Futurology dataset when compared to Bluesky.

Overall, the tables show that fine-tuning does not uniformly dominate few-shot prompting. In several early-window cases, few-shot prompting remains competitive or stronger, particularly for MAX DEPTH and some engagement targets. This indicates that retrieval is highly useful when the observed prefix contains limited cascade evidence, while fine-tuning benefits more from denser prefixes. Overall, the detailed results support the main finding. LLM-based adaptation improves over zero-shot prompting, but MMG-PopNet remains consistently stronger because it is optimized for the benchmark formulation and directly models observed cascade dynamics rather than predicting from serialized inputs alone.

Table 14: LLM Qwen3-VL-8B-Instruct comparison with MMG-PopNet on Bluesky across observation windows using MSE values.

Task	Root Only				2				10				20			
	Zero	Few	Fine-Tune	MMG-PopNet	Zero	Few	Fine-Tune	MMG-PopNet	Zero	Few	Fine-Tune	MMG-PopNet	Zero	Few	Fine-Tune	MMG-PopNet
MAX WIDTH	3.045	1.666	1.923	**0.666**	2.615	1.564	1.295	**0.577**	2.124	1.038	0.816	**0.506**	1.989	0.889	0.647	**0.459**
MAX DEPTH	1.715	0.730	0.905	**0.523**	1.577	0.790	0.835	**0.480**	1.240	0.666	0.677	**0.428**	1.048	0.561	0.604	**0.415**
STRUCTURAL VIRALITY	2.368	0.363	0.429	**0.243**	1.947	0.414	0.386	**0.225**	1.445	0.353	0.303	**0.196**	1.256	0.298	0.267	**0.191**
SIZE	6.596	2.293	2.668	**1.007**	5.294	2.289	1.983	**0.870**	3.544	1.476	1.284	**0.751**	2.842	1.194	1.065	**0.700**
UNIQUE USERS	4.051	1.867	2.009	**0.690**	3.389	1.739	1.349	**0.600**	2.557	1.085	0.837	**0.521**	2.213	0.904	0.675	**0.468**
LIKE SCORE	16.256	5.121	6.625	**1.540**	16.162	4.980	5.170	**1.445**	14.383	4.087	3.846	**1.394**	13.444	3.615	3.164	**1.327**

Table 15: LLM Qwen3-VL-8B-Instruct comparison with MMG-PopNet on r/Futurology across observation windows using MSE values.

Task	Root Only				30 min				90 min				180 min			
	Zero	Few	Fine-Tune	MMG-PopNet	Zero	Few	Fine-Tune	MMG-PopNet	Zero	Few	Fine-Tune	MMG-PopNet	Zero	Few	Fine-Tune	MMG-PopNet
Max Width	4.912	2.274	2.466	**1.581**	3.661	2.005	1.774	**1.321**	2.952	1.417	0.971	**0.794**	2.758	1.155	0.782	**0.457**
Max Depth	1.605	0.785	1.012	**0.517**	1.151	0.685	0.782	**0.496**	0.769	0.439	0.517	**0.296**	0.551	0.357	0.425	**0.242**
Structural Virality	2.164	0.264	0.343	**0.182**	1.198	0.265	0.262	**0.178**	0.667	0.188	0.179	**0.104**	0.434	0.160	0.141	**0.091**
Size	9.238	3.895	4.080	**2.356**	5.706	3.096	2.953	**1.921**	3.572	1.986	1.721	**1.141**	2.630	1.466	1.358	**0.700**
Unique Users	6.693	3.005	3.028	**1.859**	4.365	2.440	2.181	**1.551**	2.798	1.616	1.273	**0.917**	2.067	1.217	1.038	**0.552**
Like Score	21.135	8.859	8.863	**4.999**	20.018	8.178	8.572	**5.030**	17.077	6.933	6.944	**3.400**	13.981	7.052	6.016	**3.039**

Table 16: LLM Gemma-3-12b-it comparison with MMG-PopNet on Bluesky across observation windows using MSE values.

Task	Root Only				2				10				20			
	Zero	Few	Fine-Tune	MMG-PopNet	Zero	Few	Fine-Tune	MMG-PopNet	Zero	Few	Fine-Tune	MMG-PopNet	Zero	Few	Fine-Tune	MMG-PopNet
Max Width	5.366	1.312	1.993	**0.666**	3.796	1.320	1.391	**0.577**	2.545	0.969	0.812	**0.506**	2.179	0.860	0.657	**0.459**
Max Depth	2.285	0.628	0.961	**0.523**	1.655	0.659	0.800	**0.480**	1.269	0.577	0.721	**0.428**	1.105	0.533	0.603	**0.415**
Structural Virality	2.362	0.286	0.490	**0.243**	1.828	0.314	0.405	**0.225**	1.276	0.280	0.320	**0.196**	1.067	0.264	0.277	**0.191**
Size	6.577	1.625	2.819	**1.007**	5.186	1.656	2.038	**0.870**	3.406	1.200	1.355	**0.751**	2.675	1.061	1.102	**0.700**
Unique Users	4.041	1.454	1.968	**0.690**	3.132	1.348	1.395	**0.600**	1.956	0.960	0.833	**0.521**	1.491	0.825	0.672	**0.468**
Like Score	16.254	4.149	7.586	**1.540**	15.135	3.998	5.782	**1.445**	14.355	3.547	3.614	**1.394**	14.018	3.462	3.221	**1.327**

Table 17: LLM Gemma-3-12b-it comparison with MMG-PopNet on r/Futurology across observation windows using MSE values.

