IJSCIA

International Journal of Scientific Advances
ISSN: 2708-7972
Volume: 7 | Issue: 6 | Nov – Dec 2026 Available Online: www.ijscia.com
DOI: 10.51542/ijscia.v7i6.12

Stage-Aware Multimodal Popularity Prediction for TikTok Short Videos: Leakage-Safe Features and Modality Attribution

Md Nadim Hosain, Md Soyaeb Bin Kuddus, Md Emadur Rahman Ekra*, Md Daud Ibrahim

School of Electronics and Information Engineering,
Nanjing University of Information Science and Technology, Nanjing, China

E-mail: mdnadimhosain@nuist.edu.cn; mdemandekra@nuist.edu.cn; kuddusmdsoyaebbin@nuist.edu.cn; ibrahimmddaud@nuist.edu.cn

*Corresponding author details: Md Emadur Rahman Ekra; mdemandekra@nuist.edu.cn

ABSTRACT

Popularity prediction for short videos supports content moderation, creator analytics, and platform research, yet many models mix information from different moments of a video's life or evaluate on random splits that leak future information into training. In this study, we develop a stage-aware, leakage-safe multimodal framework for TikTok and evaluate it on 11,884 public TikTok posts from 1,606 creators (June–November 2024) with daily engagement snapshots and day-30 view counts. We build descriptors of creator influence, audio and sound, visual and technical attributes, semantic content, interaction structure, and temporal view dynamics from records available at prediction time, train one gradient-boosted tree model per horizon (cold start and days 1, 2, 3, and 5), and test it on the most recent five weeks of posts. Before any engagement is observed, the model identifies future top-20% posts with an area under the receiver operating characteristic curve (ROC-AUC) of 0.942 (follower-count heuristic: 0.882). After one day, it reaches 0.982 for this task and 0.864 for predicting sustained growth, where the Szabo–Huberman log-linear baseline scores 0.677. An exact Shapley decomposition of test ROC-AUC shows that creator influence carries 68% of the cold-start signal and leads day-1 growth prediction, whereas temporal view dynamics lead from day 2 onward; metadata-level content descriptors add little once the creator is known. Feature-level fusion outperforms decision-level fusion during the first days (day-1 growth ROC-AUC 0.864 vs. 0.843). The framework and protocol provide a reproducible reference for leakage-safe TikTok popularity research.

Keywords: TikTok; short-video popularity prediction; multimodal features; gradient-boosted decision trees; temporal data leakage; Shapley value attribution

1 INTRODUCTION

Short-video platforms now shape how people find news, entertainment and advice, and TikTok is one of the most widely studied of them outside China. Its For You feed distributes videos beyond a creator's followers, so a post from a small account can reach millions of viewers within hours while a post from a large account can stall. This makes popularity both hard and valuable to anticipate. Moderators need early warning about content that is about to spread, creators and brands want to know which posts merit promotion, and researchers use popularity forecasts to study attention dynamics and algorithmic amplification [1-4].

Two research traditions address this problem. The first, which goes back to Szabo and Huberman [5] and Pinto et al. [6], extrapolates final popularity from early view counts and shows that the logarithms of early and late popularity are strongly correlated. The second extracts content features from video frames, audio, captions, and user metadata and learns their relation to popularity with multimodal neural networks [7-13]. A recent Douyin study by Mi et al. [14] combined both ideas: it assembled user and author influence, audio and video quality, content semantics, and comment similarity into a feature system, sampled engagement every 15 minutes, and classified videos with a random forest.

Three problems limit what such studies tell us about TikTok. First, popularity models often use information that would not exist at the moment of prediction, such as creator statistics recorded after publication, later engagement snapshots, or label thresholds estimated on the full dataset. In addition, random splits place videos from the same weeks and creators on both sides of the split. Such leakage inflates reported accuracy and is a recognised cause of irreproducible machine-learning results [15]. Second, most studies report one prediction setting, so they do not show how the value of each modality changes as a video move from cold start to growth and decay.

Third, platform differences are often assumed away. Douyin and TikTok share a parent company but serve different markets, apply different policies, and expose different data, so findings on one platform cannot simply be transferred to the other [3, 16].

In this study, we address these problems with a stage-aware, leakage-safe framework for TikTok popularity prediction. We use a public release of TikTok posts that records, for every post, daily engagement snapshots up to day 5, daily snapshots of the creator account, and the cumulative view count on day 30 [17]. This structure allows us to define prediction horizons explicitly and to build every feature from information available at that horizon. Our contributions are as follows.

(i) We formulate TikTok popularity prediction at five observation horizons, from cold start to day 5, and for two targets: whether a post will reach the top 20% of day-30 views, and whether it will keep growing faster than typical after the observation horizon. The second target removes most of the information carried by the current view count and therefore tests whether a model understands growth rather than reading off accumulated popularity.

(ii) We design a six-modality feature system for TikTok (creator influence, audio and sound, visual and technical attributes, semantic content, interaction structure and temporal dynamics) in which creator statistics enter through as-of joins, creator track records and sound reuse are counted causally, and engagement dynamics are computed only from snapshots up to the horizon.

(iii) We evaluate gradient-boosted trees against classical early-popularity models and against linear, neural, bagging and decision-level fusion alternatives under a chronological split, with bootstrap confidence intervals, paired significance tests, creator-disjoint validation and threshold sensitivity analysis.

(iv) We decompose the test area under the receiver operating characteristic curve (ROC-AUC) exactly into Shapley contributions of the modalities at every horizon, which shows how the drivers of predictability shift across the TikTok video life cycle.

2 RELATED WORK

2.1 Early-Engagement Popularity Models

Szabo and Huberman [5] showed on Digg and YouTube that the logarithm of a post's popularity at a later time is approximately a linear function of its logarithm at an early time, so that a single regression predicts long-term popularity from early counts. Pinto et al. [6] generalised this idea to a multivariate linear model over the daily view history and reported lower errors for YouTube videos. Both models remain strong baselines because early counts summarise the outcome of the platform's own exposure decisions. Chen et al. [18] analysed 248 million Kuaishou videos and found that short-form platforms concentrate attention more sharply than long-form ones and that ordinary users have a comparatively high chance of producing a hit. This observation suggests that creator size alone should explain less on short-video platforms than on traditional video sites.

2.2 Multimodal Short-Video Popularity Prediction

Multimodal approaches learn popularity from the content itself. Cho et al. [8] combined full-frame video representations with other modalities through BiLSTM self- and co-attention to predict the popularity of YouTube Shorts, Zhong et al. [10] augmented predictions with similar historical micro-videos retrieved from a multimodal memory bank, and Hsu et al. [9] adapted vision–language features and measured inconsistency between modalities. Lu et al. [11] combined visual, acoustic, and large language model (LLM)-generated textual descriptions with metadata for micro-video popularity, and Ye et al. [12] won the 2025 Social Media Prediction (SMP) video track by feeding pretrained video features, user metadata, and contextual information into gradient-boosted regression. Engagement prediction has also moved beyond view counts: Li et al. [7] released a large Snapchat dataset with watch-percentage and engagement-continuation targets, and Sun et al. [13] showed that large multimodal models that also hear the background audio improve such predictions. The Douyin study of Mi et al. [14] is the closest to ours in scope. On 1,127 videos, it defined popularity as the growth of interactions between the midpoint and the end of a 12.5-hour observation window, split at the median, and reported an F1-score of 69.3% and an AUC of 71.3% with a random forest.

TABLE 1: Representative Studies on Short-Video and Social-Media Popularity Prediction.

Study	Platform	Modalities/Inputs	Target	Temporal control
Szabo & Huberman [5]	Digg, YouTube	early counts	later popularity	early → late
Pinto et al. [6]	YouTube	daily view history	views at target day	early → late
Ling et al. [2]	TikTok	content, creator attributes	viral vs. non-viral	none (descriptive)
Cho et al. (AMPS) [8]	YouTube Shorts	frames + other modalities	popularity	not reported
Zhong et al. (MMRA) [10]	micro-video	visual, text, retrieval	popularity score	not reported
Li et al. [7]	Snapchat	video, audio, text, metadata	watch percentage; engagement continuation	not reported
Ye et al. (MVP) [12]	SMP-Video	video features, user metadata	popularity	challenge split
Mi et al. [14]	Douyin	user, A/V quality, content, comments	interaction growth	15-min sampling, 80/20 split
Paekivi & Karjus [19]	TikTok (brands)	LLM-coded content, followers	likes, comments, shares	cross-sectional
This study	**TikTok**	**six modalities, daily snapshots**	**top-20% views; sustained growth**	**as-of features, chronological split**

2.3 Studies of TikTok

TikTok-specific research has focused on how the platform distributes attention. Klug et al. [1] documented users' beliefs that engagement, posting time, and hashtags drive recommendation, and found that engagement and posting time were indeed associated with a higher chance of being recommended. Ling et al. [2] compared viral and non-viral TikTok videos and found that the creator's follower count was the strongest indicator of virality, followed by content attributes such as close-up and medium shots, on-screen text, and point of view. Guinaudeau et al. [3] showed that TikTok views depend less on follower counts than YouTube views do, which is consistent with the For You page's reach beyond followers. Zannettou et al. [4] analysed donated viewing histories and found that users engage more with videos from creators they follow than with algorithmically recommended content. Corso et al. [16] audited the official TikTok Research API and raised caveats about the geographic coverage of its samples. Most recently, Paekivi and Karjus [19] coded about ten thousand brand and organisational TikTok videos with a large language model and found that content variables add only a modest share of explained variance beyond follower count and video age.

2.4 Research Gap

Table 1 summarises representative studies. Three gaps stand out. Few studies control temporal leakage explicitly: features, label thresholds, and normalisation statistics are rarely restricted to information available at the prediction time, and random splits are common. No study we found compares the contribution of different modalities across several observation horizons of the same videos. Finally, work that targets TikTok tends either to describe virality without forecasting it [1-4] or to model a narrow subset of accounts [19]. Our study addresses these gaps with an explicit horizon design, leakage-safe features, chronological evaluation, and an exact attribution of predictive performance to modalities.

3 PROBLEM FORMULATION

Let post i be published at time $t_{0,i}$ by creator a_i. Its cumulative view count k days after publication is $V_i(k)$; the release provides daily snapshots for $k = 0, \ldots, 5$ and the count $V_i(30)$ on day 30. An *observation horizon* $\tau \in \{C, 1, 2, 3, 5\}$ defines what a predictor may see: at cold start ($\tau = C$) only information available when the post is published, and at $\tau = k$ additionally the engagement snapshots up to day k. We study two classification targets and one regression target.

