

Predicting Continued Visibility of Music Category Videos in South Korean YouTube Trending Lists

Frankie Gao

perspectiveopen.com

Correspondence: zcfbfga@ucl.ac.uk

Submitted: 20 August 2026. Published: 20 August 2026.

Abstract

Public video counters are easy to obtain, but a large cumulative audience does not necessarily imply repeated representation in a regional trending archive. This study asks whether views, interaction rates and completed channel history predict continued visibility of Music category videos in the South Korean YouTube archive. The source contains 1,825 Music rows; excluding 19 rows with corrupted identifiers leaves 801 usable videos from November 2017 to June 2018. Duplicate and coverage-aware reconstruction yields 691 eligible first-appearance windows. The primary outcome is representation on at least three of seven calendar dates, including the first observed appearance. A secondary landmark task predicts next-day archive presence. Chronological development selects a channel-history random forest and a daily momentum random forest without sharing videos between fitting and later evaluation cohorts. On 213 later first appearances, the selected first forest achieves Brier 0.2000 and ROC AUC 0.687, versus 0.2239 and 0.500 for training prevalence. Its Brier improvement over prevalence has a positive paired channel-cluster interval. On 493 daily landmarks from 228 new videos, the forest achieves equal-video Brier 0.2264 and AUC 0.640; its Brier improvement over prevalence is 0.0145 with a 95 percent interval of 0.0057 to 0.0236. A pooled 20 percent monitoring allocation identifies 27 positive videos among 43 selected, versus 16 with view-count ranking; this historical comparison does not establish a causal promotion benefit. Comment and like rates have different conditional coefficient directions, and strict coverage restriction weakens first-appearance gains. The contribution is an auditable two-time-point forecasting and monitoring design, rather than a causal promotional claim or a model of the present YouTube Charts system.

Keywords: YouTube; music category; regional trending archive; continued visibility; engagement rates; chronological validation; monitoring capacity

1 Introduction

A music promotion team often faces a practical question after a video becomes visible on a platform list: should it continue to receive close attention over the next few days, or is the appearance likely to be brief? The answer can influence which clips are included in a daily monitoring report, which audience conversations are examined, and how an analyst allocates limited research time. These are decisions about observing content. They are distinct from deciding whether to buy advertising, estimating commercial returns, or concluding that a specific interaction causes a platform to distribute a video more widely.

Cumulative views alone are an incomplete description of this decision. They combine prior exposure, time since publication and multiple occasions on which a video was watched. A clip can have a substantial worldwide audience before its first recorded appearance in a particular regional archive. Likewise, the same number of likes or comments means something different at 50,000 views than at five million views. A

useful analysis therefore needs to distinguish audience scale from interaction density, and a first observation from a later observation for which a short history has become available.

Research on online popularity has shown that early measurements can contain information about later audience accumulation. Szabo and Huberman analysed the prediction of later content popularity from initial access measurements [5]. Borghol and colleagues studied content-agnostic factors using sets of similar YouTube videos, emphasising that popularity differences are not explained only by the media content [6]. These studies motivate considering early counters and context. They do not establish that a model of cumulative popularity will predict membership in a regional list, where the observed outcome is also shaped by list construction and collection coverage.

The present study narrows the prediction target to what the source can actually support: subsequent representation in a historical archive. The Korean file of the Trending YouTube Video Statistics dataset is suitable for a modest, reproducible investigation because it records dates, video identifiers, category labels and numerical counters [1]. Its Music category is broader than official music videos or Korean pop releases. Treating every category-10 entry as an official K-pop music video would produce an unsupported population claim. The manuscript consequently uses Music category videos as the population label throughout.

Three research questions guide the analysis. First, do interaction rates and completed channel outcomes improve prediction at the first observed appearance beyond views and publication age? Second, does a later daily observation create useful additional information through lifecycle and counter-change features? Third, can the first-appearance scores rank videos more effectively than views alone when an analyst can monitor only a fraction of the cohort? All three questions concern videos already represented in the archive. Predicting whether an arbitrary newly uploaded video will enter a list is a different problem and is not evaluated.

The design adds practical and methodological elements to a simple classifier comparison. It constructs two information sets at explicitly different prediction times, keeps missing collection days out of the outcome definition, builds channel history only from completed earlier follow-up, and evaluates repeated daily observations with equal total weight per video. It also translates predicted probabilities into a monitoring-capacity exercise. This combination supplies an interpretable research workflow and an operational demonstration without claiming a new learning algorithm or a universally superior forecasting method.

2 Methods

2.1 Source provenance and measurement

The preserved Korean CSV and category dictionary were downloaded through the Kaggle dataset API on 2 October 2026. A metadata read on 3 October 2026 confirmed dataset version 115, the CC0 Public Domain licence label and a last-update date of 3 June 2019 [1]. The study preserves the original files, their sizes, retrieval record and SHA-256 digests. The complete Korean CSV contains 34,567 records and 15,876 distinct identifier tokens. One token, #NAME?, is not a usable YouTube identifier and occurs in 443 records across different titles and channels. Category 10 is labelled Music in the supplied category dictionary. Filtering on that numerical category produces 1,825 raw records and 802 distinct tokens, rather than 802 reliably distinct videos.

Identity validation requires a full eleven-character identifier containing letters, digits, underscores or hyphens. Nineteen Music rows contain the invalid token #NAME? and span seven distinct title, channel-name and publication-time combinations. Treating this token as one video would conflate unrelated clips and their follow-up. All nineteen rows are retained in a separate source-audit table and excluded from modelling; no video identity is reconstructed from a title. The usable subset contains 1,806 raw rows and 801 valid video identifiers. Coverage counts likewise exclude invalid identifiers in all categories, leaving 15,875 distinct valid IDs in the full archive.

The original collector repository describes the regional collection workflow [2]. Its request uses the YouTube Data API mostPopular chart with a regionCode parameter, which determines the regional chart request [3]. The video statistics object supplies video-level counters such as views, likes and comments [4]. Accordingly, a Korean list entry and its cumulative counters are different measurements: the list is regionally requested, while the numerical counters in these records are not a measurement of Korean-only viewers. They also do not measure unique fans, purchases, streaming subscriptions, or campaign-attributable conversions.

Trending dates use the source format year.day.month and were parsed explicitly, rather than inferred by an automatic date parser. The archive spans 14 November 2017 to 14 June 2018, or 213 calendar dates. Records exist for 205 dates. No regional rows exist on 10 and 11 January or on 8 to 13 April 2018. Missing regional dates were preserved as unobserved, not converted to a day on which every video exited. This distinction matters both for seven-day labels and for the one-day landmark outcome.

Daily coverage is also uneven on dates that have records. The archive contains 51 to 192 unique valid-ID videos per observed date, with a median of 158. A recorded date is therefore evidence of a collection snapshot, not proof that every eligible regional item was captured. Figure 1 makes this limitation visible before any model results are presented. The study predicts representation in this archive; an absent record cannot independently establish removal from the complete platform list. Stricter daily-coverage restrictions are used as sensitivity analyses, rather than assuming that all observed dates are equally complete.