Task	Root Only				30 min				90 min				180 min			
	Zero	Few	Fine-Tune	MMG-PopNet	Zero	Few	Fine-Tune	MMG-PopNet	Zero	Few	Fine-Tune	MMG-PopNet	Zero	Few	Fine-Tune	MMG-PopNet
Max Width	7.837	1.984	2.831	**1.581**	4.027	1.948	1.900	**1.321**	3.058	1.350	1.274	**0.794**	2.680	1.026	1.017	**0.457**
Max Depth	2.439	0.687	1.299	**0.517**	1.311	0.650	0.891	**0.496**	0.974	0.457	0.589	**0.296**	0.766	0.321	0.509	**0.242**
Structural Virality	2.204	0.241	0.429	**0.182**	0.969	0.237	0.296	**0.178**	0.637	0.177	0.196	**0.104**	0.449	0.132	0.170	**0.091**
Size	9.238	3.226	4.869	**2.356**	5.575	2.948	3.305	**1.921**	3.415	1.930	2.202	**1.141**	2.510	1.361	1.816	**0.700**
Unique Users	6.693	2.497	3.470	**1.859**	4.319	2.283	2.576	**1.551**	2.676	1.507	1.731	**0.917**	1.951	1.074	1.406	**0.552**
Like Score	21.116	7.747	12.986	**4.999**	19.077	8.055	13.202	**5.030**	18.979	6.803	12.123	**3.400**	18.766	6.938	10.649	**3.039**

Table 18: LLM GPT-4o-mini comparison with MMG-PopNet on Bluesky across observation windows using MSE values.

Task	Root Only			2			10			20		
	Zero	Few	MMG-PopNet	Zero	Few	MMG-PopNet	Zero	Few	MMG-PopNet	Zero	Few	MMG-PopNet
Max Width	2.742	1.666	**0.666**	2.268	1.562	**0.577**	1.632	1.486	**0.506**	1.386	1.395	**0.459**
Max Depth	1.577	0.590	**0.523**	1.398	0.598	**0.480**	1.055	0.539	**0.428**	0.906	0.484	**0.415**
Structural Virality	0.799	0.275	**0.243**	0.746	0.281	**0.225**	0.614	0.248	**0.196**	0.532	0.223	**0.191**
Size	5.531	1.880	**1.007**	4.545	1.789	**0.870**	3.096	1.464	**0.751**	2.446	1.307	**0.700**
Unique Users	3.472	1.690	**0.690**	2.767	1.530	**0.600**	1.796	1.266	**0.521**	1.392	1.115	**0.468**
Like Score	12.566	4.002	**1.540**	14.130	3.852	**1.445**	13.828	3.478	**1.394**	13.415	3.212	**1.327**

Table 19: LLM GPT-4o-mini comparison with MMG-PopNet on r/Futurology across observation windows using MSE values.

Task	Root Only			30 min			90 min			180 min		
	Zero	Few	MMG-PopNet	Zero	Few	MMG-PopNet	Zero	Few	MMG-PopNet	Zero	Few	MMG-PopNet
Max Width	4.853	2.179	**1.581**	3.388	2.173	**1.321**	2.449	1.813	**0.794**	2.118	1.522	**0.457**
Max Depth	1.476	0.697	**0.517**	1.126	0.623	**0.496**	0.749	0.431	**0.296**	0.551	0.360	**0.242**
Structural Virality	0.847	0.237	**0.182**	0.692	0.221	**0.178**	0.526	0.180	**0.104**	0.405	0.158	**0.091**
Size	8.418	3.451	**2.356**	5.444	3.163	**1.921**	3.226	2.208	**1.141**	2.238	1.612	**0.700**
Unique Users	6.123	2.648	**1.859**	4.056	2.511	**1.551**	2.460	1.758	**0.917**	1.689	1.318	**0.552**
Like Score	20.291	7.341	**4.999**	20.546	7.930	**5.030**	20.385	6.932	**3.400**	19.958	7.070	**3.039**

Zero-Shot Prompt The zero-shot prompt is constructed once per query with no retrieved examples. The placeholder <WINDOW> is replaced by the observation window size, and <EARLY_CONVERSATION_TREE_JSON> is replaced by the serialised conversation tree.

Prompt Template — Zero-Shot

```
You are analyzing a social media conversation tree to predict its
final growth metrics. You will be provided with the first <WINDOW>
 of a social media conversation thread. This conversation thread
continued after this.

Based on the content, timing, and structural patterns of these
early interactions, predict the FINAL state of this conversation
tree when it reaches saturation (that is, stops growing).

METRIC DEFINITIONS:
- max_width: Maximum number of replies at any single depth level.
- max_depth: Maximum depth of the conversation tree.
- structural_virality: Average distance between all pairs of nodes
 (Low = Broadcast, High = Viral).
- num_posts: Total number of posts in the final conversation (root
 + replies).
- num_unique_users: Total number of unique users in the final
conversation.
- root_score: Engagement score of the intial root post.

You will be given a NESTED JSON TREE. Each reply node has text,
time_since_root, time_since_parent, and children. The root node
has text, timestamp, and children.

OUTPUT FORMAT:
Return a single JSON object. Do not include markdown formatting.
{"max_width": <number>, "max_depth": <number>, "
structural_virality": <number>, "num_posts": <number>, "
num_unique_users": <number>, "root_score": <number>}

EARLY CONVERSATION TREE:
<EARLY_CONVERSATION_TREE_JSON>

INSTRUCTIONS:
Using only the conversation tree above as evidence, predict the
FINAL values.
Do NOT include any explanation or extra fields.
Output exactly one JSON object with these keys only: max_width,
max_depth, structural_virality, num_posts, num_unique_users,
root_score.

Your JSON Prediction:
```

Below is a minimal example of the input that populates <EARLY_CONVERSATION_TREE_JSON> when only the root post has been observed.