3.1 Popularity

A post is popular when its day-30 views reach the 80th percentile $q_{0.8}$ of the training posts:

$$y_i^{\text{pop}} = \mathbb{I}[V_i(30) \geq q_{0.8}] \quad (1)$$

We use the top 20% rather than a median split because the view distribution is extremely skewed (Section 6.3), and the upper tail is what moderators and marketers care about. Section 8.7 reports results for the top 10% and the top 50% as well.

3.2 Sustained Growth

For $\tau \geq 1$, the log-growth of a post after the horizon is

$$G_i(\tau) = \ln\big(1 + V_i(30)\big) - \ln\big(1 + V_i(\tau)\big) \quad (2)$$

and the growth label compares it with the training median $\tilde{G}(\tau)$:

$$y_i^{\mathrm{gr}}(\tau) = \mathbb{I}\left[G_i(\tau) \geq \tilde{G}(\tau)\right] \quad (3)$$

This target asks whether a post will keep gaining views faster than a typical post that has reached the same horizon. Because $V_i(\tau)$ enters with a negative sign, a predictor cannot succeed by reading off the accumulated popularity, which makes this task considerably harder than the first. It is the TikTok analogue of the interaction-growth target of Mi et al. [14], defined here on views and on a 30-day rather than a 12.5-hour window.

3.3 Regression

We also predict $\ln(1 + V_i(30))$ directly and report errors on the logarithmic scale.

Every quantity that depends on the data distribution, including $q_{0.8}$, $\tilde{G}(\tau)$, imputation values, scaling constants, hyper-parameters and decision thresholds, is estimated on the training period only. A feature is admissible at horizon τ only if it can be computed from records time-stamped no later than $t_{0,i}$ + τ days. We call a design that meets these conditions *leakage-safe*.

4 MULTIMODAL FEATURE ENGINEERING

We organise the descriptors into six modalities (Figure 1, Table 2). The first four describe the post and its creator at publication and are available at every horizon. The last two summarise post-publication engagement and exist only for $\tau \geq 1$. Count variables are log-transformed with $\ln(1 + x)$ to compress their heavy tails.

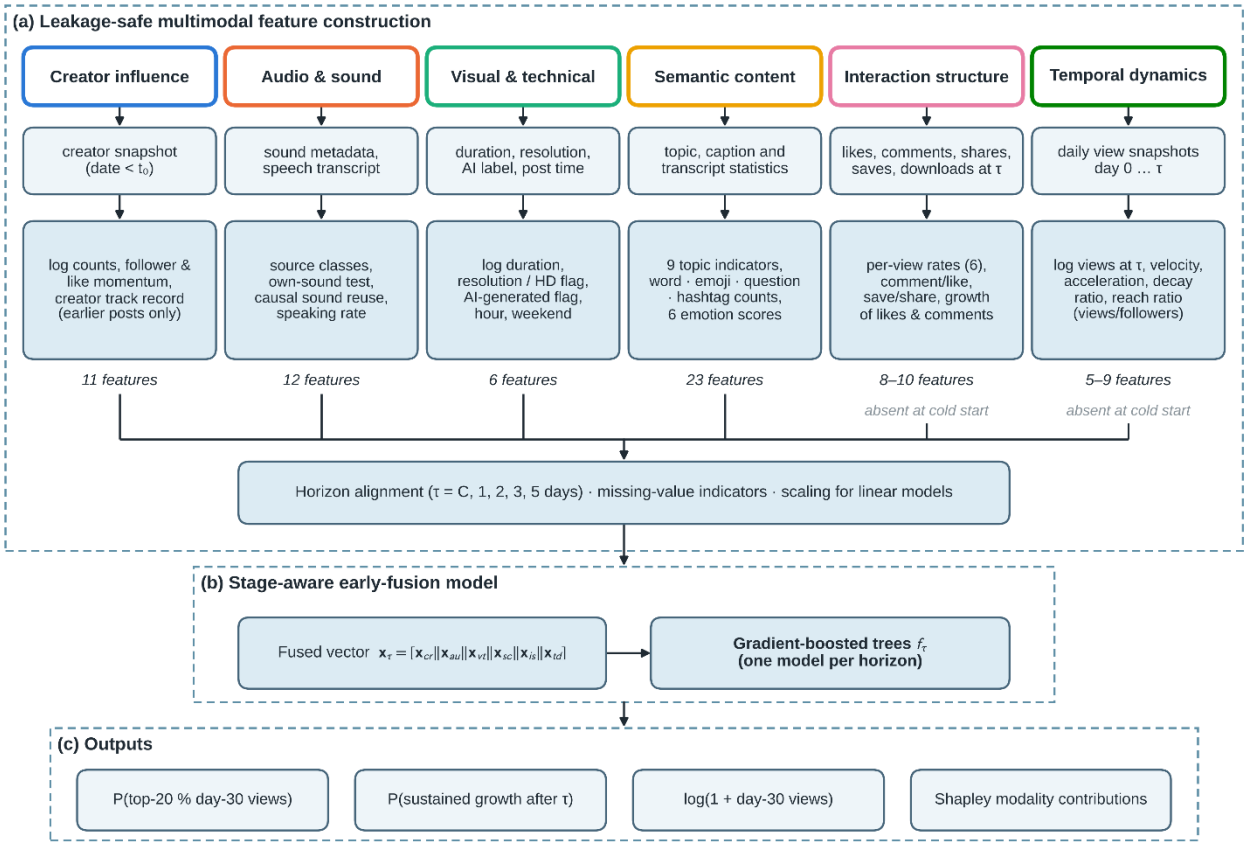


FIGURE 1: Stage-Aware Multimodal Popularity Prediction Framework. (a) Six modality blocks are built from records available at the prediction time; interaction-structure and temporal-dynamics blocks exist only after publication. (b) The blocks are concatenated, and a gradient-boosted tree model is trained for each horizon. (c) The models output popularity and growth probabilities, a log-view estimate, and Shapley modality contributions.

4.1 Creator Influence

The release provides one snapshot of each creator account per day. To keep the creator description free of the post's own early engagement, we join each post to the latest snapshot dated strictly before its publication day,

$$d_i^* = \max\{d \in \mathcal{D}_{a_i} : d < \mathrm{date}(t_{0,i})\} \quad (4)$$

Where $\mathcal{D}_a$ is the set of snapshot dates of creator a. From this snapshot we take the follower count F, following count, total likes received L, and number of posted videos N, and derive the follower-to-following ratio and the average likes per video. Creator momentum captures accounts that are growing at publication time:

$$m_i^F = \frac{7}{\Delta d}\left[\ln\left(1 + F_{a_i}(d_i^*)\right) - \ln\left(1 + F_{a_i}(d_i^* - 7)\right)\right] (5)$$

With an analogous term for total likes, where Δd is the number of days between the two snapshots. We also describe the creator's track record with the mean log day-30 views of the creator's earlier posts in the pool whose 30-day outcome was already known,

$$h_i = \frac{1}{|\mathcal{H}_i|} \sum_{j \in \mathcal{H}_i} \ln\left(1 + V_j(30)\right) \quad (6)$$

$$\mathcal{H}_i = \{j : a_j = a_i,\ t_{0,j} < t_{0,i} - 30\,\mathrm{d}\} \quad (7)$$

Together with the mean log day-1 views of earlier posts at least two days old and the number of earlier posts. For interpretability and for linear models, we also combine the two creator signals that matter most into a creator influence index (CII):

$$\mathrm{CII}_i(\alpha) = \alpha\, z_i^F + (1 - \alpha)\, z_i^L \quad (8)$$

Where z_i^F and z_i^L are the values of $\ln(1 + F_i)$ and $\ln(1 + L_i/N_i)$ standardised with training-set statistics. The Douyin study [14] weighted follower and like counts for authors by a grid search on its evaluation metric and fixed the weights of user features by the analytic hierarchy process; we select α on the training period only (Section 8.4).

4.2 Audio and Sound

Sound is central to TikTok: creators either record original audio or attach a track from the platform's library, and popular sounds propagate across many videos.

The release records how the sound was chosen (40 interface sources such as "original", "single_song", "edit_page_recommend" or "duet"), the sound identifier, and the identifier of the sound's owner. Figure 2 shows how we turn these fields into 12 descriptors. We map the 40 sources to eight classes (original, library song, recommended, search, favourites, remix, template and other), flag posts whose sound belongs to the posting creator, and measure the momentum of a sound by counting earlier posts that used it:

$$R_i = \ln\left(1 + \left|\{j : s_j = s_i,\ t_{0,j} < t_{0,i}\}\right|\right) \quad (9)$$

Where s_i is the sound identifier. The count is causal: it only looks backwards in time. Two speech descriptors complete the block: the speaking rate supplied with the release (words per second of speech) and an indicator of whether the post contains speech at all. The release contains no waveforms, so perceptual audio quality, for example a NISQA mean opinion score [20], cannot be computed. Figure 2(c) shows where such a module would enter the framework; we treat it as future work and do not report any audio-quality result.

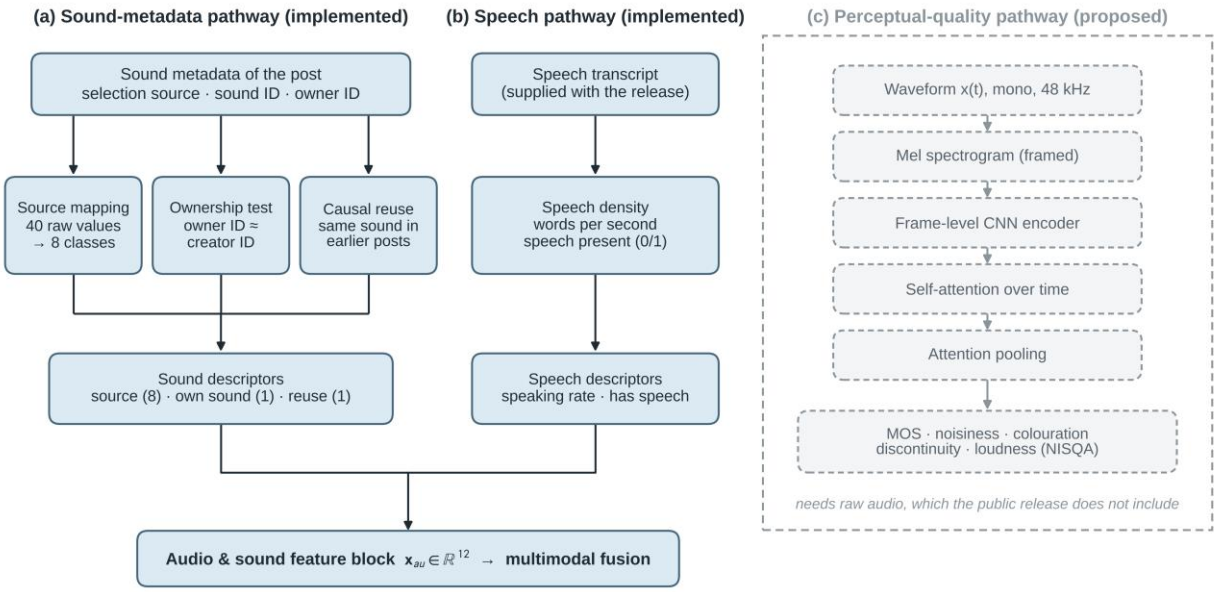


FIGURE 2: Audio and Sound Descriptor Extraction. (a) Sound metadata are mapped to source classes, an own-sound flag, and a causal reuse count. (b) Speech density is derived from the transcript statistics supplied with the release. (c) A NISQA-style perceptual-quality branch (dashed) shows where raw-audio descriptors would enter; it was not executed because the public release contains no audio.