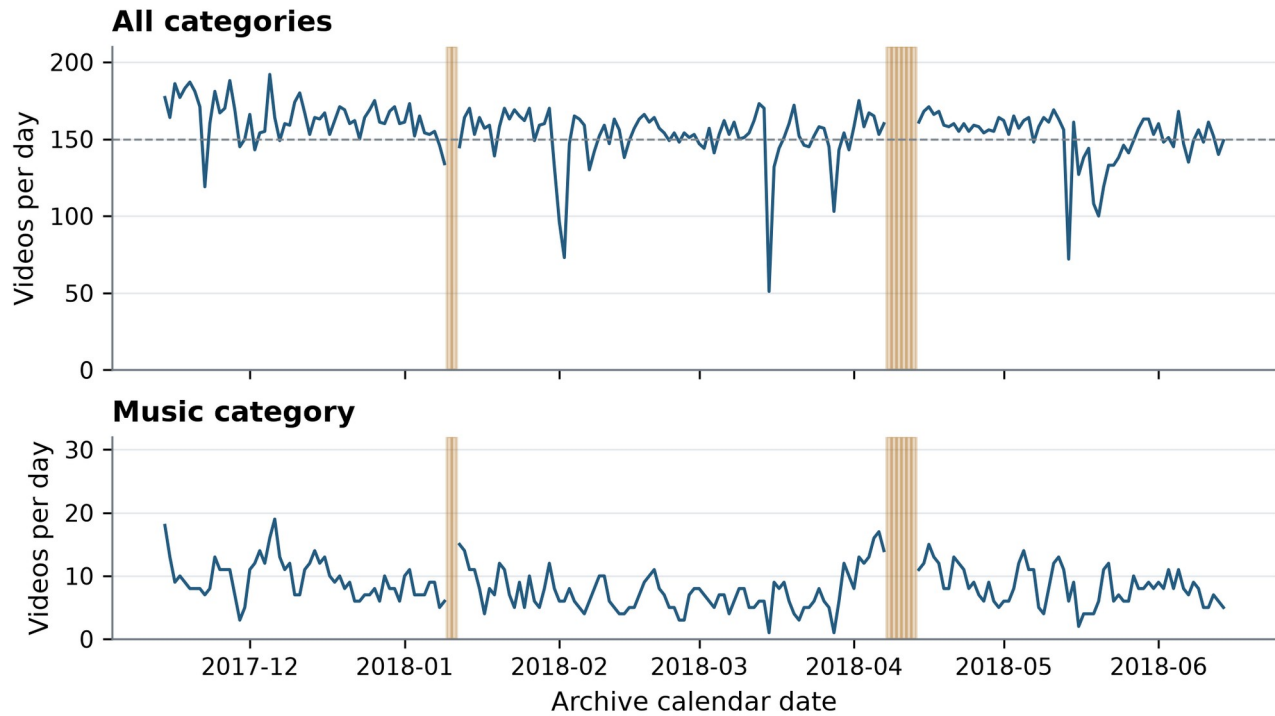


Figure 1. Regional archive coverage. Daily unique valid-ID video counts in the complete Korean archive and the Music subset, 14 November 2017 to 14 June 2018. Invalid source identifier tokens are excluded from counts. Gold bands mark eight dates without any regional records. Missing dates remain gaps, rather than zero observations. Even recorded dates contain only 51 to 192 valid video IDs, so coverage is not demonstrably complete.

2.2 Duplicate handling and cohort reconstruction

After identity validation, the Music subset contains 101 redundant rows when video identifier and trending date are used as the unit. For each video-date pair, the primary reconstruction retains the record with the highest view count, breaking ties with its original input position. This rule leaves 1,705 unique video-days. Eleven duplicated groups contain conflicting view, like or comment counters. The raw file remains unchanged, and retaining the first source row is evaluated separately. The primary rule is a transparent resolution of repeated daily records; it does not recover the unrecorded intraday observation time.

The first appearance is the earliest date on which a video identifier occurs in this archive. It is not necessarily the video's first-ever platform-list appearance. To reduce left-edge truncation, videos first recorded during the opening seven archive dates are excluded from the primary and daily cohorts. The primary cohort additionally requires the complete seven-calendar-day follow-up to fall inside the archive and every regional follow-up date to be observed. Videos with a negative day-level publication age would also be excluded, although no further primary exclusions arise from that condition.

Applying exclusions in a fixed sequence removes 38 opening-week videos, 17 videos without complete ending follow-up and 55 videos whose seven-day window intersects a missing regional date. The resulting 691 eligible primary videos include 226 positives. Of these, 467 have completed follow-up before the final-cohort boundary, 213 first appear on or after that boundary, and 11 cross the boundary and are embargoed from model fitting and evaluation. The exclusions prevent the easiest forms of incomplete-outcome coding; they do not resolve partial coverage inside recorded snapshots.

Table 1. Data reconstruction and analysis denominators

Stage	Records or videos	Interpretation
Complete Korean CSV	34,567 rows	All source categories
Raw Music subset	1,825 rows; 802 ID tokens	Category ID 10; includes invalid token
Invalid Music identity excluded	19 rows	Corrupted token spans 7 metadata groups
Usable Music source	1,806 rows; 801 valid IDs	No inferred replacement identifiers
Deduplicated Music observations	1,705 video-days	101 redundant usable rows removed
Opening archive week excluded	38 videos	First seven archive dates
Ending follow-up excluded	17 videos	Incomplete seven-day horizon
Missing-date windows excluded	55 videos	No assumed outcomes on absent dates
Eligible first-appearance cohort	691 videos	467 fit, 11 embargo, 213 later evaluation
Eligible daily landmarks	1,601 rows	1,087 fit, 21 excluded, 493 later evaluation

Exclusion counts are sequential and mutually exclusive for the primary cohort. Daily landmarks use a separate next-day eligibility rule. Source: preserved Korean archive and local reconstruction.

2.3 Two prediction times and outcomes

The primary prediction is made after the counters in a video's first recorded snapshot are available. Let A denote an indicator that a video appears on a specified archive date and t denote its first observed date. The primary outcome is one when the video is represented on at least three distinct dates during t through t plus six days. The first date is included, so the target requires at least two further observed dates. Re-entry after an unrecorded day counts toward this total. This is a thresholded measure of continued visibility, rather than survival time until an irreversible exit.

$$Y_{i,7} = 1\{R_{i,7} \geq 3\}, R_{i,7} = A_{i,t} + A_{i,t+1} + \dots + A_{i,t+6} \quad (1)$$

The secondary prediction is made at every eligible observed daily landmark, using information available at that snapshot. Its outcome is archive presence on the next calendar date. A landmark is used only when the next regional snapshot exists. The test cohort contains only videos whose first observed appearance is on or after 15 April 2018, so a long-running development video cannot reappear as a test observation. These rules produce 1,087 development landmarks from 514 videos and 493 later landmarks from 228 new videos; a further 21 landmarks are excluded at the fitting boundary or because they belong to earlier-entry videos after that boundary.

$$Z_{i,t} = A_{i,t+1} \quad (2)$$

Figure 2 distinguishes the two questions and their denominators. Some videos eligible for next-day prediction are ineligible for seven-day prediction because their later follow-up is incomplete. Conversely, a primary video contributes only its first snapshot to the seven-day task, while it can contribute multiple landmarks to the daily task. Performance values from the two tasks must therefore be interpreted against

their own outcome frequency, horizon and evaluation unit, rather than treated as scores on the same classification problem.