Example Input — Root-Only Tree

```
{
  "text": "Nvidia partner says it can cut data center energy use by
   50% as AI boom strains power grid",
  "timestamp": "2024-08-27 11:17:18",
  "children": []
}
```

Few-Shot RAG Prompt The few-shot prompt augments the zero-shot template with k retrieved examples (default $k = 4$), selected via embedding-based nearest-neighbour search over the training set. Each placeholder <RETRIEVED_EXAMPLE_TREE_i_JSON> and <RETRIEVED_EXAMPLE_i_TARGETS_JSON> is replaced by the corresponding retrieved tree and its ground-truth final metrics; <TARGET_TREE_JSON> is the query conversation.

Prompt Template — Few-Shot RAG

```
You are analyzing a social media conversation tree to predict its
final growth metrics. You will be provided with the first <WINDOW>
 of a social media conversation thread. This conversation thread
continued after this.

Based on the content, timing, and structural patterns of these
early interactions, predict the FINAL state of this conversation
tree when it reaches saturation (that is, stops growing).

METRIC DEFINITIONS:
- max_width: Maximum number of replies at any single depth level.
- max_depth: Maximum depth of the conversation tree.
- structural_virality: Average distance between all pairs of nodes
 (Low = Broadcast, High = Viral).
- num_posts: Total number of posts in the final conversation (root
 + replies).
- num_unique_users: Total number of unique users in the final
conversation.
- root_score: Engagement score of the intial root post.

Each conversation is represented as a NESTED JSON TREE.

I will show you several examples of early trees and their final
metrics, then ask you to predict for a new tree.

====================
FEW-SHOT EXAMPLES
====================

--- Example 1 ---
Early conversation tree:
<RETRIEVED_EXAMPLE_TREE_1_JSON>
Final metrics:
<RETRIEVED_EXAMPLE_1_TARGETS_JSON>

--- Example 2 ---
Early conversation tree:
<RETRIEVED_EXAMPLE_TREE_2_JSON>
Final metrics:
<RETRIEVED_EXAMPLE_2_TARGETS_JSON>

[... k examples total ...]

====================
NEW CONVERSATION (PREDICT THIS)
====================
Early conversation tree:
<TARGET_TREE_JSON>
Final metrics:
```

G Modality Ablation Details

In Section 5.5, to evaluate the specific contributions of each modality in the MMG-PopNet model, we conducted comprehensive ablation studies. By removing one input source at a time, we measured the relative Mean Squared Error (MSE) increase over the full model. The full MMG-PopNet architecture combines four distinct information sources: text (node-level post text), image (root-post visual content), topology (reply-tree structure), and temporal (node-level timing features).

Implementation Details: To isolate each modality, we modified the model architecture and inputs as follows (w/o means "without"):

- **w/o Text (Semantic Ablation):** and its corresponding projection layer are completely disabled. To maintain architectural compatibility with the downstream Graph Neural Network (GNN), the text embedding for every node is replaced with a zero-initialized vector of the exact same dimensionality. Transformer parameters are frozen and excluded from optimization. This isolates the contribution of actual semantic content. By passing zero-vectors into an otherwise unchanged text feature slot, any performance drop directly reflects the loss of node-level language cues.
- **w/o Image (Visual Ablation):** The CLIP vision encoder is bypassed and all raw image pixels are ignored. However, the graph-level image fusion slot is preserved. Every cascade is instead assigned an identical, learnable "dummy" image embedding that is optimized alongside the rest of the network parameters.
- **w/o Topology (Structural Ablation):** We completely disable the neighbor aggregation mechanism of the bidirectional GraphSAGE network by ignoring the reply edges entirely. Instead of message passing, the network treats the cascade as a disconnected set of independent nodes. To ensure any performance drop is strictly due to the loss of structural connectivity and not a reduction in parameter capacity, the GraphSAGE layers are substituted with standard MLPs. These MLPs are applied independently to each node and strictly match the layer count and hidden feature dimensions of the original GNN. This isolates the value of explicit cascade interactions. Because the replacement MLPs retain the exact same per-node transformation capacity as the original model, this ablation cleanly measures the impact of structural message passing.
- **w/o Temporal (Timing Ablation):** The explicit node-level cascade timing features, specifically time since root and time since parent, are stripped from the input feature matrix before entering the GNN. The nodes are processed using only their text features and topological connections. This specifically targets localized reaction speeds and the pace of the cascade. Global posting metadata may remain, but the precise timing of individual user interactions is entirely removed.

Tables 20,21,22 report the detailed modality-to-target sensitivity analysis across the r/Gaming and r/Futurology datasets. Table 20 aggregates the MSE performance for both datasets to highlight the trade-offs between semantic and structural signals. Table 21 and 22provide the detailed mean and standard deviation across three random seed runs (42, 1042, and 2042) for r/Gaming and r/Futurology, respectively.