4.3 Visual and Technical Attributes

The release describes each post's video stream by its duration and encoded resolution and states whether TikTok labels the post as AI-generated. We use log duration, resolution, a high-definition flag (≥ 720p), and the AI-generated flag. We add two posting-context variables, the hour of posting and a weekend indicator, to this block because they are properties of the upload rather than of its content. Frame-level visual quality and object or scene descriptors would require the video files and are not available in the release.

4.4 Semantic Content

Content semantics enter through the descriptors that the release computes from the caption and the speech transcript: one of nine topic labels (for example, beauty and fashion, dance and music, or cooking and food), the caption language, the number of words, emojis, questions and hashtags, and six emotion scores (anger, joy, surprise, sadness, disgust and fear). We add the total emotion mass, an indicator of joy as the dominant emotion, and an indicator of missing text statistics, which are absent for 33.2% of posts. This gives 23 semantic descriptors. The release does not document the models used for the topic and emotion annotations,

so we treat these descriptors as given and interpret them with caution.

4.5 Interaction Structure

After publication, viewers can like, comment on, share, save (TikTok's "collect"), download, or forward a post to WhatsApp. The mix of these actions describes how an audience engages, independently of how large it is. For each action c we compute its smoothed rate per view at horizon τ,

$$e_i^c(\tau) = \ln\big(1 + C_i^c(\tau)\big) - \ln\big(1 + V_i(\tau)\big) \quad (10)$$

Where $C_i^c(\tau)$ is the cumulative count of action c. We add the comment-to-like and save-to-share ratios and, for $\tau \geq 2$, the day-on-day growth of likes and comments. Shares, saves and downloads signal that viewers want to keep or pass on a post, which differs from the passive approval expressed by a like. The release contains no comment text, so we cannot measure comment sentiment or the duplication of comments that Mi et al. [14] used to detect paid engagement; repeated comments are in any case not proof of manipulation.

4.6 Temporal Dynamics

Let $\ell_i(k) = \ln(1 + V_i(k))$. At horizon τ we use $\ell_i(\tau)$ and $\ell_i(1)$, the view velocity and acceleration,

$$v_i(\tau) = \ell_i(\tau) - \ell_i(\tau - 1) \quad (11)$$
$$\kappa_i(\tau) = v_i(\tau) - v_i(\tau - 1) \quad (12)$$

The decay ratio of daily gains

$$\Delta V_i(k) = V_i(k) - V_i(k-1),$$
$$\rho_i(\tau) = ln\big(1 + \Delta V_i(\tau)\big) - ln\big(1 + \Delta V_i(\tau - 1)\big) \quad (13)$$

The growth since day 1 and the reach ratio,

$$r_i(\tau) = \ell_i(\tau) - \ln(1 + F_i) \quad (14)$$

Which compares the audience a post has reached with the creator's follower base. A large reach ratio indicates distribution beyond followers, the mechanism that distinguishes the For You page from follower-based feeds [3]. The decay ratio is negative when daily gains shrink, which marks the transition from burst to saturation (Section 8.1). Day-0 snapshots exist for only 37.5% of posts, so we keep an indicator of their availability and leave the day-0-based velocity missing otherwise.

TABLE 2: Multimodal Feature Inventory (Number of Features Per Horizon).

Modality	C	τ ≥ 1	Main descriptors	Origin
Creator influence	11	11	log followers, following, total likes, videos; likes per video; follow ratio; follower and like momentum; track record (3)	derived (as-of, causal)
Audio & sound	12	12	8 source classes; own sound; causal sound reuse; speaking rate; speech present	metadata + derived
Visual & technical	6	6	log duration; resolution; HD flag; AI-generated flag; posting hour; weekend	metadata
Semantic content	23	23	9 topics; English flag; word, emoji, question, hashtag counts; text-missing flag; 6 emotions; emotion mass; joy dominant	release annotations
Interaction structure	–	8–10	rates per view of likes, comments, shares, saves, downloads, WhatsApp shares; comment/like; save/share; like and comment growth (τ ≥ 2)	derived from snapshots
Temporal dynamics	–	5–9	log views at τ and day 1; velocity; previous velocity, acceleration, decay ratio, growth since day 1 (τ ≥ 2); reach ratio; day-0 indicator	derived from snapshots
Total	**52**	**70–76**		

5 PREDICTION AND ATTRIBUTION FRAMEWORK

5.1 Overview

Figure 1 summarises the framework. For every horizon τ, we align the six feature blocks to the information available at that horizon, concatenate them into one vector $\mathbf{x}_\tau$ (feature-level or *early* fusion) and train a separate gradient-boosted tree model f_τ. Training one model per horizon lets the model weight the modalities differently at each stage of the life cycle, which is why we call the framework stage-aware. The outputs are the two class probabilities, the regression estimate, and an exact Shapley attribution of predictive performance to the modalities.

5.2 Gradient-Boosted Decision Trees

We use histogram-based gradient-boosted decision trees (GBDT) in the formulation popularised by LightGBM [21]. The model is an additive ensemble of M regression trees,

$$F_M(\mathbf{x}) = F_0 + \eta \sum_{m=1}^{M} T_m(\mathbf{x}), \qquad \hat{p} = \sigma\big(F_M(\mathbf{x})\big) \quad (15)$$

Where F_0 is the log-odds of the training base rate, η the learning rate, and σ the logistic function. The trees minimise the penalised logistic loss.

$$\mathcal{L} = -\sum_i [y_i \ln \hat{p}_i + (1 - y_i)\ln(1 - \hat{p}_i)] + \frac{\lambda}{2}\sum_{m,l} w_{ml}^2 \quad (16)$$

Where w_{ml} is the value of leaf l in tree m and λ an L2 penalty. Each new tree is fitted with a second-order approximation of the loss around the current model, using the gradients and Hessians

$$g_i = \hat{p}_i - y_i, \qquad h_i = \hat{p}_i(1-\hat{p}_i) \qquad (17)$$

For a leaf containing the sample set I_l, the optimal value and the gain of splitting a node into a left part L and a right part R are:

$$w_l^* = -\frac{G_l}{H_l+\lambda}, \qquad G_l = \sum_{i\in I_l} g_i,\ H_l = \sum_{i\in I_l} h_i (18)$$

$$\Delta = \frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right] \qquad (19)$$

Trees grow leaf-wise: the leaf with the largest gain is split next until L leaves are reached or a leaf would hold fewer than $n_{\min}$ samples. Continuous features are binned into at most 255 histogram bins, and missing values receive their own bin whose split direction is learned, so the model needs no imputation. Training stops when the loss on an internal 15% validation split has not improved for 40 rounds (Figure 3).

We chose GBDT because our inputs are heterogeneous tabular descriptors with heavy tails and structured missingness, a setting in which tree ensembles still match or outperform deep networks [22]; because tree splits capture threshold effects and interactions, such as a large follower base paired with weak early reach, without hand-crafted cross terms; and because a fit takes seconds, which makes the exhaustive Shapley analysis below feasible.

5.3 Baselines and Comparison Models

We compare the GBDT with two classical early-popularity models, three learned alternatives and a decision-level fusion design. The Szabo–Huberman (S-H) model [5] regresses final popularity on the log count at the horizon,

$$\ln\big(1+V_i(30)\big) = \beta_0 + \beta_1\,\ell_i(\tau) + \varepsilon_i \qquad (20)$$

and the multivariate linear (ML) model of Pinto et al. [6] uses the whole observed trajectory,

$$\ln\big(1+V_i(30)\big) = \beta_0 + \sum_{k=0}^{\tau}\theta_k\,\ell_i(k) + \gamma\,u_i + \varepsilon_i (21)$$

Where u_i indicates a missing day-0 snapshot. Both are fitted by least squares on the training period. For classification, we use their predicted log views (popularity) or predicted log growth (growth) as scores. At cold start, where no views exist, the baseline ranks posts by the creator's follower count. The learned alternatives receive the same features as the GBDT: L2-regularised logistic regression (LR), a two-layer perceptron (MLP, 128 and 64 units, early stopping) and a random forest (RF) [23], the learner used in the Douyin study [14]. LR and MLP inputs are median-imputed with missing indicators, winsorised at the 0.5th and 99.5th training percentiles, and standardised. We also evaluate a soft-voting average of GBDT and RF.

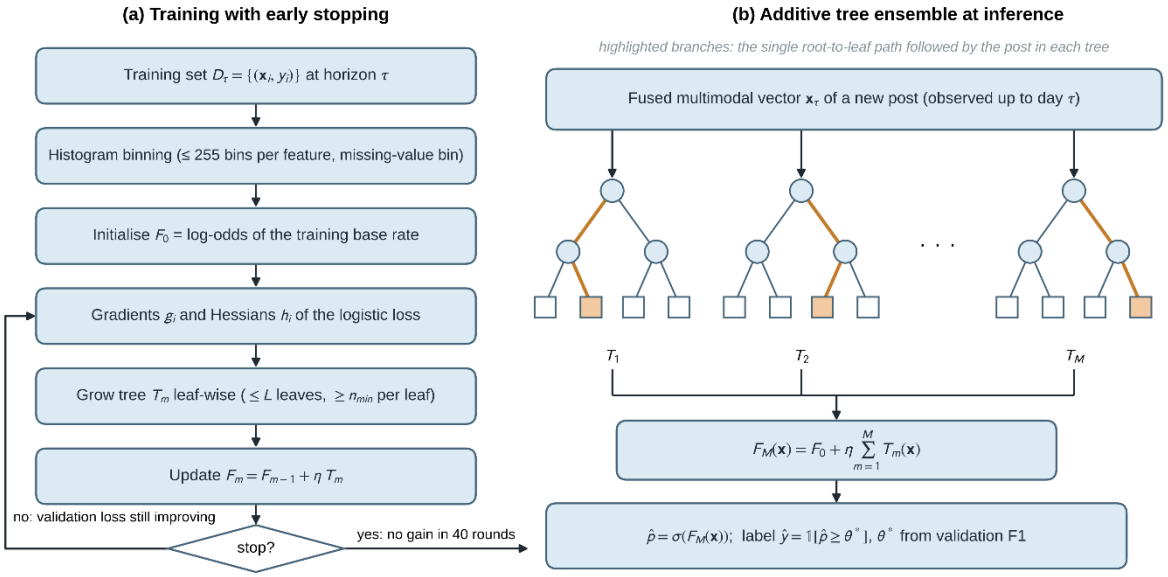


FIGURE 3: Gradient-Boosted Tree Model Used at Each Horizon. (a) Training: histogram binning, Newton-step trees fitted to gradients and Hessians of the logistic loss, and early stopping on an internal validation split. (b) Inference: each tree routes the post to one leaf, the leaf values are summed and mapped to a probability, and the decision threshold θ* is chosen on the validation period.