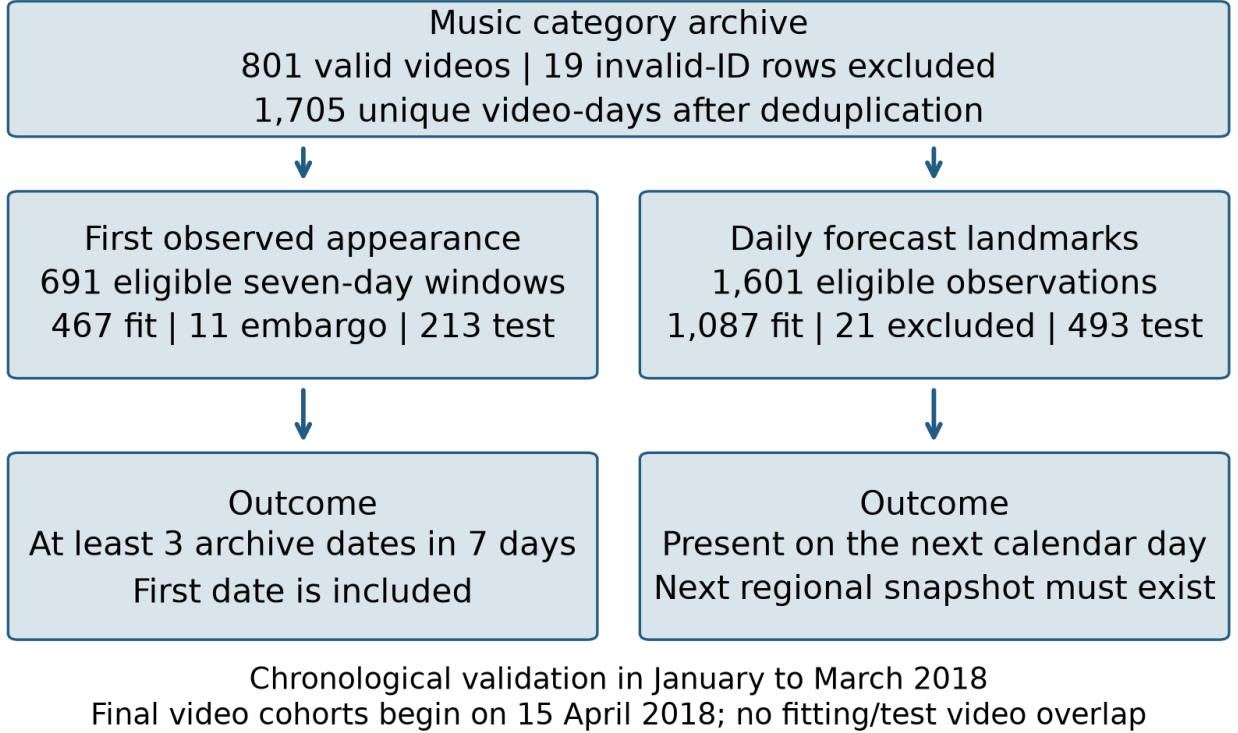


Figure 2. Two prediction times and two outcomes. The first-appearance task estimates seven-day continued archive visibility. The daily task estimates the next recorded calendar day and excludes carry-over videos from the later test cohort. Eligibility and split counts have different denominators. Channel history uses only completed earlier seven-day outcomes; current-video future appearances are never predictors.

2.4 Predictors and the information boundary

The simplest learned first-appearance comparator uses the natural logarithm of one plus cumulative views, the natural logarithm of one plus day-level publication age, and sine and cosine encodings of the archive weekday. Publication age subtracts the UTC publication calendar date from the trending date. The source does not establish the collector timezone precisely, so this variable is a day-level approximation, not exact hours since release. Weekday encodings add a small calendar context without an ordinal assumption that Sunday is farther from Monday than from Saturday.

Interaction predictors are likes and comments per 1,000 cumulative views, transformed with log one plus the rate. In Equation 3, L denotes likes, C comments and V views. No video has zero views, and numerical counters are nonnegative. When a counter is marked disabled, its rate is treated as missing and its disable flag is retained. The imputer then learns the relevant training median. A zero rate from an enabled counter remains zero. This keeps unavailable feedback distinct from observed zero feedback. The removal or error flag is retained only for data-quality filtering because the reprocessed archive does not verify its contemporaneous availability; it is excluded from every predictive feature list.

$$r_{like} = \frac{1000 L}{V}; r_{comment} = \frac{1000 C}{V} \quad (3)$$

Channel context is calculated without directly encoding channel or video identity. At each prediction date, only eligible earlier primary videos whose full seven-day follow-up has ended strictly before that date are used. A Beta one-one smoothed global outcome rate, denoted g in Equation 4, supplies the prior. For a recorded channel name with n matured videos and s positives, its smoothed rate is (s plus five times the global prior) divided by (n plus five). The model receives this rate and log one plus n . An unseen channel therefore begins at the contemporaneous global prior, instead of being assigned an extreme success rate based on a single observation.

$$q_{channel} = \frac{s + 5g}{n + 5} \quad (4)$$

During later evaluation, completed earlier test-cohort outcomes can enter the history state only after their seven-day window has closed. This is chronological information updating, not model refitting or use of an outcome before it is known. Model coefficients, trees and imputation parameters remain fixed. The protocol uses exact video identifiers to separate fitting and evaluation videos, but channel names may recur. Channel-name overlap is therefore assessed explicitly, and name changes are recognised as a limitation of the grouping variable.

Daily models add log one plus days since first observation, consecutive recorded-day streak and number of observed dates so far. Momentum comprises an indicator that the immediately preceding calendar date is recorded for that video and changes in log one plus views, likes and comments since that date. A change is missing when the earlier record is separated by more than one day, rather than mislabelled as one-day growth. Disabled current or preceding interaction counters also make the corresponding change missing. Negative changes remain valid recorded differences in the primary analysis and are removed in a separate sensitivity check.

Table 2. Predictor groups and availability rules

Group	Variables	Timing or interpretation
Scale and calendar	Log views, log age, weekday sine and cosine	Available at the current archived snapshot
Interaction density	Log likes and comments per 1,000 views; two disable flags	Missing rates imputed from training only
Matured channel history	Log earlier entry count; smoothed earlier outcome rate	Only completed earlier seven-day outcomes
Observed lifecycle	Log chart age, streak and observed-date count	Computed from current and preceding records
Recorded momentum	Previous-day flag; three log-counter changes	Exact preceding calendar date only
Excluded from prediction	Removal flag, identifiers, future date totals and outcomes	Metadata retained for audit, never model inputs

Views and interaction counters are global video statistics in a regionally requested list. Lifecycle and momentum are used only in daily specifications.