The detailed results confirm the main trend in Section 5.5. The full MMG-PopNet model consistently achieves the best average MSE across every target, confirming that each modality contributes valuable predictive information. However, the sensitivity to specific modalities varies distinctly between the two communities

On **r/Gaming**, temporal information is the most influential signal for most structural and participation targets. Temporal information is the most influential signal for structural and participation metrics. Removing temporal features causes the sharpest performance degradation for MAX WIDTH (21.2%), UNIQUE USERS (21.0%), SIZE (18.7%), MAX DEPTH, and STRUCTURAL VIRALITY. This indicates that the pace and timing of early responses are critical for predicting how gaming discussions grow and branch. Conversely, removing image features has the smallest overall effect, indicating that root visual content plays a more modest, complementary role in this specific community.

On **r/Futurology**, the ablation effects are more distributed across modalities. Temporal and topology features remain important for structural targets, but text and image removals also produce visible degradation across several targets. This suggests that r/Futurology popularity prediction depends on a broader mixture of semantic, temporal, and structural signals, rather than being dominated by a single modality.

Table 20: **Modality-to-Target Sensitivity Analysis.** MSE performance of the full MMG-PopNet model versus single-modality ablations across the r/Gaming and r/Futurology datasets. This table isolates the contribution of each modality across the six distinct predictive targets, highlighting the trade-offs between semantic and structural signals. Lower is better. The best value is **bolded**; the second-best is underlined.

Task	Model	r/Gaming				r/Futurology				Avg
		20	50	90	Avg	30	90	180	Avg	
MAX WIDTH	MMG-PopNet	**1.054**	**0.652**	0.509	**0.739**	**1.093**	0.620	0.348	**0.687**	**0.713**
	w/o Text	1.079	0.674	**0.487**	0.747	1.202	**0.617**	**0.319**	0.713	0.730
	w/o Image	1.056	0.653	0.509	**0.739**	1.145	0.650	0.340	0.711	0.725
	w/o Temporal	1.083	1.184	0.641	0.969	1.127	0.711	0.441	0.759	0.864
	w/o Topology	1.061	0.706	0.551	0.773	1.202	0.693	0.377	0.757	0.765
MAX DEPTH	MMG-PopNet	**0.266**	0.196	0.152	**0.205**	**0.413**	**0.271**	**0.202**	**0.295**	**0.250**
	w/o Text	0.269	0.196	**0.150**	**0.205**	0.457	**0.271**	0.205	0.311	0.258
	w/o Image	0.267	**0.194**	0.155	**0.205**	0.422	0.280	0.204	0.302	0.254
	w/o Temporal	0.269	0.296	0.163	0.243	0.418	0.285	0.211	0.305	0.274
	w/o Topology	0.277	0.214	0.192	0.228	0.440	0.295	0.218	0.318	0.273
SIZE	MMG-PopNet	1.272	**0.784**	0.598	**0.885**	**1.703**	**0.949**	0.524	**1.059**	**0.972**
	w/o Text	1.293	0.813	**0.562**	0.889	1.899	0.962	**0.509**	1.123	1.006
	w/o Image	1.272	**0.784**	0.603	0.886	1.783	0.995	0.520	1.099	0.993
	w/o Temporal	1.302	1.391	0.745	1.146	1.749	1.086	0.649	1.161	1.154
	w/o Topology	**1.267**	0.807	0.644	0.906	1.854	1.050	0.548	1.151	1.028
STRUCTURAL VIRALITY	MMG-PopNet	**0.076**	0.057	0.045	**0.059**	**0.143**	**0.097**	**0.072**	**0.104**	**0.082**
	w/o Text	0.078	0.058	0.045	0.060	0.158	0.098	**0.072**	0.109	0.085
	w/o Image	0.077	**0.056**	**0.044**	**0.059**	0.147	0.100	**0.072**	0.106	0.083
	w/o Temporal	**0.076**	0.117	0.045	0.079	0.146	0.102	0.073	0.107	0.093
	w/o Topology	0.082	0.068	0.061	0.070	0.152	0.107	0.078	0.112	0.091
UNIQUE USERS	MMG-PopNet	1.170	**0.730**	0.556	**0.819**	**1.334**	**0.747**	0.407	**0.829**	**0.824**
	w/o Text	1.197	0.764	**0.525**	0.829	1.485	0.756	**0.392**	0.878	0.853
	w/o Image	1.172	0.735	0.561	0.822	1.402	0.789	0.406	0.866	0.844
	w/o Temporal	1.198	1.331	0.697	1.075	1.380	0.863	0.516	0.920	0.997
	w/o Topology	**1.166**	0.771	0.597	0.845	1.467	0.830	0.436	0.911	0.878
LIKE SCORE	MMG-PopNet	**4.659**	**4.237**	**4.003**	**4.300**	**4.527**	**3.032**	**2.632**	**3.397**	**3.848**
	w/o Text	5.215	5.016	4.397	4.876	5.347	3.680	3.237	4.088	4.482
	w/o Image	4.714	4.327	4.023	4.355	4.783	3.271	2.674	3.576	3.965
	w/o Temporal	4.726	5.041	4.284	4.683	4.597	3.454	3.166	3.739	4.211
	w/o Topology	4.702	4.348	4.113	4.387	4.812	3.292	2.728	3.611	3.999

Across both datasets, text is consistently the most important modality for LIKE SCORE. Removing textual semantics increases the average LIKE SCORE MSE from 4.300 to 4.876 on r/Gaming and from 3.397 to 4.088 on r/Futurology. These results are consistent with the text-centered nature of the dataset platforms of the benchmark., where discussion content is primarily text based. This suggests a broader hypothesis that the dominant modality may shift with platform design. On platforms where images or videos are the primary medium of interaction, visual features may play a larger role in predicting engagement and cascade growth.