To test whether the modalities should interact inside one model, we implement decision-level (*late*) fusion. One GBDT is trained per modality, and a logistic meta-learner combines their out-of-fold logits:

$$\hat{p}^{\mathrm{LF}} = \sigma\Big(\omega_0 + \sum_{g\in\mathcal{M}} \omega_g\ \mathrm{logit}\big(\hat{p}_g\big)\Big) \qquad (22)$$

Where $\hat{p}_g$ is the probability from the model of modality $g \in \mathcal{M}$ and the out-of-fold predictions come from five-fold cross-fitting on the training period.

5.4 Shapley Attribution of Modalities

Per-feature attributions are hard to read when features are correlated and do not say how much each data source contributes to predictability. We therefore treat the modalities as players in a cooperative game whose value $v(S)$ is the test ROC-AUC of a GBDT trained on the modalities in coalition S, with $v(\emptyset) = 0.5$

for a random ranking. The Shapley value [24] of modality g is

$$\phi_g = \sum_{S \subseteq \mathcal{M} \setminus \{g\}} \pi(S)[v(S \cup \{g\}) - v(S)] \quad (23)$$

$$\pi(S) = \frac{|S|!\,(|\mathcal{M}|-|S|-1)!}{|\mathcal{M}|!} \quad (24)$$

and the contributions add up to the total improvement over chance,

$$\sum_{g \in \mathcal{M}} \phi_g = v(\mathcal{M}) - v(\varnothing), \qquad v(\varnothing) = 0.5 \quad (25)$$

With four modalities at cold start and six afterwards, we evaluate all $2^{|\mathcal{M}|}$ coalitions exactly (16 or 64 models per horizon and target) instead of sampling them. Each coalition model is retrained, not merely re-evaluated with features masked, so the attribution reflects how well a model can do with a given set of data sources. Because the Shapley axioms divide information shared by several players among them, Fryer et al. [24] caution against reading Shapley values as feature-selection criteria; we therefore report them alongside single-modality and leave-one-out results. All hyper-parameters are fixed to those selected for the full model, and no decision is taken on the basis of these test-set values.

6 DATASET AND PREPROCESSING

6.1 Data Source and Access

TikTok offers an official Research API, but access is limited to researchers at non-profit academic institutions in the United States, the European Economic Area, the United Kingdom, Switzerland, and a few other regions, and requires a vetted application [25]. The API reports engagement counts as of the query time, so daily trajectories have to be collected prospectively by repeated queries. We therefore use a public research release, the TikTok Video Engagement 200K dataset [17], distributed under a CC BY-NC 4.0 licence. The release links three tables by post and creator identifiers (Figure 4): a video table with post metadata and the day-30 view count, a daily engagement table with cumulative views, likes, comments, shares, saves, downloads and WhatsApp shares for days 0 to 5 after posting, and a daily creator table with follower, following, like and video counts. We use the 15,001-post subset redistributed in a public GitHub repository [17], which omits raw caption and transcript text but keeps the descriptors derived from them. Because the uploader does not document the collection procedure, we first checked the internal consistency of the data.

6.2 Integrity Checks

TikTok post identifiers embed the upload time in their upper 32 bits. For 99.59% of the 15,001 posts, this embedded time agreed with the recorded creation time to within one minute once a constant offset was removed. The offset was exactly five hours for posts created before 3 November 2024 and six hours afterwards, which matches the end of daylight-saving time in the United States and indicates that creation times are recorded in US Central time. The embedded time precedes the recorded publication time by a median of 8 s, consistent with an identifier assigned at upload. Cumulative counts decreased between consecutive snapshots in only 0.02% of view transitions and at most 0.34% of transitions for any action, and the day-30 view count was at least the last observed count for 99.99% of posts. These checks indicate genuine platform records rather than synthetic values, although they cannot reveal how the posts were sampled.

6.3 Sample and Target Distribution

Of the 12,000 posts with a day-30 label, we retain the 11,884 posts (1,606 creators) that have at least one engagement snapshot by day 1; the remaining 116 posts lack any early observation. Posts were created between 24 June and 9 November 2024, and 99.3% of them were published between 17:00 and 23:59 Central time, which indicates that the release sampled evening posts. The median creator has 13,860 followers and contributes 4 posts. Day-30 views are extremely skewed (Table 3): the median post reached 668 views, the 80th percentile 5,273, the 99th percentile 546,896, and the largest post 16,310,546. The 80th percentile of the training period, 5,338 views, defines the popularity label.

TABLE 3: Characteristics of the Analytic Sample.

Property	Value
Posts/creators	11,884 / 1,606
Posting period	24 Jun – 9 Nov 2024
Engagement snapshots (pool)	79,489 post-days
Creator snapshots (pool)	252,166 creator-days
Snapshot coverage, day 0 / days 1–5	37.5% / 98.2–99.4%
Day-30 views: median / mean	668 / 29,355
Day-30 views: P90 / P99 / max	21,725 / 546,896 / 16,310,546
Popularity threshold (training P80)	5,338 views
Median duration / 540p share	20.4 s / 92.7%
Original sound / English caption	38.2% / 64.2%
Largest topics	beauty & fashion 22.0%, other 16.3%, life hacks 15.4%
Training/test posts	9,507 / 2,377
Test posts by unseen creators	112

P80, P90, P99: 80th, 90th and 99th percentiles; max: maximum. Pool: posts and creators of the release subset used here.

6.4 Preprocessing and Splitting

Cumulative counts are forward filled from the latest earlier snapshot when a day is missing, which never uses later information. Day-0 snapshots are mostly missing because the first crawl usually followed an evening upload on the next day, so day-0 features stay missing. Missing values are left to the native handling of the tree models and imputed with training medians plus missing-value indicators for the linear and neural models. We do not remove outliers, because viral posts are the phenomenon of interest; instead, we compress counts logarithmically and winsorise the inputs of linear and neural models at training percentiles.

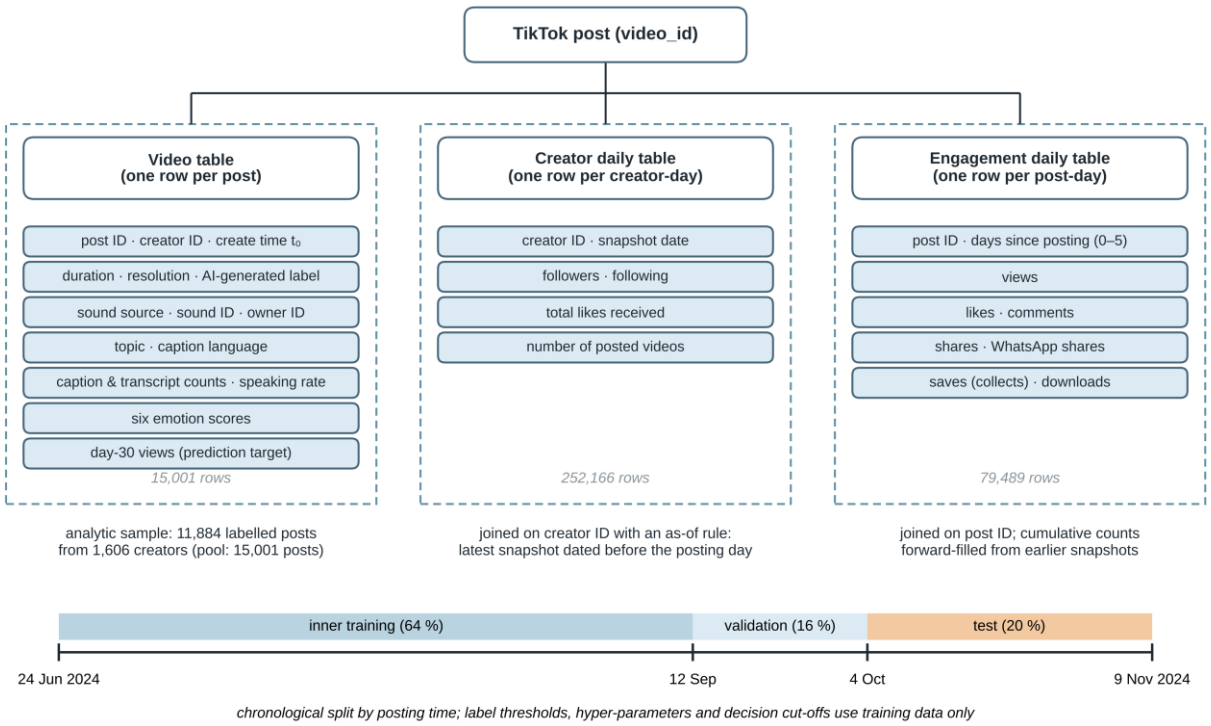


FIGURE 4: Data Tables, Joins, and Chronological Split.

We split the posts chronologically by creation time (Figure 4). The first 80% (9,507 posts, 24 June – 4 October 2024) form the training period and the last 20% (2,377 posts, 4 October – 9 November 2024) the test period. Within the training period, the last 20% (from 12 September) serves as a validation period for hyper-parameter and threshold selection, after which the final models are refitted on the whole training period. Because creators keep posting, 2,265 test posts come from creators who also appear in training, a realistic situation for a deployed model; Section 8.7 evaluates generalisation to unseen creators separately.

7 EXPERIMENTAL SETUP

7.1 Evaluation Metrics

For classification, we report accuracy, precision, recall and F1-score,

$$\mathrm{P} = \frac{TP}{TP+FP}, \qquad \mathrm{R} = \frac{TP}{TP+FN} \tag{26}$$

$$F_1 = \frac{2\,TP}{2\,TP+FP+FN}, \qquad \mathrm{Acc} = \frac{TP+TN}{N_{\text{test}}} \tag{27}$$

Together with the ROC-AUC, which equals the probability that a random positive post is ranked above a random negative post,

$$\mathrm{AUC} = \frac{1}{|y^+||y^-|}\sum_{p\in y^+}\sum_{n\in y^-}\mathbb{I}\left[\hat{s}_p > \hat{s}_n\right] \tag{28}$$

and the area under the precision–recall curve (PR-AUC, computed as average precision), which is more informative than ROC-AUC when positives are rare. Threshold-dependent metrics use the cut-off that maximises F1 on the validation period,

$$\theta^* = \operatorname*{argmax}_{\theta} F_1(\theta; \mathcal{D}_{\text{val}}) \tag{29}$$

So that no test information influences the decision rule. For regression, we report the mean absolute error (MAE) and root-mean-square error (RMSE) of $\ln(1 + V(30))$, the coefficient of determination R^2 and Spearman's rank correlation ρ.