2 5 Chronological development and model fitting

The final-cohort boundary is 15 April 2018. Development validation uses January, February and March 2018 as three expanding chronological blocks. In every fit, all training target horizons end strictly before the next validation start. Primary validation is indexed by first prediction date. Daily validation uses videos first appearing in that month and landmarks whose next-day outcome ends within it. No video identifier is shared between a training fold and its validation cohort. The latest final-fitting target date is 7 April, and the earliest final prediction date is 15 April.

Logistic regression provides a simple penalised probability model; random forests provide a flexible comparator [9]. Each logistic specification searches inverse penalty strengths 0.1, 1 and 10. Each forest uses 300 trees and searches minimum leaf sizes 5 and 15 with either the square-root feature count or all features available at a split. This yields 30 specification-parameter candidates and 90 validation fits. The main seed is 42. Five forest seeds are assessed as a stability diagnostic without selecting a replacement seed on the final results.

Median imputation is fitted only on the relevant training observations. Logistic inputs are standardised using those imputed training observations; forest inputs are not standardised. Every video contributes total fitting weight one, with weights rescaled to mean one for estimation. There is no class oversampling or class-weight balancing. For each specification, hyperparameters minimise the pooled validation Brier score. The best learned specification for each task is then selected by that same score, while the prevalence baseline remains an explicit comparator. The selected models are the channel-history forest with minimum leaf size 5 and the daily momentum forest with minimum leaf size 15, both using square-root feature sampling. The channel-history logistic with C equal to 1 is retained as an interpretable comparator.

This is a retrospective analysis of an archive that had already been inspected during feasibility work. The final configuration was written before final-cohort fitting and scoring in the corrected run, but this does not make the study preregistered or the historical data prospectively untouched. The final corrected run excludes removal status after the availability audit and all invalid-ID records after source-identity inspection. The complete development search is rerun on the corrected cohort; results from the earlier mixed-identity run are not reported as final evidence. Sensitivities and additional phase diagnostics are labelled as such and do not replace the development-selected primary specifications. The chronological design addresses concrete temporal and identity leakage risks; it cannot eliminate every source of retrospective design choice [7,8].

2 6 Evaluation and uncertainty

The primary probability score is the Brier score, the weighted mean squared difference between a binary outcome and its predicted probability. A lower value indicates less probability error. Log loss, ROC AUC and average precision provide complementary information. Proper probability scores motivate evaluating the probabilities themselves rather than reporting only thresholded correctness [10]. Average precision is included because ranking quality and outcome prevalence matter when only a limited number of videos will be reviewed [11]. Accuracy and balanced accuracy at probability 0.5 are retained in the results archive but are not the selection criterion.

$$Brier = \frac{\sum w(y - p)^2}{\sum w} \quad (5)$$

For daily observations, the weight for a row equals one divided by that video's number of eligible evaluation landmarks. A video appearing five times therefore contributes the same total weight as a video appearing once. This estimates performance when videos receive equal importance and their observed landmarks are averaged within video. Equal-row scores answer a different question by placing more importance on videos with more recorded dates. Both are reported. Conditional subgroup analyses recompute video weights within the subgroup, so their outcome rates cannot be directly averaged into the global weighted rate.

Paired uncertainty calculations resample whole channel-name clusters 2,000 times and keep each model and its comparator attached to the same sampled observations. Improvement is defined as comparator Brier minus model Brier. Reported 95 percent intervals are the 2.5th and 97.5th percentiles of that distribution. Calendar dependence is assessed separately with circular seven-day and fourteen-day date blocks. These resamples condition on the fitted models; they do not rerun development selection or fully represent uncertainty across alternative model-development pipelines. Channel and date intervals are separate diagnostics rather than a combined two-way cluster procedure.

The monitoring exercise ranks the 213 final primary videos by their first-snapshot score and selects the ceiling of a fixed capacity fraction times the cohort size. The central setting is 20 percent, giving 43 selections. View-count, like-rate and comment-rate ranking supply simple comparators. Precision is the positive fraction among selected videos; capture is the fraction of all 72 positive videos selected. A paired channel bootstrap recomputes each ranking and capacity in each resample. The main exercise pools the final cohort retrospectively. A separately labelled daily-batch diagnostic restricts competition to videos observed on the same calendar day, which better exposes the integer cost of sparse daily queues.

3 Results

3 1 Cohort structure and descriptive patterns

Among 691 complete primary windows, 226 videos are recorded on at least three of the first seven dates, an outcome rate of 32.7 percent. Figure 3 shows that the outcome is concentrated at the shorter end of the first-week distribution. The development cohort has 152 positives among 467 videos; the later cohort has 72 among 213, or 33.8 percent. Outcome frequencies are therefore comparable at the aggregate first-appearance level, although composition and regional coverage can still vary by date and channel.

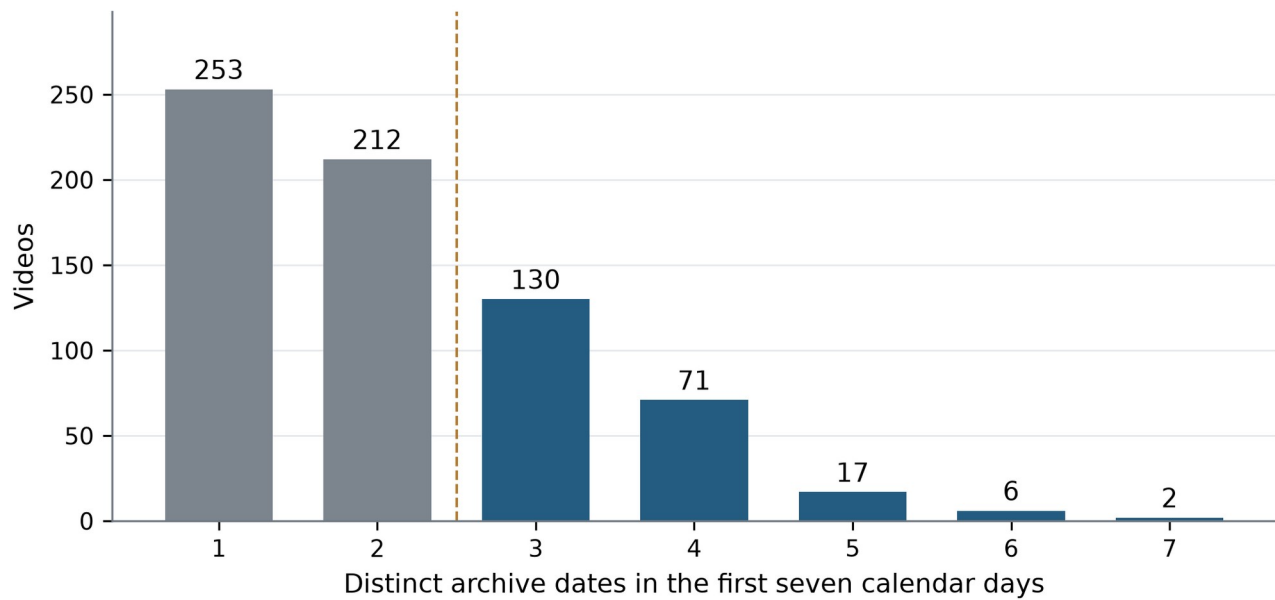


Figure 3. Observed seven day visibility. Distribution across 691 eligible valid-ID videos. The dashed boundary defines the primary positive outcome: at least three recorded dates, including the first. Reappearances count; the figure does not establish continuous ranking. There are 226 positive videos overall, including development, cutoff-embargo and final evaluation cohorts.