Table 21: **Modality-to-Target Sensitivity Analysis on r/Gaming.** MSE performance of full MMG-PopNet model versus single-modality ablations on the r/Gaming dataset. Results are averaged across three random seeds: 42, 1042, and 2042, and are reported as mean ± standard deviation. Lower is better. The best value is **bolded**; the second-best is underlined.

Task	Model	20	50	90	Avg
MAX WIDTH	MMG-PopNet	**1.054 ± 0.008**	**0.652 ± 0.013**	0.509 ± 0.014	**0.739 ± 0.009**
	w/o Text	1.079 ± 0.015	0.674 ± 0.003	**0.487 ± 0.030**	0.747 ± 0.013
	w/o Image	1.056 ± 0.012	0.653 ± 0.006	0.509 ± 0.022	**0.739 ± 0.006**
	w/o Temporal	1.083 ± 0.008	1.184 ± 0.708	0.641 ± 0.007	0.969 ± 0.234
	w/o Topology	1.061 ± 0.008	0.706 ± 0.004	0.551 ± 0.012	0.773 ± 0.007
MAX DEPTH	MMG-PopNet	**0.266 ± 0.003**	0.196 ± 0.003	0.152 ± 0.006	**0.205 ± 0.001**
	w/o Text	0.269 ± 0.002	0.196 ± 0.003	**0.150 ± 0.001**	**0.205 ± 0.000**
	w/o Image	0.267 ± 0.004	**0.194 ± 0.003**	0.155 ± 0.005	**0.205 ± 0.002**
	w/o Temporal	0.269 ± 0.002	0.296 ± 0.153	0.163 ± 0.002	0.243 ± 0.052
	w/o Topology	0.277 ± 0.003	0.214 ± 0.007	0.192 ± 0.005	0.228 ± 0.001
SIZE	MMG-PopNet	1.272 ± 0.012	**0.784 ± 0.004**	0.598 ± 0.023	**0.885 ± 0.010**
	w/o Text	1.293 ± 0.016	0.813 ± 0.008	**0.562 ± 0.028**	0.889 ± 0.012
	w/o Image	1.272 ± 0.014	**0.784 ± 0.005**	0.603 ± 0.020	0.886 ± 0.004
	w/o Temporal	1.302 ± 0.014	1.391 ± 0.820	0.745 ± 0.004	1.146 ± 0.277
	w/o Topology	**1.267 ± 0.013**	0.807 ± 0.011	0.644 ± 0.022	0.906 ± 0.011
STRUCTURAL VIRALITY	MMG-PopNet	**0.076 ± 0.003**	0.057 ± 0.000	0.045 ± 0.003	**0.059 ± 0.001**
	w/o Text	0.078 ± 0.001	0.058 ± 0.002	0.045 ± 0.003	0.060 ± 0.001
	w/o Image	0.077 ± 0.003	**0.056 ± 0.002**	**0.044 ± 0.003**	**0.059 ± 0.002**
	w/o Temporal	**0.076 ± 0.001**	0.117 ± 0.096	0.045 ± 0.001	0.079 ± 0.032
	w/o Topology	0.082 ± 0.004	0.068 ± 0.004	0.061 ± 0.003	0.070 ± 0.001
UNIQUE USERS	MMG-PopNet	1.170 ± 0.010	**0.730 ± 0.011**	0.556 ± 0.019	**0.819 ± 0.010**
	w/o Text	1.197 ± 0.016	0.764 ± 0.005	**0.525 ± 0.024**	0.829 ± 0.011
	w/o Image	1.172 ± 0.016	0.735 ± 0.006	0.561 ± 0.017	0.822 ± 0.003
	w/o Temporal	1.198 ± 0.007	1.331 ± 0.820	0.697 ± 0.007	1.075 ± 0.273
	w/o Topology	**1.166 ± 0.012**	0.771 ± 0.006	0.597 ± 0.016	0.845 ± 0.008
LIKE SCORE	MMG-PopNet	**4.659 ± 0.023**	**4.237 ± 0.138**	**4.003 ± 0.048**	**4.300 ± 0.064**
	w/o Text	5.215 ± 0.017	5.016 ± 0.038	4.397 ± 0.009	4.876 ± 0.013
	w/o Image	4.714 ± 0.028	4.327 ± 0.136	4.023 ± 0.055	4.355 ± 0.044
	w/o Temporal	4.726 ± 0.020	5.041 ± 0.988	4.284 ± 0.023	4.683 ± 0.335
	w/o Topology	4.702 ± 0.050	4.348 ± 0.051	4.113 ± 0.042	4.387 ± 0.047

Table 22: **Modality-to-Target Sensitivity Analysis on r/Futurology.** MSE performance of full MMG-PopNet model versus single-modality ablations on the r/Futurology dataset. Results are averaged across three random seeds: 42, 1042, and 2042, and are reported as mean ± standard deviation. Lower is better. The best value is **bolded**; the second-best is underlined.