7.2 Hyper-Parameter Selection

Each learned model is tuned separately for every target and horizon by grid search on the validation period, with ROC-AUC as the criterion (Table 4). The GBDT grid crosses three learning rates, four leaf limits, and two minimum leaf sizes (24 configurations) with L2 penalty λ = 1 and early stopping. The random forest is tuned with 300 trees and refitted with 500. Logistic regression and the MLP are tuned over their regularisation strengths. All random procedures use the seed 42.

7.3 Statistical Analysis and Implementation

We estimate 95% confidence intervals (CIs) for ROC-AUC, PR-AUC and F1 with 1,000 bootstrap resamples of the test set and compare the GBDT with each alternative by a two-sided paired bootstrap test on the ROC-AUC difference (2,000 resamples). To check that results do not depend on the creators shared

between training and test periods, we also run five-fold cross-validation grouped by creator, so that no creator appears in both training and validation folds, and we evaluate the chronological test set separately for posts from creators unseen in training. All experiments run in Python 3.13 with scikit-learn 1.9.1, pandas 3.0 and NumPy 2.5 on a two-core CPU; a full GBDT fit takes 1–4 s.

8 RESULTS

8.1 Life-Cycle Structure of TikTok Views

Figure 5 shows how views accumulate. The typical post gains most of its audience almost immediately: by the first daily snapshot, the median post had already collected 68% of its day-30 views, and 72% of posts had passed half of their final count. By day 5, the median share reached 86%. Popular posts follow a different curve. Posts that ended in the top 20% had collected a median of only 39% of their final views by day 1 and 74% by day 5, against 73% and 88% for the other posts. Their daily gains also decay more slowly in relative terms (Figure 5(b)). Popularity on TikTok is therefore not only a large initial burst but also a longer tail of continued distribution, which is what the growth target measures. The three stages in Figure 5 (a cold-start burst within the first day, deceleration over days 1–5, and a long tail to day 30) motivate the horizons we study.

TABLE 4: Hyper-Parameter Search Spaces and Selected Values (Validation ROC-AUC).

Model	Search space	Selected: popularity, C	Selected: popularity, τ = 1	Selected: growth, τ = 1
GBDT	η ∈ {0.02, 0.05, 0.1}; L ∈ {7, 15, 31, 63}; n_{min} ∈ {20, 60}	η = 0.1, L = 7, n_{min} = 20	η = 0.05, L = 63, n_{min} = 20	η = 0.02, L = 31, n_{min} = 20
RF	max_features ∈ {0.2, 0.4, √p}; min_leaf ∈ {2, 8}	max_features = 0.4, min_leaf = 8	max_features = 0.4, min_leaf = 8	max_features = 0.2, min_leaf = 2
LR	C ∈ {0.01, 0.1, 1}	C = 0.1	C = 0.1	C = 1
MLP	α ∈ {10^{-4}, 10^{-3}, 10^{-2}}	α = 0.0001	α = 0.001	α = 0.001

η: learning rate; L: maximum leaves per tree; n_{min}: minimum samples per leaf; max_features: share of features tried per split (√p = square root of the number of features). Selected values for the other horizons are listed in the released result files.

8.2 Comparison of Models

Table 5 compares all models at cold start and at τ = 1, and Figure 6 shows the ROC curves. At cold start, the GBDT reaches a ROC-AUC of 0.942 and a PR-AUC of 0.861, against 0.882 and 0.701 for ranking by follower count (p < 0.001). Its F1-score of 0.784 is 0.124 higher than that of the heuristic. With a positive rate of about 20%, these figures mean that a pre-publication screening model would find about 78% of the eventual top posts while 79% of its alerts are correct. Logistic regression and the MLP come close in ROC-AUC (0.937 and 0.935) but remain significantly below the GBDT (p = 0.046 and p = 0.032), while the random forest (0.940) and late fusion (0.942) are statistically indistinguishable from it.

After one day, popularity becomes easy to predict even for the classical models: the S-H log-linear model reaches 0.971, because the day-1 count already contains 68% of the typical post's final views. The GBDT still improves on it to 0.982 (Δ = 0.011, p < 0.001), raises the PR-AUC from 0.944 to 0.958 and the precision from 0.838 to 0.925. The random forest is equivalent here (|Δ| < 0.001, p = 0.915).

The growth task separates the models clearly. The S-H and multivariate linear models reach only 0.677 and 0.685: early view counts say little about whether a post will keep growing. Logistic regression lifts the ROC-AUC to 0.807 and the MLP to 0.829. The tree ensembles do best: the random forest reaches 0.853 and the GBDT 0.864 (95% CI [0.849, 0.879]), with an F1-score of 0.783 and a PR-AUC of 0.843. The advantage of the GBDT over the random forest (Δ = 0.011) and over late fusion (Δ = 0.021) is significant (p < 0.001). Averaging the GBDT with the random forest brings no gain (Δ = −0.001, p = 0.380), so we keep the single GBDT as the proposed model.

When each modality is modelled separately and only the modality-level predictions are combined, the growth ROC-AUC falls from 0.864 to 0.843. Growth therefore depends on interactions between modalities, for example between a post's early reach and its creator's follower base, which a stacked combination of separate models cannot represent.

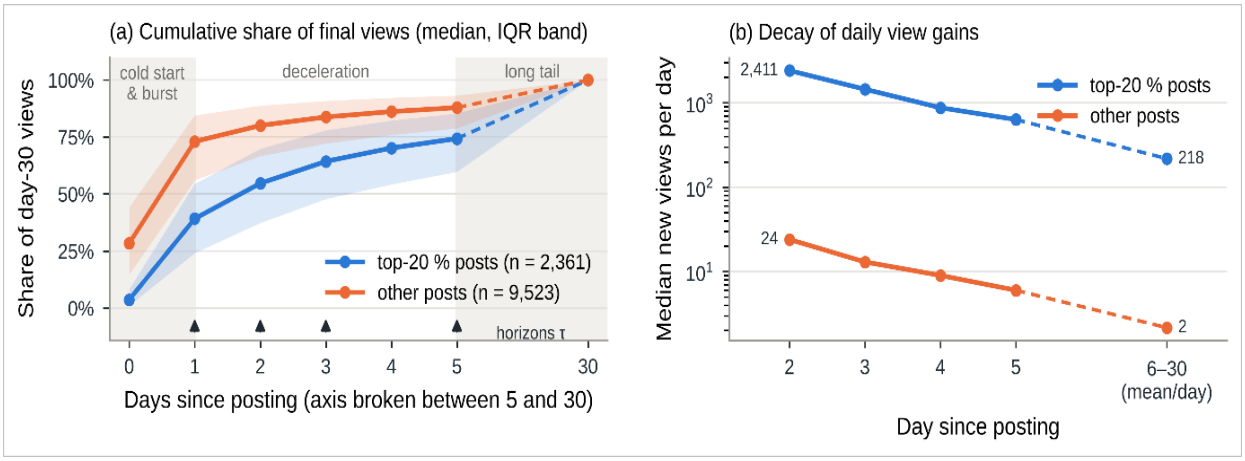


FIGURE 5: Empirical Life Cycle of 11,884 TikTok Posts**.** (a) Median share of day-30 views accumulated by each day, with interquartile bands, for posts that reached the top 20% and for the others; day 0 is observed for 38% of posts only. Arrows mark the prediction horizons. (b) Median daily view gains on days 2–5 and mean daily gain over days 6–30 (log scale). Dashed segments bridge the broken time axis.

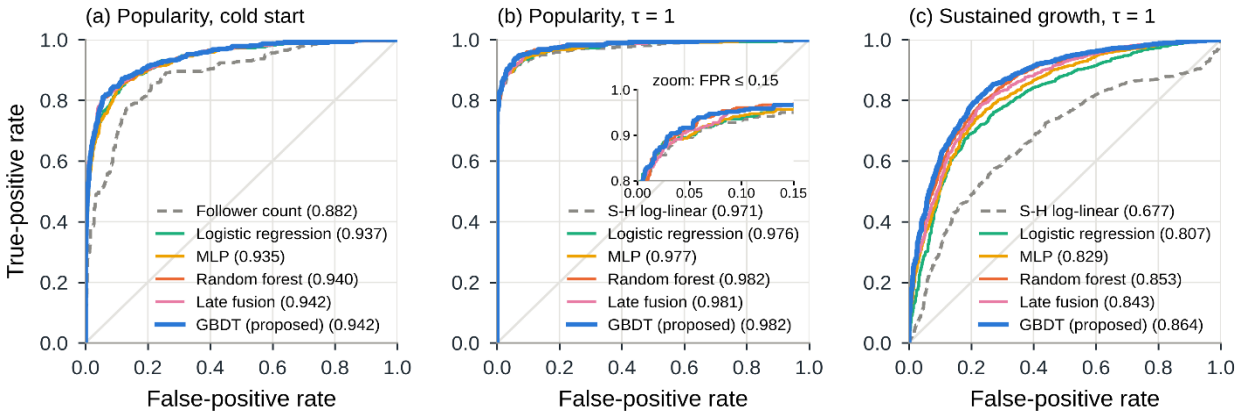


FIGURE 6: ROC Curves on the Chronological Test Set. (a) Popularity at cold start, with the follower-count heuristic as baseline. (b) Popularity after one day, with the Szabo–Huberman (S-H) model as baseline; the inset enlarges the low false-positive region. (c) Sustained growth after one day. Areas under the curve are given in parentheses.

8.3 Performance Across Horizons

Table 6 follows the models through the life cycle. Popularity prediction improves quickly with each additional day of data, from 0.942 at cold start to 0.982 on day 1 and 0.994 on day 5, and the gap to the S-H baseline narrows from 0.011 on day 1 to 0.008 on day 5. Growth prediction behaves differently. The GBDT reaches 0.864 on day 1, 0.882 on day 2, 0.895 on day 3, and 0.897 on day 5, while the S-H baseline stays between 0.666 and 0.679. Later horizons help because the velocity, acceleration, and decay ratio can only be computed from several snapshots. The regression results tell the same story: the RMSE of the log day-30 views falls from 1.026 at cold start to 0.502 on day 1, compared with 0.642 for the S-H model, and the Spearman correlation between predicted and observed values reaches 0.964. The GBDT has the lowest RMSE on days 1–3, whereas the random forest is slightly better at cold start (1.021 vs. 1.026) and on day 5 (0.319 vs. 0.361). On day 5, the GBDT's RMSE appears to be inflated by a few large errors: its MAE is similar to that of the random forest (GBDT 0.149 vs. RF 0.147) and its median absolute percentage error on the raw scale is lower (GBDT 8.1% vs. RF 8.5%).