Across eligible primary videos, the median first-snapshot view count is 220,390 for positive outcomes and 103,527 for negative outcomes. The corresponding medians are 5.37 versus 4.23 comments per 1,000 views, but 48.65 versus 50.41 likes per 1,000 views. Publication age has a median of 1 day in the positive group and 1 day in the negative group. These descriptive summaries overlap substantially; they do not demonstrate separable classes or adjusted effects. They suggest why interaction composition and audience scale should be examined separately instead of merged into a single engagement total.

The usable Music snapshots and first-appearance records both contain 354 recorded channel names. The unit for deduplication and cohort separation is a validated source video identifier, rather than a title or channel name. Channel names are not verified stable account identifiers, so channel-history quantities and cluster intervals inherit that measurement limitation. The invalid-ID exclusion also shows why counting distinct text tokens alone is insufficient to establish video identity.

3 2 First appearance prediction

The selected first-appearance forest has pooled development Brier 0.1970 and later-cohort Brier 0.2000. Its later ROC AUC is 0.687, average precision 0.551 and mean probability 34.5% against a 33.8% observed outcome rate. Training prevalence gives Brier 0.2239, and the views-and-age logistic comparator gives 0.2206. The selected-model point reduction relative to prevalence is 10.7%, a modest rather than decisive reduction in probability error.

The engagement-only forest gives later first-appearance Brier 0.2003, very close to the selected channel-history forest at 0.2000. The latter has the lowest pooled development Brier among the declared specifications. This is a selection based on earlier chronological validation, rather than a choice of whichever model looks most favourable on the later cohort. The near-identical later forest scores do not establish an independent predictive benefit from channel history.

Table 3. Chronological development and later cohort model performance

Task and model	Dev Brier	Later Brier	ROC AUC	Average precision
First Training prevalence	0.2155	0.2239	0.500	0.338
First LR views and age	0.2113	0.2206	0.572	0.420
First LR engagement	0.2121	0.2092	0.642	0.496
First RF engagement	0.2041	0.2003	0.683	0.574
First LR channel history	0.2001	0.2070	0.658	0.511
First RF channel history	0.1970	0.2000	0.687	0.551
Daily Training prevalence	0.2368	0.2409	0.500	0.403
Daily LR lifecycle	0.2316	0.2354	0.589	0.462
Daily LR static engagement	0.2353	0.2301	0.617	0.520
Daily LR momentum	0.2323	0.2317	0.622	0.497
Daily RF momentum	0.2303	0.2264	0.640	0.540

First task: 213 later videos and 72 positives. Daily task: 493 later rows from 228 videos; equal-video outcome rate 40.3 percent. All final parameters are selected using development observations; RF channel history is the selected first model; the logistic and engagement-only forest remain comparators.

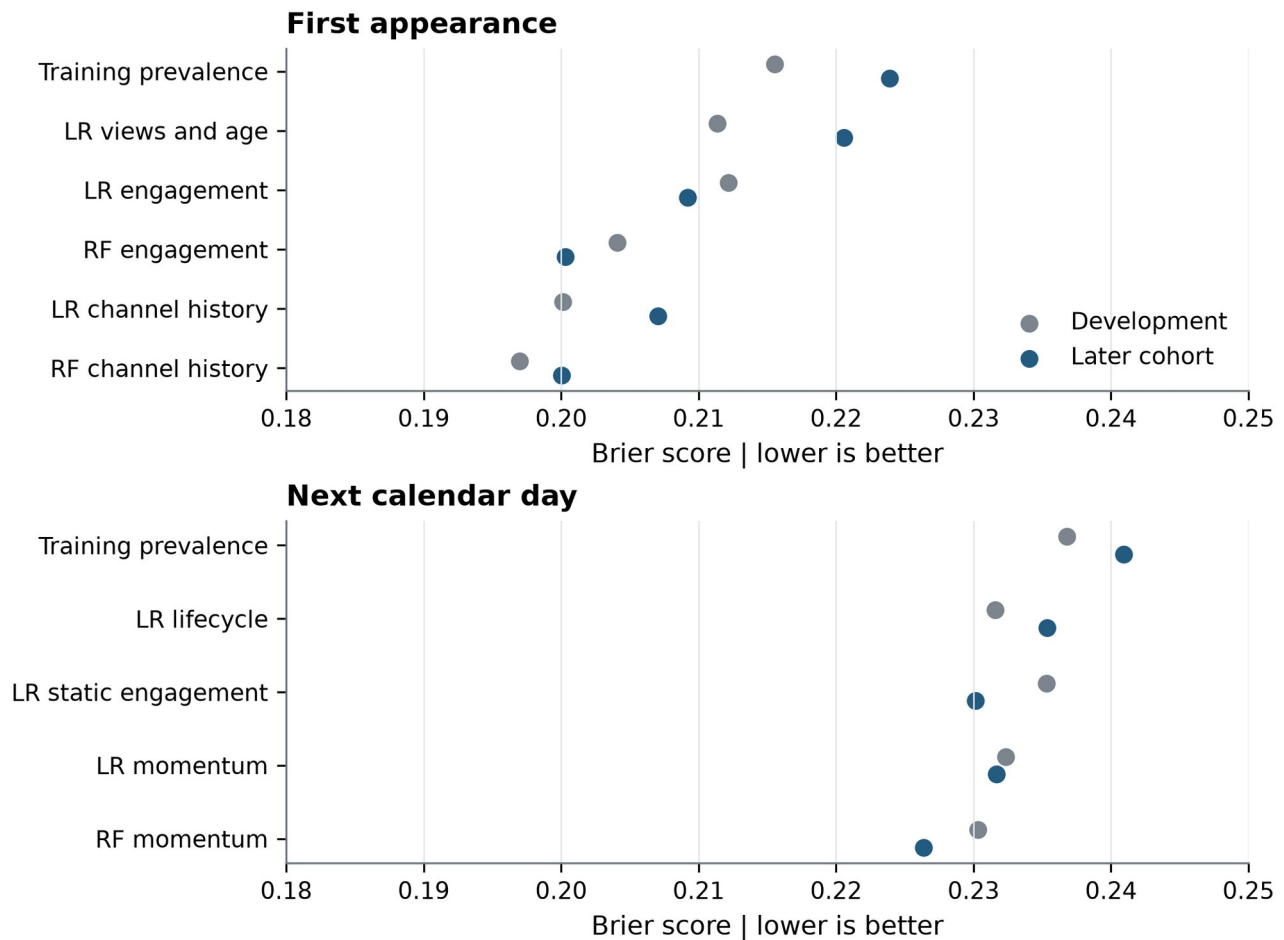


Figure 4. Development and later cohort probability errors. Each learned specification is tuned using three chronological monthly validation blocks and assessed on videos first observed from 15 April 2018. The development-selected learned models are RF channel history and RF momentum. Scores give every video one total unit of weight. Development and later-cohort values are descriptive comparisons across different cohorts, not paired significance tests.