Task	Model	30	90	180	Avg
MAX WIDTH	MMG-PopNet	**1.093 ± 0.013**	0.620 ± 0.020	0.348 ± 0.008	**0.687 ± 0.009**
	w/o Text	1.202 ± 0.020	**0.617 ± 0.005**	**0.319 ± 0.016**	0.713 ± 0.011
	w/o Image	1.145 ± 0.015	0.650 ± 0.010	0.340 ± 0.008	0.711 ± 0.007
	w/o Temporal	1.127 ± 0.022	0.711 ± 0.012	0.441 ± 0.013	0.759 ± 0.005
	w/o Topology	1.202 ± 0.009	0.693 ± 0.011	0.377 ± 0.022	0.757 ± 0.005
MAX DEPTH	MMG-PopNet	**0.413 ± 0.001**	**0.271 ± 0.003**	**0.202 ± 0.005**	**0.295 ± 0.003**
	w/o Text	0.457 ± 0.012	**0.271 ± 0.002**	0.205 ± 0.001	0.311 ± 0.004
	w/o Image	0.422 ± 0.005	0.280 ± 0.006	0.204 ± 0.002	0.302 ± 0.004
	w/o Temporal	0.418 ± 0.009	0.285 ± 0.006	0.211 ± 0.004	0.305 ± 0.004
	w/o Topology	0.440 ± 0.005	0.295 ± 0.007	0.218 ± 0.003	0.318 ± 0.003
SIZE	MMG-PopNet	**1.703 ± 0.015**	**0.949 ± 0.017**	0.524 ± 0.010	**1.059 ± 0.012**
	w/o Text	1.899 ± 0.055	0.962 ± 0.005	**0.509 ± 0.013**	1.123 ± 0.021
	w/o Image	1.783 ± 0.022	0.995 ± 0.019	0.520 ± 0.006	1.099 ± 0.010
	w/o Temporal	1.749 ± 0.039	1.086 ± 0.022	0.649 ± 0.006	1.161 ± 0.009
	w/o Topology	1.854 ± 0.025	1.050 ± 0.014	0.548 ± 0.019	1.151 ± 0.007
STRUCTURAL VIRALITY	MMG-PopNet	**0.143 ± 0.002**	**0.097 ± 0.001**	**0.072 ± 0.002**	**0.104 ± 0.000**
	w/o Text	0.158 ± 0.003	0.098 ± 0.002	**0.072 ± 0.001**	0.109 ± 0.001
	w/o Image	0.147 ± 0.002	0.100 ± 0.002	**0.072 ± 0.001**	0.106 ± 0.000
	w/o Temporal	0.146 ± 0.005	0.102 ± 0.003	0.073 ± 0.002	0.107 ± 0.002
	w/o Topology	0.152 ± 0.005	0.107 ± 0.003	0.078 ± 0.001	0.112 ± 0.001
UNIQUE USERS	MMG-PopNet	**1.334 ± 0.012**	**0.747 ± 0.014**	0.407 ± 0.012	**0.829 ± 0.010**
	w/o Text	1.485 ± 0.040	0.756 ± 0.004	**0.392 ± 0.010**	0.878 ± 0.015
	w/o Image	1.402 ± 0.017	0.789 ± 0.017	0.406 ± 0.004	0.866 ± 0.009
	w/o Temporal	1.380 ± 0.030	0.863 ± 0.009	0.516 ± 0.006	0.920 ± 0.007
	w/o Topology	1.467 ± 0.021	0.830 ± 0.013	0.436 ± 0.019	0.911 ± 0.001
LIKE SCORE	MMG-PopNet	**4.527 ± 0.003**	**3.032 ± 0.039**	**2.632 ± 0.096**	**3.397 ± 0.035**
	w/o Text	5.347 ± 0.050	3.680 ± 0.038	3.237 ± 0.029	4.088 ± 0.035
	w/o Image	4.783 ± 0.066	3.271 ± 0.007	2.674 ± 0.074	3.576 ± 0.024
	w/o Temporal	4.597 ± 0.080	3.454 ± 0.075	3.166 ± 0.015	3.739 ± 0.003
	w/o Topology	4.812 ± 0.055	3.292 ± 0.055	2.728 ± 0.097	3.611 ± 0.004

H Qualitative Case Study

This qualitative case study examines MMG-PopNet predictions for the r/Futurology dataset under a 180-minute observation window, focusing on the SIZE target, which measures the final number of nodes in a social cascade. The scatter plot compares predicted SIZE against actual SIZE. Both axes are shown in log scale and are displayed using powers of ten, such as 10^1, 10^2, and 10^3, to make the heavy-tailed popularity distribution easier to interpret. In this view, points near the red dashed diagonal indicate more accurate predictions, while points above or below the diagonal indicate over-prediction or under-prediction.

For the qualitative analysis, several cascades are selected from different regions of the scatter plot to provide a wide range of examples, including accurate predictions, over-predictions, and under-predictions. These selected examples are shown as colored nodes on the scatter plot. For each selected case, interpretability is applied to the root post content. Specifically, GradSAM [75] token explanations are used to highlight influential text tokens, and Grad-CAM [76, 77] image explanations are used to highlight influential image regions. This helps provide intuition about how the root post's text and image content may have contributed to the predicted cascade SIZE, while also showing that final popularity depends on additional temporal and interaction dynamics captured by the full model.