The relative standing of the comparison models also changes with the horizon. The multivariate linear model, which exploits the whole observed trajectory, gains most from longer observation: its growth ROC-AUC rises from 0.685 on day 1 to 0.841 on day 5, whereas the single-snapshot S-H model does not improve. The advantage of early over late fusion for growth is largest at the first horizons (Δ = 0.021, 0.011 and 0.010 on days 1, 2 and 3, all p < 0.001) and disappears by day 5 (Δ = 0.003, p = 0.336), when the temporal block alone carries most of the signal. The same holds for the random forest, which is significantly weaker than the GBDT for growth on days 1–3 but equivalent on day 5 (p = 0.615); on day 5 the GBDT–RF average is marginally better than the GBDT alone (Δ = 0.002, p = 0.032). Cross-modal interactions thus matter most when the engagement history is still short, which is exactly when an early warning is useful.

8.4 Hyper-Parameter Sensitivity

Figure 7(a, b) shows the validation ROC-AUC of the GBDT over the learning rate and the number of leaves. Across all 24 configurations, the validation ROC-AUC varies between 0.980 and 0.983 for popularity and between 0.874 and 0.881 for growth.

The model is insensitive to these choices, so the gains reported above do not come from fine-tuning. Figure 7(c) shows the training-period ROC-AUC of the creator influence index as a function of the follower weight α. Neither component alone is optimal: followers alone (α = 1) give 0.864 and likes per video alone (α = 0) give 0.873, whereas the balanced index with α* = 0.45 reaches 0.888. On the test period, this index alone achieves 0.903, compared with 0.882 for follower count, which indicates that the audience a creator has and the response that audience gives are complementary.

TABLE 5: Test-Set Performance at Cold Start and After One Day
(Chronological Split; Threshold-Dependent Metrics Use the Validation-Optimal F1 Cut-Off).

Model	Acc	Prec	Rec	F1	ROC-AUC [95% CI]	PR-AUC	*p* vs. GBDT
Popularity (top 20% of day-30 views), cold start							
Follower-count heuristic	0.845	0.572	0.782	0.661	0.882 [0.865, 0.900]	0.701	< 0.001
Logistic regression	0.915	0.818	0.717	0.764	0.937 [0.923, 0.949]	0.846	0.046
MLP	0.910	0.826	0.673	0.742	0.935 [0.922, 0.947]	0.837	0.032
Random forest [23]	0.919	0.793	0.784	0.789	0.940 [0.927, 0.952]	0.858	0.355
Late fusion (stacked GBDTs)	0.923	0.829	0.758	0.792	0.942 [0.930, 0.955]	0.865	0.911
GBDT, early fusion (proposed)	**0.917**	**0.789**	**0.780**	**0.784**	**0.942 [0.930, 0.954]**	**0.861**	**–**
Popularity (top 20% of day-30 views), τ = 1 day							
S-H log-linear [5]	0.946	0.838	0.893	0.865	0.971 [0.961, 0.981]	0.944	< 0.001
Multivariate linear [6]	0.951	0.874	0.874	0.874	0.971 [0.961, 0.981]	0.944	< 0.001
Logistic regression	0.957	0.901	0.871	0.886	0.976 [0.967, 0.985]	0.951	0.003
MLP	0.950	0.856	0.891	0.873	0.977 [0.968, 0.985]	0.949	0.036
Random forest [23]	0.960	0.925	0.861	0.892	0.982 [0.975, 0.989]	0.958	0.915
Late fusion (stacked GBDTs)	0.955	0.927	0.834	0.878	0.981 [0.974, 0.987]	0.954	0.367
GBDT, early fusion (proposed)	**0.959**	**0.925**	**0.856**	**0.889**	**0.982 [0.975, 0.989]**	**0.958**	**–**
Sustained growth after the horizon, τ = 1 day							
S-H log-linear [5]	0.602	0.549	0.800	0.651	0.677 [0.655, 0.700]	0.636	< 0.001
Multivariate linear [6]	0.652	0.615	0.670	0.641	0.685 [0.664, 0.706]	0.646	< 0.001
Logistic regression	0.676	0.603	0.883	0.717	0.807 [0.791, 0.824]	0.771	< 0.001
MLP	0.745	0.693	0.809	0.746	0.829 [0.813, 0.845]	0.802	< 0.001
Random forest [23]	0.763	0.696	0.869	0.773	0.853 [0.838, 0.868]	0.829	< 0.001
Late fusion (stacked GBDTs)	0.756	0.696	0.842	0.762	0.843 [0.828, 0.859]	0.815	< 0.001
GBDT, early fusion (proposed)	**0.790**	**0.752**	**0.816**	**0.783**	**0.864 [0.849, 0.879]**	**0.843**	**–**

Acc: accuracy; Prec: precision; Rec: recall. 95% confidence intervals from 1,000 bootstrap resamples of the test set; *p*: two-sided paired bootstrap test of the ROC-AUC difference with the proposed GBDT (2,000 resamples). Test set: 2,377 posts; positive rate 19.3% (popularity) and 46.4% (growth).

8.5 Modality Attribution Across the Life Cycle

Figure 8 decomposes the improvement of the test ROC-AUC over chance into exact Shapley contributions, and Table 7 lists the ROC-AUC of each modality on its own and the loss when it is removed. At cold start, creator influence contributes 0.301 (68%) of the total improvement of 0.442, followed by visual and technical attributes with 0.054 (12%), semantic content with 0.051 (12%), and audio and sound with 0.036 (8%). Creator influence alone reaches a ROC-AUC of 0.941, almost the value of the full model (0.942), and removing it costs 0.216. The content modalities carry real signals on their own (ROC-AUC between 0.617 and 0.655), but this signal is largely redundant with the creator description: removing any one of them changes the cold-start ROC-AUC by at most 0.002. Their Shapley shares therefore reflect information they share with the creator block, which the Shapley value divides among the players, rather than unique information.

TABLE 6: ROC-AUC, PR-AUC And LOG-View RMSE of the Proposed GBDT and the Classical Baselines at Every Horizon.

Horizon	Popularity: ROC-AUC		GBDT PR-AUC	Growth: ROC-AUC		GBDT PR-AUC	Regression: RMSE of ln(1 + *V*(30))		
	baseline	GBDT		S-H	GBDT		baseline	RF	GBDT
cold start	0.882	**0.942**	0.861	–	–	–	1.373	**1.021**	1.026
day 1	0.971	**0.982**	0.958	0.677	**0.864**	0.843	0.642	0.527	**0.502**
day 2	0.985	**0.991**	0.977	0.679	**0.882**	0.872	0.487	0.392	**0.383**
day 3	0.988	**0.991**	0.981	0.667	**0.895**	0.889	0.467	0.333	**0.329**
day 5	0.986	**0.994**	0.985	0.666	**0.897**	0.887	0.444	**0.319**	0.361

Baselines: follower-count heuristic at cold start and Szabo–Huberman log-linear model afterwards; the growth target is undefined at cold start. Regression: RMSE of ln(1 + day-30 views) for the baseline, the random forest (RF) and the GBDT; the lowest RMSE per horizon is in bold.

After publication, popularity is explained by three blocks of similar weight. At τ = 1, temporal dynamics contribute 0.143 (30%), interaction structure 0.137 (28%) and creator influence 0.125 (26%), and these shares hardly change up to day 5. Interaction structure alone reaches 0.968. Part of this signal is the scale of exposure rather than audience response: sparse actions such as WhatsApp shares are zero for most posts on day 1, so their smoothed per-view rate largely mirrors the view count. Growth prediction shows a clear shift across the life cycle. On day 1, creator influence is the largest contributor (0.142, 39%), ahead of temporal dynamics (0.097, 27%) and interaction structure (0.082, 23%), and it is the only block whose removal costs much (0.049). From day 2 onward, temporal dynamics take the lead, with 37% on day 2, 41% on day 3 and 42% on day 5, and removing them costs 0.052 on day 3 and 0.052 on day 5, while the creator shares declines to 27%. As the trajectory lengthens, the post's own velocity, acceleration and decay increasingly reveal what the creator's profile could only anticipate. The three content modalities together account for at most 12% of the growth signal at any horizon.

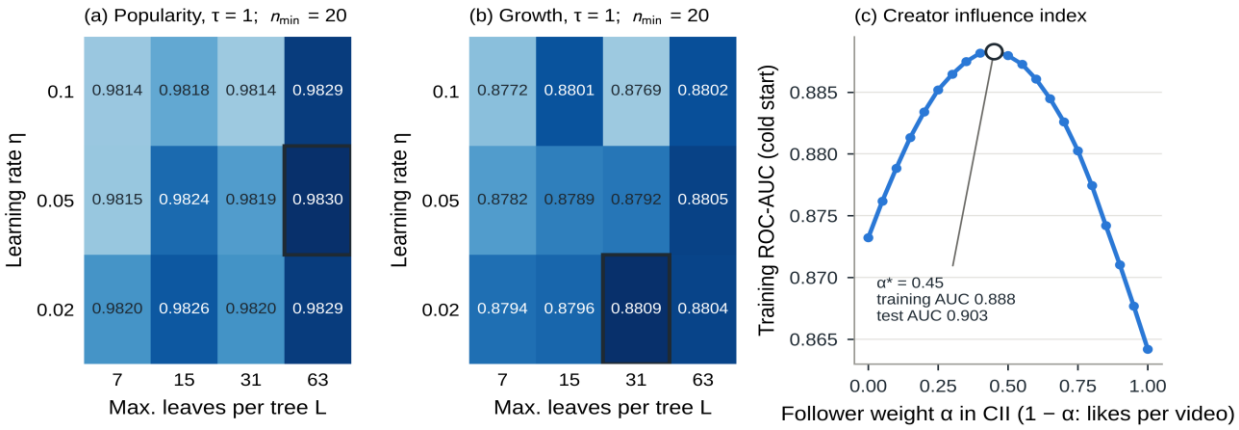


FIGURE 7: Hyper-Parameter Sensitivity. (a, b) Validation ROC-AUC of the GBDT at τ = 1 over learning rate and maximum number of leaves, at the selected minimum leaf size; the outlined cell is the selected configuration. (c) Training-period ROC-AUC of the creator influence index CII(α) at cold start as a function of the follower weight α.