Channel-cluster resampling gives a selected-first-forest Brier reduction of 0.0239, with a 95 percent interval from 0.0063 to 0.0421 against prevalence. Against views and age, the reduction is 0.0206, with interval 0.0016 to 0.0391. Both intervals are positive conditional on the fitted predictions. The logistic channel-history comparator has Brier 0.2070 and a prevalence-comparison interval that includes zero. These differing results discourage treating either the learning family or history variables as universally superior. Ranking discrimination and probability accuracy remain separate evidence.

3 3 Daily prediction and the size of the model gain

The 493 later daily rows contain 265 positive next-day outcomes, an equal-row rate of 53.8 percent. Giving each of the 228 videos one total unit changes the outcome rate to 40.3 percent. The selected daily forest achieves Brier 0.2264, AUC 0.640 and average precision 0.540 under the latter estimand. Prevalence has Brier 0.2409. The relative point improvement is 6.0%. The lifecycle logistic, static-engagement logistic and momentum logistic scores remain close, so the result supports a limited gain rather than an unusually powerful forecast.

The forest Brier reduction against prevalence is 0.0145, with a channel-cluster 95 percent interval of 0.0057 to 0.0236. Against lifecycle logistic, the interval is 0.0001 to 0.0178. Figure 5 places these comparisons beside the first-appearance forest result. Calendar-block sensitivities are positive against prevalence for both selected models, but they do not eliminate retrospective selection, cross-channel dependence or coverage uncertainty.

Table 4. Selected model paired channel cluster Brier improvements

Task	Comparator	Reduction	95 percent lower	95 percent upper
onset	Training prevalence	0.0239	0.0063	0.0421
onset	LR views and age	0.0206	0.0016	0.0391
daily	Training prevalence	0.0145	0.0057	0.0236
daily	LR lifecycle	0.0090	0.0001	0.0178

Positive reductions favour the selected learned model. There are 121 first-appearance and 127 daily channel-name clusters; 2,000 paired resamples. Intervals condition on fitted models and are not adjusted for the complete exploratory comparison set.

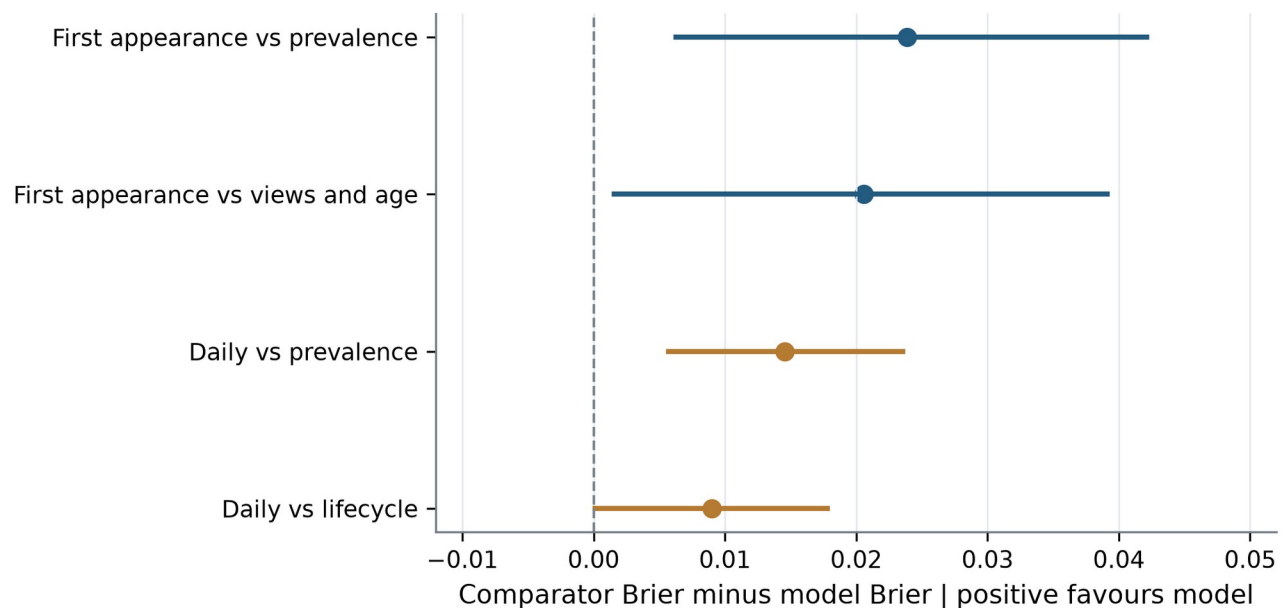


Figure 5. Paired uncertainty in probability score improvements. Points are observed Brier-score reductions; horizontal intervals are 95 percent percentile intervals from 2,000 paired channel-cluster resamples of the final cohorts. The selected first-appearance forest intervals against prevalence and views-and-age logistic, and daily forest intervals against prevalence and lifecycle logistic, are positive conditional on the fitted predictions. Intervals condition on fitted models and do not correct for all exploratory comparisons.

3 4 Interaction composition and model interpretation

The channel-history logistic comparator assigns the log like-rate predictor a standardised coefficient of -0.624 and the log comment-rate predictor a coefficient of 0.581. These opposite directions remain conditional on views, publication age, calendar encoding, counter availability and matured channel history. This comparator is retained for interpretation and is not an explanation of the selected forest. Figure 6

shows the continuous coefficients on the same training-standard-deviation scale. The signs should not be read as a recommendation to suppress likes or manufacture comments.

A ratio can be high because its numerator is large or because its denominator is relatively small. The already-visible cohort can also induce a conditional relationship different from the relationship among all uploaded videos. A clip with an intensely responsive existing fan community may have a high like density at relatively limited broad exposure, while discussion and viewing breadth may jointly track repeated visibility. These are plausible interpretations, not mechanisms established by this dataset. Neither sentiment, independent viewers, watch time nor promotional exposure is observed, so positive comments and controversial comments cannot be distinguished.