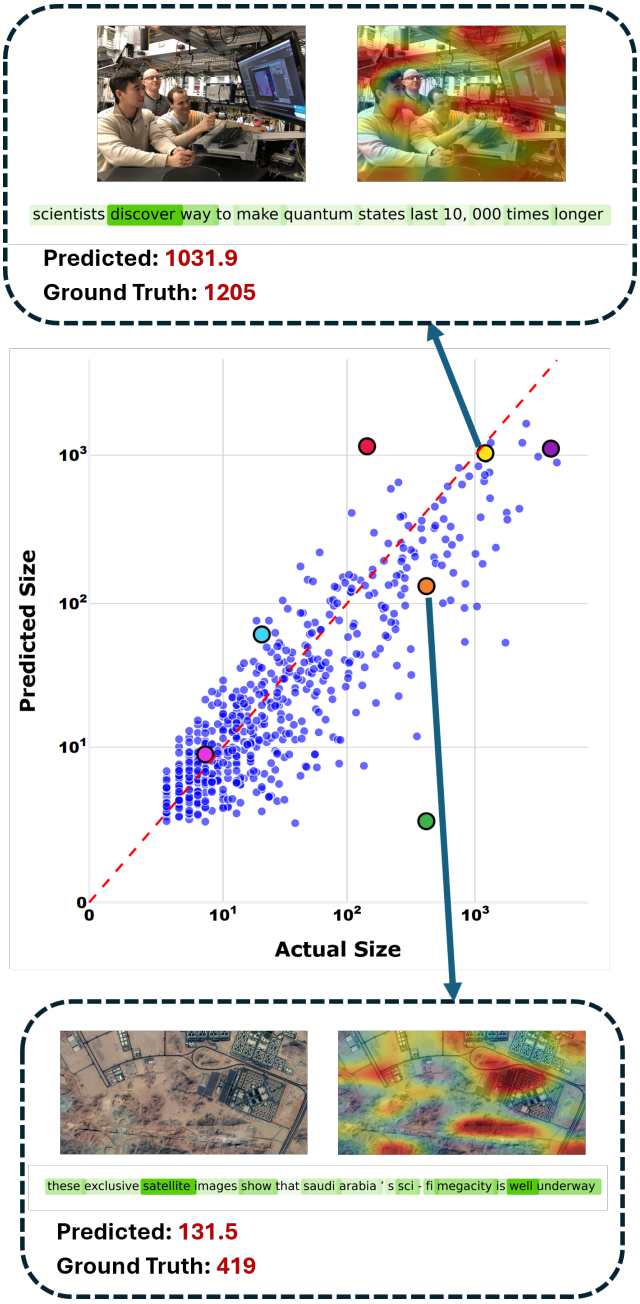


Figure 7: The top example shows a relatively accurate high-popularity prediction, where the predicted SIZE is close to the actual SIZE. GradSAM highlights root-post tokens such as "discover", "quantum states", and "longer" suggesting that the model attends to scientific novelty and breakthrough-oriented language. Grad-CAM emphasizes parts of the laboratory image, including the people and equipment, which may provide visual cues of scientific credibility. The bottom example shows an under-predicted cascade, where the actual SIZE is much larger than the predicted SIZE. GradSAM highlights tokens such as "exclusive," "satellite," "megacity," and "well underway," while Grad-CAM focuses on several regions of the satellite image. This case suggests that although the root post contains visually and textually salient signals, the model may underestimate posts whose later popularity is driven by broader public interest in large-scale infrastructure or geopolitical topics.