TABLE 7: Single-Modality ROC-AUC (Alone) and Loss When the Modality is Removed (Δ Without), Test Set.

Modality	Popularity, cold start		Popularity, τ = 1		Growth, τ = 1		Growth, τ = 3	
	alone	Δ without	alone	Δ without	alone	Δ without	alone	Δ without
Creator influence	0.941	0.216	0.939	0.004	0.830	0.049	0.809	0.022
Audio & sound	0.617	0.001	0.601	0.001	0.555	0.002	0.529	0.000
Visual & technical	0.655	0.001	0.636	0.000	0.583	0.002	0.575	0.000
Semantic content	0.651	−0.002	0.651	0.000	0.584	−0.001	0.577	0.000
Interaction structure	–	–	0.968	−0.001	0.754	0.007	0.773	0.002
Temporal dynamics	–	–	0.978	0.002	0.783	0.007	0.874	0.052
All modalities	**0.942**		**0.982**		**0.864**		**0.895**	

All values come from the coalition models of the Shapley analysis. Δ without = ROC-AUC (all modalities) − ROC-AUC (all but the modality), i.e., the loss caused by removing the modality; a negative value means the model is slightly better without it. Interaction structure and temporal dynamics are unavailable at cold start.

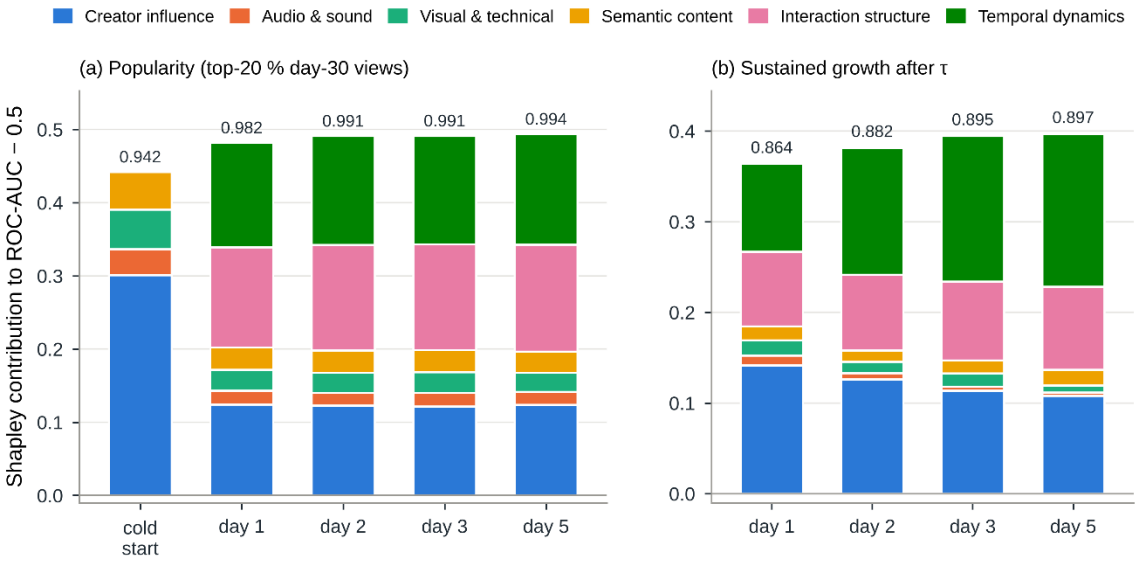


FIGURE 8: Exact Shapley Decomposition of Test Roc-Auc Over Modalities at Each Horizon (All 16 or 64 Coalitions Retrained). Bar height is the improvement over a random ranking (ROC-AUC − 0.5); the number above each bar is the ROC-AUC of the model with all modalities. (a) Popularity. (b) Sustained growth.

8.6 Feature-Level Importance

Permutation importance on the test set (Figure 9) refines this picture at the level of single features. For popularity at τ = 1, the log view count at the horizon dominates: permuting it lowers the ROC-AUC by 0.201, whereas no other feature costs more than about 0.002. The next features, creator track record (day-1 views), likes per video, WhatsApp-share rate, log followers, help to separate posts with similar day-1 counts. The growth model relies on different information. Its most important features are log followers (0.033), log video count (0.026) and likes per video (0.021), followed by creator track record (day-1 views), creator track record (day-30 views), view velocity, follower momentum, like rate, log views at τ, reach ratio.

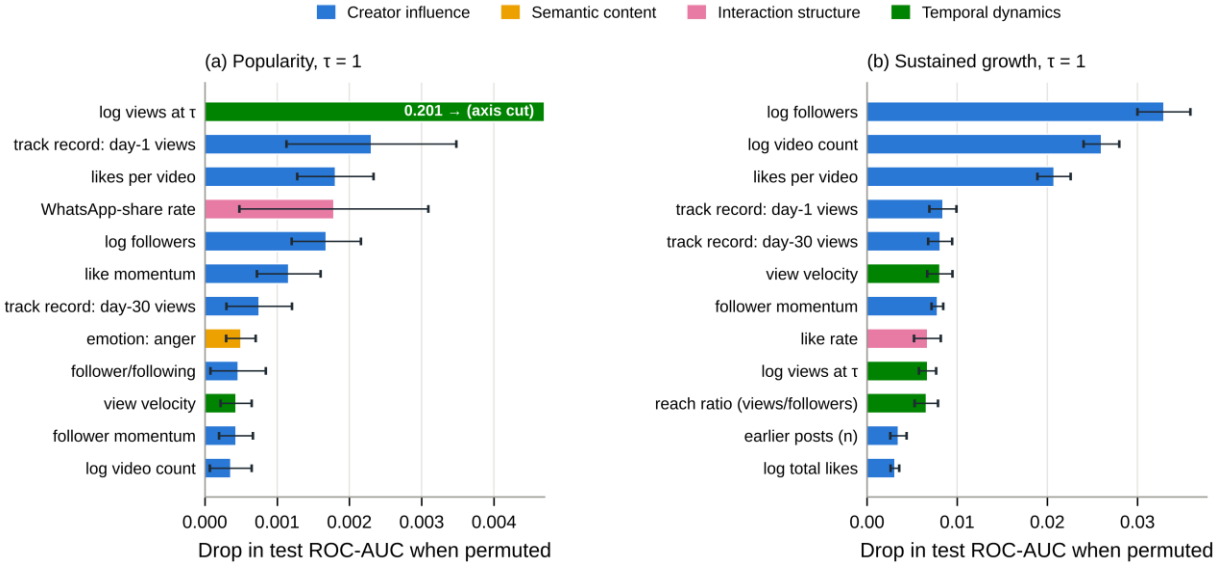


FIGURE 9: Permutation Importance of the Twelve Most Important Features for the GBDT at T = 1 (Mean Decrease in Test ROC-AUC Over Ten Permutations; Error Bars: ±1 S.D.). Colours denote modalities. (a) Popularity. (b) Sustained growth.

The direction of these effects is clear in the training data. Among posts in the lowest follower quintile, only 26% grew faster than the median after day 1, against 86% in the highest quintile; for likes per video the corresponding shares are 21% and 83%. The reach ratio acts in the opposite direction: posts whose day-1 views were smallest relative to the follower base grew most afterwards (71% above the median in the lowest quintile, 44% in the highest). A plausible reading is that TikTok attention has two sources with different time profiles. Recommendation-driven exposure arrives in an early burst and saturates quickly, whereas an established audience keeps returning to a creator's posts for weeks. A post that has already been carried far beyond its creator's followers on day 1 has used up much of its burst, while a post from a large, responsive account still has its follower demand ahead of it. Permutation importance understates features that are strongly correlated with others, such as the creator's follower and like counts, so we read it together with the group-level Shapley analysis.

8.7 Robustness

- ***Creator-Disjoint Validation***

When the folds are grouped by creator, so that no creator contributes posts to both training and evaluation, the GBDT still reaches a mean ROC-AUC of 0.930 (s.d. 0.014) at cold start, 0.986 (0.004) for popularity at $\tau = 1$ and 0.858 (0.010) for growth at $\tau = 1$, against 0.865, 0.983 and 0.704 for the corresponding baselines and 0.926, 0.985 and 0.817 for logistic regression. On the chronological test set, the 112 posts by creators never seen in training obtain ROC-AUCs of 0.948 (cold start), 0.982 (popularity, $\tau = 1$) and 0.900 (growth, $\tau = 1$), compared with 0.942, 0.982 and 0.862 for posts by known creators. These estimates rest on few posts but indicate that the models do not merely memorise individual creators.

- ***Popularity Threshold***

We re-ran the GBDT with the hyper-parameters selected for the top-20% task, changing only the threshold. At cold start the advantage over the follower heuristic holds for every definition: the ROC-AUC is 0.943, 0.942 and 0.910 for the top 10%, 20% and 50%, against 0.883, 0.882 and 0.795, and for the rarest class the PR-AUC rises from 0.539 to 0.760. At $\tau = 1$ the ROC-AUC values are 0.983, 0.982 and 0.986 against 0.977, 0.971 and 0.969 for the S-H baseline. For the top 10% at $\tau = 1$, however, the PR-AUC of the GBDT (0.928) does not exceed that of the S-H baseline (0.930): once a day of views is known, the extreme tail is already well separated by the view count, and a model tuned for another threshold gains nothing in precision there. Users interested only in the extreme tail after publication should tune for that threshold or keep the simple baseline.

- ***Comparison with the Douyin Study***

Mi et al. [14] reported an F1-score of 69.3% and an AUC of 71.3% for Douyin interaction growth. Both studies split growth at the median, but our target uses views instead of combined interactions and a 30-day instead of a 12.5-hour window, and the data come from another platform and a larger sample (11,884 vs. 1,127 videos), so the numbers are not directly comparable. Even so, the S-H baseline on our growth target (ROC-AUC 0.677) is of a similar magnitude to their reported random-forest AUC of 0.713, and the proposed GBDT adds about 0.187 in ROC-AUC on top of it.

9 DISCUSSIONS

9.1 What Predicts TikTok Popularity, and When

Our results separate two questions that are often conflated. Whether a post will end up popular is largely decided by the time of the first daily snapshot, and from then on the current view count is a strong predictor that any reasonable model exploits. Whether a post will keep growing is a different and harder question. Here the classical extrapolation models are only moderately better than chance, and the multimodal GBDT improves the ROC-AUC by about 0.19–0.23 at every horizon. This question matters in practice: on day 1, a moderator already knows a post's current popularity and wants to know where it is going.