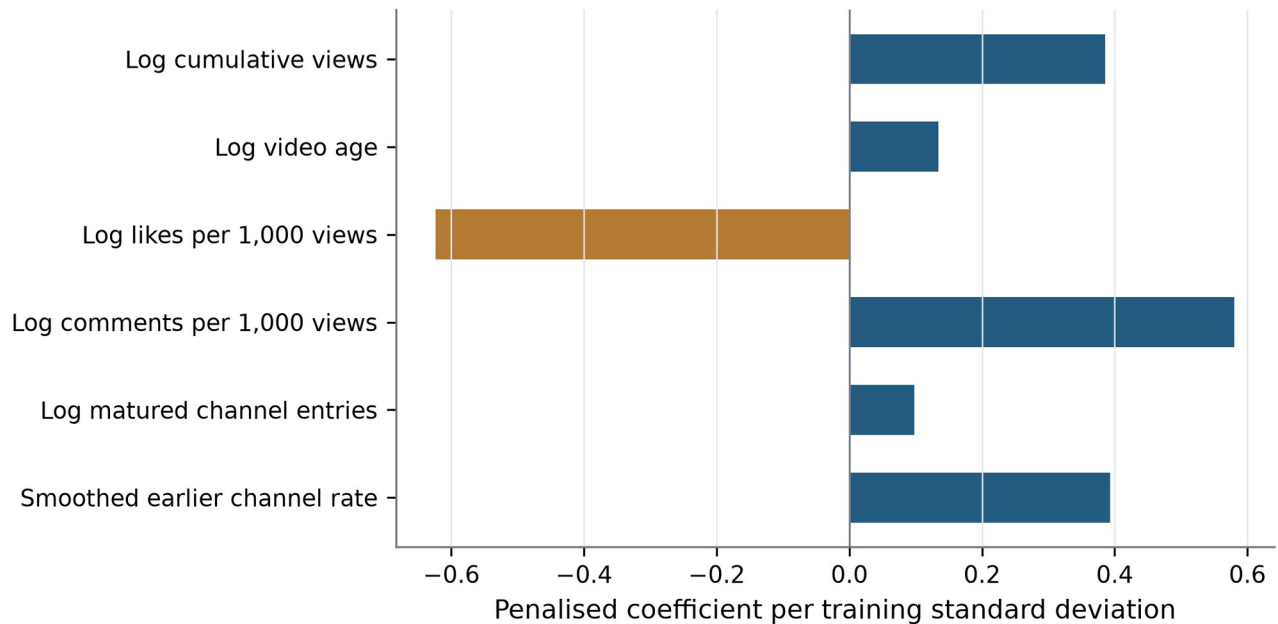


Figure 6. Different conditional directions for likes and comments. Continuous-feature coefficients from the development-fitted first-appearance channel-history logistic comparator, with C equal to 1. This comparator is not the selected forest. All displayed inputs are standardised after training-only median imputation. The negative like-rate and positive comment-rate coefficients describe conditional predictive associations among already observed Music videos. They are not causal effects, tests of statistical significance, or evidence that generating comments improves visibility.

The fixed-parameter removals give a more useful predictive check than interpreting a coefficient or forest importance as causal evidence. Removing engagement inputs from the selected first forest worsens later Brier; removing channel history has only a very small effect, so history is not shown to supply a large independent gain. The full daily model, however, improves slightly when static engagement rates are removed. Removing recorded momentum changes daily Brier only modestly. This contrast indicates that the value of an input group depends on the task, other features and estimator. It also limits any claim that the forest advantage is specifically explained by one-day growth.

3 5 Probability calibration and observation phase

Figure 7 compares mean predictions with observed outcomes in five bins. First-appearance forest probabilities provide a graded risk ordering, with imperfect agreement between intermediate predictions and observed risk. The daily forest is reasonably close to the diagonal at an aggregate level, with visible local deviations. These bins are descriptive checks with modest counts; a connected line is not evidence

that each point estimates a smooth population curve. The final outcomes are not used to recalibrate either model.

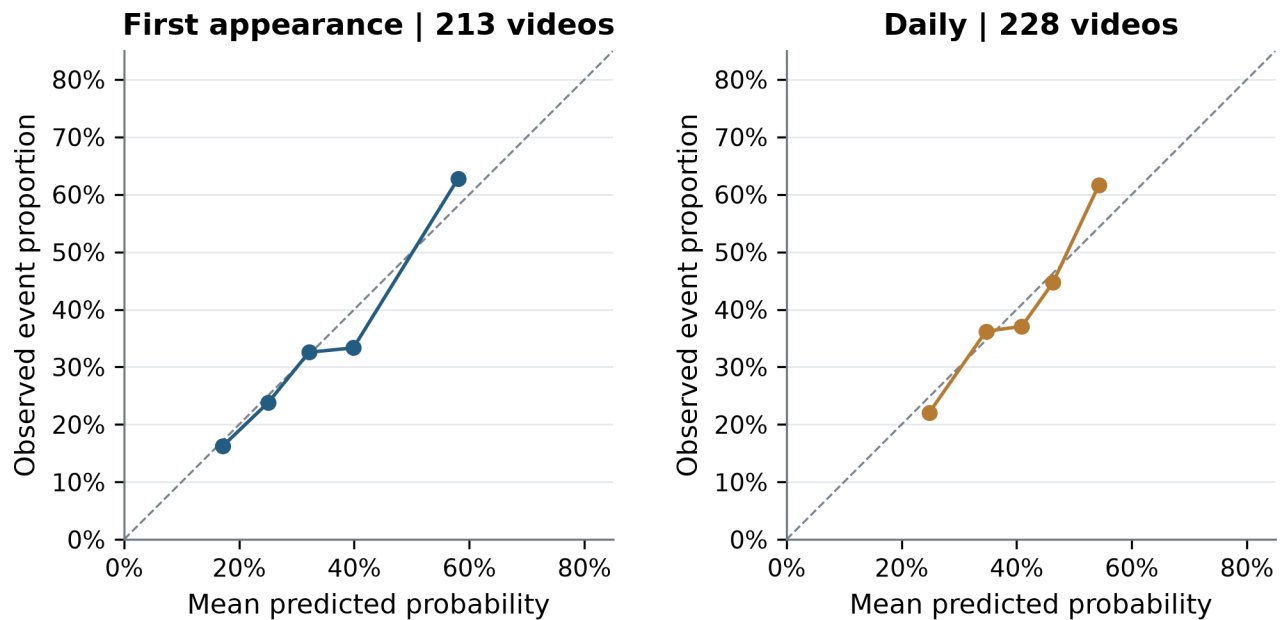


Figure 7. Probability calibration in the later cohorts. Five approximately equal-row prediction bins for the selected learned models. Daily event proportions and mean probabilities use the original equal-video weights within each bin; bin sizes are 98 to 99 rows and videos may span bins. First-appearance bins contain 42 to 43 distinct videos. Lines guide the eye and are not fitted calibration curves. No recalibration uses the final outcomes.

The daily task also has a phase-specific weakness obscured by an aggregate calibration plot. Evaluating first observations separately gives a next-day event rate of 66.2 percent among 228 videos, while evaluating later observations separately gives an equal-video rate of 28.1 percent among 150 videos. The selected forest understates the first-appearance rate and overstates the later rate. The same video can occur in both groups, and weights are recalculated within each, so these percentages cannot be averaged directly into the overall 40.3 percent rate.

A post-result diagnostic uses two training-only conditional prevalence rates, one for first observations and one for later observations, estimated with the original equal-video training weights. It gives final overall Brier 0.2394, compared with 0.2409 for global prevalence and 0.2264 for the forest. This modest improvement over a single prior does not account for the full forest gain. The diagnostic is deliberately kept separate from the development selection. It illustrates why a phase-specific baseline is useful while avoiding a new final-cohort-optimised classifier. The weights used to estimate its phase rates target the global landmark estimand, whereas the separately reported phase evaluation gives videos equal importance within that phase.