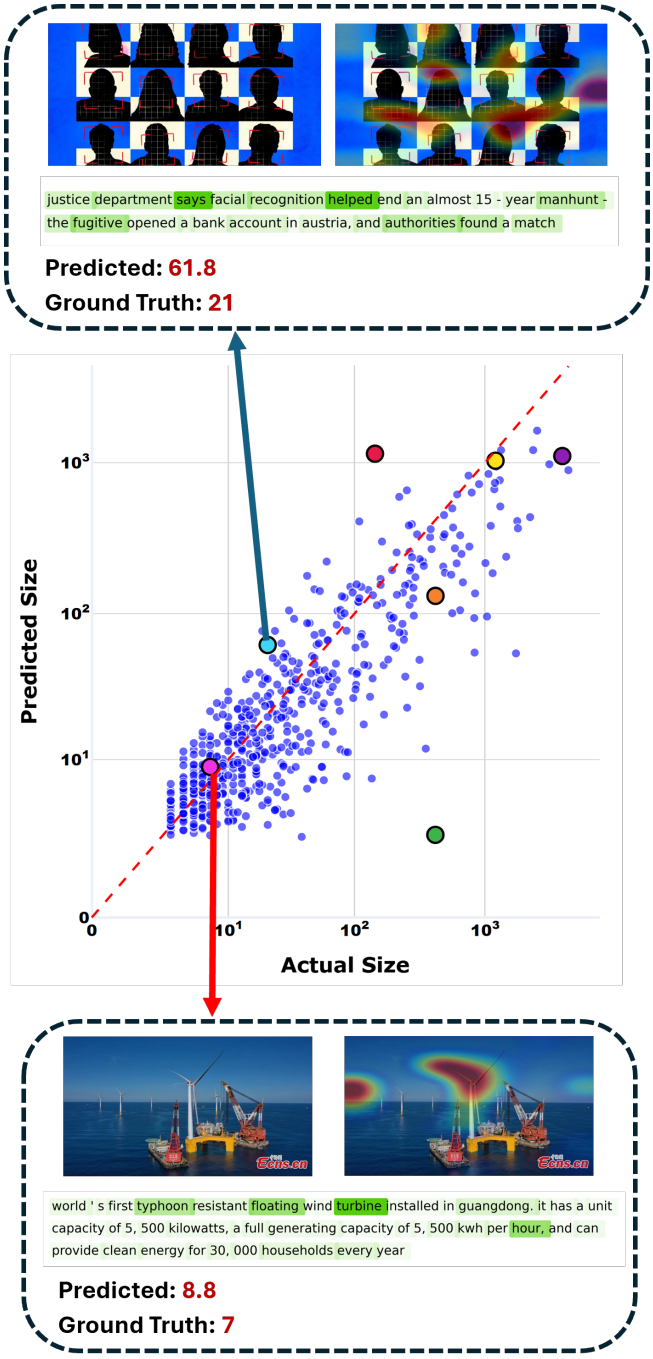


Figure 8: The top example shows an over-predicted cascade, where the predicted SIZE is larger than the actual SIZE. GradSAM highlights root-post tokens such as "says," "facial recognition," "helped," "fugitive," and "match," suggesting that the model attends to crime, surveillance, and authority-related language. Grad-CAM highlights multiple regions across the facial-recognition image, which may reinforce the post's technology and public-safety framing. The bottom example shows a relatively accurate low-popularity prediction, where the predicted SIZE is close to the actual SIZE. GradSAM highlights tokens such as "floating," "wind turbine," and "hour," while Grad-CAM focuses on parts of the offshore turbine and surrounding scene. This comparison suggests that root-post content can provide meaningful cues for SIZE prediction, but attention to salient technology-related terms does not always translate into high cascade growth.

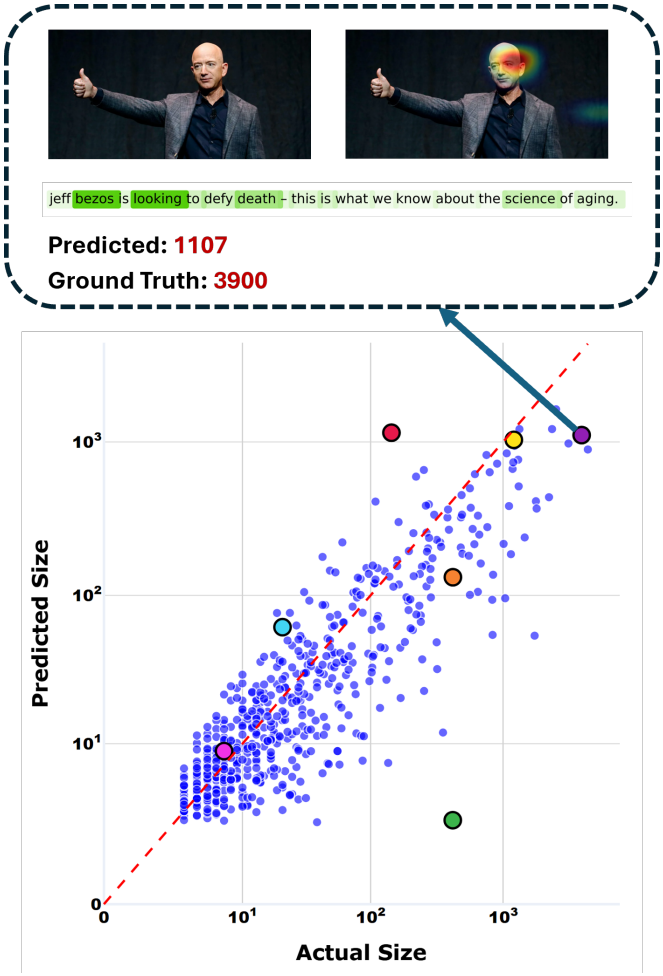


Figure 9: The highlighted example shows an under-predicted cascade, where the actual SIZE is much larger than the predicted SIZE. GradSAM highlights root-post tokens such as "bezos," looking," death," and science of aging," suggesting that the model attends to the named entity and the longevity-related framing of the post. Grad-CAM focuses strongly on the face of Jeff Bezos, indicating that the visual explanation is concentrated on him in the image. This case suggests that the root post contains salient celebrity and science-related cues, but the model still underestimates the eventual discussion volume, possibly because later cascade growth is driven by broader public debate around wealth, longevity, and aging beyond the root content alone.

I Limitations.

Despite the unified design of MMG-Pop and the strong empirical performance of MMG-PopNet, this work has several limitations.

First, the benchmark is constructed from Bluesky and Reddit, which provide diverse but still incomplete coverage of social media ecosystems. Platform-specific moderation policies, recommendation algorithms, user demographics, and interaction norms can substantially affect popularity dynamics. Therefore, conclusions drawn from these datasets may not fully generalize to platforms such as X/Twitter, TikTok, Instagram, YouTube, or private messaging communities.

Second, our formulation represents social cascades primarily as tree-structured reply or interaction graphs. This abstraction captures explicit propagation paths, but it may omit broader network exposure effects, algorithmic ranking effects, cross-platform diffusion (e.g., getting high engagement on one social platform due to a viral event on second social platform), and unobserved impressions. A post may become popular not only because of its visible reply tree, but also because of recommendation systems, external sharing, creator reputation, or coordinated amplification that is not directly observable in the collected cascade.

Third, although MMG-PopNet jointly models text, image, temporal, and structural signals, the available modalities are uneven across platforms and communities. Root visual content contributes modestly in our experiments, but this may partly reflect dataset composition rather than the intrinsic value of visual signals. Similarly, richer video, audio and user-history features are not fully modeled. Future extensions should consider broader media types and more complete user-context features while carefully protecting user privacy.

Fourth, the work of popularity prediction has the potential of being misused by nefarious parties to identify the signals that provide them the highest social engagement to spread hateful or toxic messages or content on the social media. So, one has to be careful and mindful in using such techniques to ensure well being of all.