9.2 Creator Influence on a For You Platform

TikTok is known for distributing content beyond followers [3], and the For You page does let small accounts break out. Our results show, however, that the creator remains the most informative single source both before publication and for the question of whether a post will keep growing after its first day. A plausible reading is that large and responsive audiences sustain views long after the recommendation burst has faded, whereas posts that owe their early views mainly to wide algorithmic exposure tend to saturate. Both findings agree with Ling et al. [2], for whom follower count was the strongest indicator of virality, and with Paekivi and Karjus [19], who found that follower count and video age explain more engagement variance than content variables for brand accounts, and they qualify the folk theory that every TikTok post has an equal chance [1]. For platform analysts, they imply that early-warning systems should model a post's views relative to its creator's audience rather than in absolute terms.

9.3 Why Early Fusion

For growth prediction, early fusion outperformed decision-level fusion and bagged trees during the first three days. The likely reason is that the informative quantities are relational: an early view count means something different for an account with 2,000 followers than for one with 2 million, and a high share rate means more when views are still accelerating. Gradient-boosted trees learn such relations from the joint feature vector, whereas modality-specific models combined afterwards see each piece in isolation. This argues for joint modelling in future multimodal systems, whether they use trees on descriptors or neural encoders on raw content.

9.4 Transfer from Douyin

We modelled the TikTok framework on the Douyin study [14] but did not assume that its components transfer, and three design decisions differ. We measure popularity by views rather than by combined interactions, because views are the quantity that TikTok's For You distribution allocates. We follow posts for 30 days with daily snapshots rather than for 12.5 hours with 15-minute sampling; the life-cycle analysis shows that popular TikTok posts keep accumulating views long after the first day, so a short window would miss the outcome that distinguishes them. Finally, we do not treat repeated comments as evidence of paid engagement, since short, duplicated comments can also arise organically; detecting manipulation needs comment text and account-level evidence that the public release does not contain.

9.5 Implications for Practice

A cold-start model with a ROC-AUC of 0.942 can rank incoming posts for review before any engagement exists, and a day-1 growth model can flag posts that are about to keep spreading. Both are computationally light (on a two-core CPU, a fit takes 1–4 s and scoring takes less than 0.1 ms per post) and use only metadata and early engagement counts,

and the Shapley analysis indicates which data sources are worth collecting at each stage.

10 LIMITATIONS AND FUTURE WORK

Several limitations qualify our conclusions. First, the data come from a single public release whose collection procedure is undocumented. Our integrity checks show that identifiers, timestamps and counts are internally consistent, but they cannot reveal the sampling frame. The posts were published almost exclusively in the evening (US Central time) between June and November 2024, and the release appears to favour English-language and creator-economy content, so the results may not generalise to other regions, posting times or periods. Replicating the study with data collected prospectively through the TikTok Research API [25], for example by querying the same posts daily, is the most important next step. The full release of about 210,000 posts [17] would also allow larger and more diverse test sets.

Second, the public subset contains no raw video, audio or comment text. Our audio, visual, and semantic modalities therefore rely on metadata and on annotations supplied with the release whose models are undocumented. Perceptual audio quality (Figure 2(c)), frame-level visual quality, object and scene content, comment sentiment and comment originality could not be measured, and the modest contribution of the content modalities may partly reflect these coarse descriptors rather than the true value of content. Third, daily snapshots are coarse: the effective age of a post at "day 1" varies with the posting and crawling times, and day-0 snapshots exist for only a minority of posts. Finer sampling would sharpen the cold-start and burst stages. Fourth, our definition of popularity uses views; other targets such as watch time or engagement continuation [7] capture different aspects of audience attention.

Future work should (i) add content encoders on raw frames, audio and comments, including a NISQA-style audio-quality branch, and test whether they add information beyond the descriptors studied here; (ii) model the engagement trajectory as a point process or sequence once finer snapshots are available; (iii) let the modality weights adapt continuously with the age of a post instead of training separate models per horizon; and (iv) study fairness: creator-centred predictors may reinforce existing visibility gaps if they are used to allocate promotion.

CONCLUSION

In this study, we developed a stage-aware, leakage-safe multimodal framework for TikTok popularity prediction and evaluated it on 11,884 public TikTok posts with daily engagement snapshots. Built from creator influence, audio and sound, visual and technical, semantic, interaction-structure and temporal-dynamics descriptors that are all observable at prediction time, a gradient-boosted tree model per horizon identified future top-20% posts with an ROC-AUC of 0.942 before any engagement and 0.982 after one day, and predicted sustained growth after day 1 with 0.864, far above classical early-popularity models. An exact Shapley decomposition showed that creator influence dominates before publication and for growth prediction on day 1, that the post's own view dynamics take over from day 2, and that metadata-level content descriptors add little beyond the creator; joint modelling of the modalities matters most early in a post's life. The data, code and protocol give future TikTok studies a reproducible baseline against which richer content models can be measured.

ACKNOWLEDGMENT

The authors thank the School of Electronics and Information Engineering, Nanjing University of Information Science and Technology, for providing research facilities and academic support during this work.

REFERENCES

[1] Klug, D., Qin, Y., Evans, M., & Kaufman, G. (2021). Trick and please. A mixed-method study on user assumptions about the TikTok algorithm. *Proceedings of the 13th ACM Web Science Conference (WebSci)*, pp. 84–92. https://doi.org/10.1145/3447535.3462512

[2] Ling, C., Blackburn, J., De Cristofaro, E., & Stringhini, G. (2022). Slapping cats, bopping heads, and Oreo shakes: Understanding indicators of virality in TikTok short videos. *Proceedings of the 14th ACM Web Science Conference (WebSci)*, pp. 164–173. https://doi.org/10.1145/3501247.3531551

[3] Guinaudeau, B., Munger, K., & Votta, F. (2022). Fifteen seconds of fame: TikTok and the supply side of social video. *Computational Communication Research*, 4(2), 463–485. https://doi.org/10.5117/CCR2022.2.004.GUIN

[4] Zannettou, S., Nemes-Nemeth, O., Ayalon, O., Goetzen, A., Gummadi, K. P., Redmiles, E. M., & Roesner, F. (2024). Analyzing user engagement with TikTok's short format video recommendations using data donations. *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI)*. https://doi.org/10.1145/3613904.3642433

[5] Szabo, G., & Huberman, B. A. (2010). Predicting the popularity of online content. *Communications of the ACM*, 53(8), 80–88. https://doi.org/10.1145/1787234.1787254

[6] Pinto, H., Almeida, J. M., & Gonçalves, M. A. (2013). Using early view patterns to predict the popularity of YouTube videos. *Proceedings of the 6th ACM International Conference on Web Search and Data Mining (WSDM)*, pp. 365–374. https://doi.org/10.1145/2433396.2433443

[7] Li, D., Li, W., Lu, B., Li, H., Ma, S., Krishnan, G., & Wang, J. (2024). Delving deep into engagement prediction of short videos. *Proceedings of the*

European Conference on Computer Vision (ECCV). https://doi.org/10.1007/978-3-031-72949-2_17

[8] Cho, M., Jeong, D., & Park, E. (2024). AMPS: Predicting popularity of short-form videos using multi-modal attention mechanisms in social media marketing environments. *Journal of Retailing and Consumer Services*, 78, Article 103778. https://doi.org/10.1016/j.jretconser.2024.103778

[9] Hsu, C.-C., Lee, C.-M., Lin, Y.-F., Chou, Y.-S., Jian, C.-Y., & Tsai, C.-H. (2024). Revisiting vision-language features adaptation and inconsistency for social media popularity prediction. *Proceedings of the 32nd ACM International Conference on Multimedia (MM)*, pp. 11464–11469. https://doi.org/10.1145/3664647.3689000

[10] Zhong, T., Lang, J., Zhang, Y., Cheng, Z., Zhang, K., & Zhou, F. (2024). Predicting micro-video popularity via multi-modal retrieval augmentation. *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*. https://doi.org/10.1145/3626772.3657929

[11] Lu, J., Xiao, M., Wang, W., Du, Y., Wu, Z., & Hua, C. (2025). Multi-modal and metadata capture model for micro video popularity prediction. *arXiv preprint* arXiv:2502.17038.

[12] Ye, L., Zhang, Y., Wu, Y., Chen, Y.-P. P., Yu, J., Yang, W., & Song, Z. (2025). MVP: Winning solution to SMP Challenge 2025 video track. *arXiv preprint* arXiv:2507.00950.

[13] Sun, W., Cao, L., Cao, Y., Zhang, W., Wen, W., Zhang, K., Chen, Z., Lu, F., Min, X., & Zhai, G. (2025). Engagement prediction of short videos with large multimodal models. *arXiv preprint* arXiv:2508.02516.

[14] Mi, S., Zhang, Q., Zhang, S., & Li, G. (2026). Research on short video popularity prediction based on multimodal features: A case study of Douyin platform (in Chinese). *Frontiers of Data & Computing*, 8(1), 183–194. https://doi.org/10.11871/jfdc.issn.2096-742X.2026.01.015

[15] Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, Article 100804. https://doi.org/10.1016/j.patter.2023.100804

[16] Corso, F., Pierri, F., & De Francisci Morales, G. (2024). What we can learn from TikTok through its Research API. *Companion Publication of the 16th ACM Web Science Conference (WebSci Companion)*, pp. 110–114. https://doi.org/10.1145/3630744.3663611

[17] lingbow. (n.d.). *TikTok Video Engagement 200K* [Dataset]. Hugging Face. CC BY-NC 4.0. Public 15,001-post subset redistributed by A. Alshahrani in "TikTok video engagement prediction—day-30 view count forecast," GitHub repository, commit ad9029a. https://huggingface.co/datasets/lingbow/tiktok-video-engagement-200k and https://github.com/Abr036/TikTok_Video_Engagement_Prediction (accessed November 4, 2026)

[18] Chen, Z., Liu, P., Piao, J., Xu, F., & Li, Y. (2024). Shorter is different: Characterizing the dynamics of short-form video platforms. *arXiv preprint* arXiv:2410.16058.

[19] Paekivi, S., & Karjus, A. (2026). Modeling engagement with brand and organizational TikTok videos using machine-assisted theory-ensemble annotation. *arXiv preprint* arXiv:2606.16053.

[20] Mittag, G., Naderi, B., Chehadi, A., & Möller, S. (2021). NISQA: A deep CNN-self-attention model for multidimensional speech quality prediction with crowdsourced datasets. *Proceedings of Interspeech*, pp. 2127–2131. https://doi.org/10.21437/Interspeech.2021-299

[21] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 3146–3154.

[22] Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*.

[23] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

[24] Fryer, D., Strümke, I., & Nguyen, H. (2021). Shapley values for feature selection: The good, the bad, and the axioms. *IEEE Access*, 9, 144352–144360. https://doi.org/10.1109/ACCESS.2021.3119110

[25] TikTok. (n.d.). Research API. TikTok for Developers. https://developers.tiktok.com/products/research-api (accessed November 4, 2026)