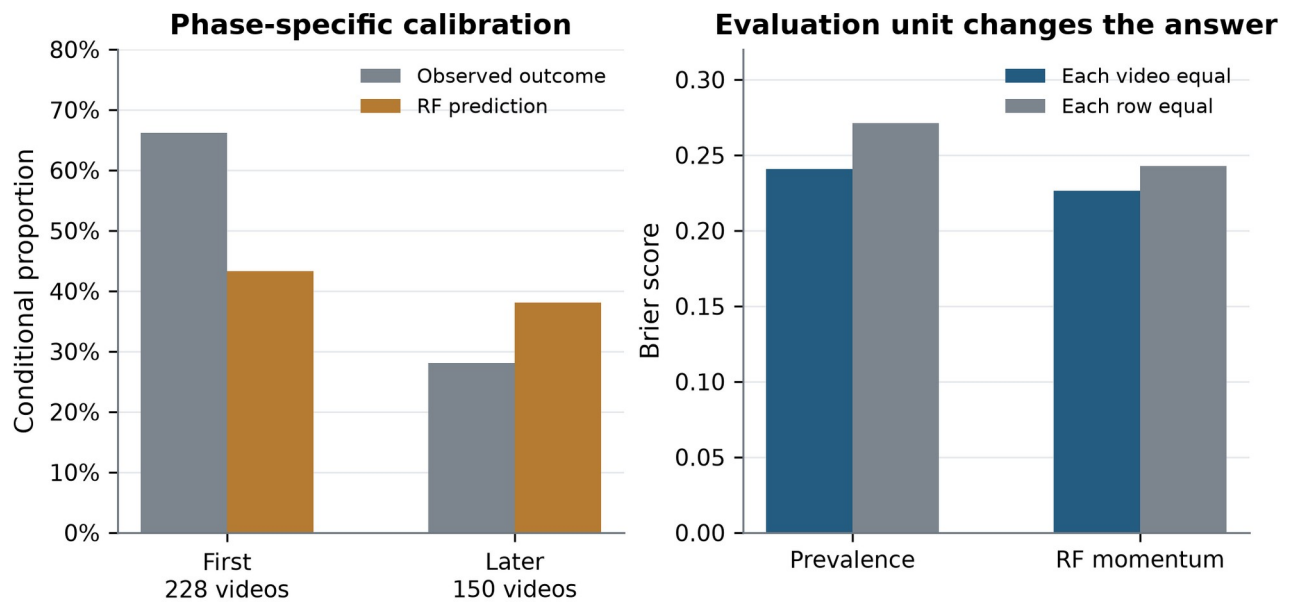


Figure 8. Observation phase and the choice of evaluation unit. Left: first and later appearances evaluated separately, assigning each video one unit within its phase. The same video can contribute to both phases. RF predictions understate the first-appearance event rate and overstate the later rate. These conditional subgroup rates do not average directly to the overall rate because weights are recomputed within subgroups. Right: equal-video and equal-row scores for the identical 493 final landmarks; the estimands differ.

3.6 Monitoring under limited capacity

At the pooled 20 percent setting, the selected first-appearance model finds 27 positive outcomes among 43 chosen videos, giving precision 62.8% and capturing 37.5 percent of the 72 positives. View-count ranking finds 16 positives among the same number of selections, precision 37.2% and capture 22.2 percent. Comment-rate ranking supplies a useful transparent comparison, while ranking solely by like density does not establish a consistent advantage. The score is intended to identify a queue for closer observation, not to determine advertising spend or to label lower-ranked clips as unworthy of support.

The model-minus-views precision difference is 25.6 percentage points. Its paired channel-bootstrap interval is 2.3 to 42.5 percentage points. Its positive lower endpoint supports a better allocation within this historical cohort conditional on the scores and clustering scheme; it does not establish assured superiority in a future queue or a causal intervention benefit. Figure 9 reports the full capacity profile to show that the choice of 20 percent is not the entire story. Small-capacity points have especially limited denominators.

Table 5. Pooled monitoring at the 20 percent setting

Ranking strategy	Selected	Positive	Precision	Positive capture
View count	43	16	37.2%	22.2%
Like rate	43	16	37.2%	22.2%
Comment rate	43	20	46.5%	27.8%
LR views and age	43	17	39.5%	23.6%
RF engagement	43	26	60.5%	36.1%

Ranking strategy	Selected	Positive	Precision	Positive capture
RF channel history	43	27	62.8%	37.5%

The development-selected first model is RF channel history. Strategies use 43 of 213 videos. Pooled ranking is a retrospective cohort exercise; it is not a record of an implemented campaign. Binomial intervals in the machine-readable capacity table are not used as the primary inferential evidence.

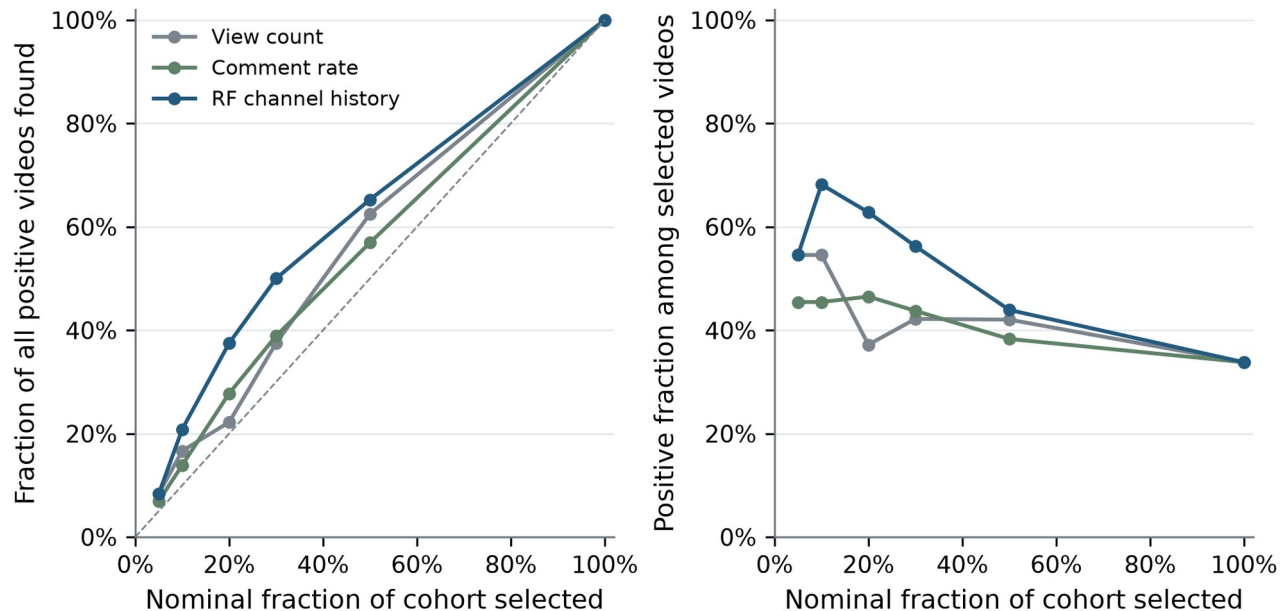


Figure 9. Capacity limited monitoring in the final first appearance cohort. Retrospective pooled-cohort ranking across 213 videos, with 72 positive outcomes. At the pre-specified 20 percent setting, rounding selects 43 videos: the selected model finds 27 positives and view-count ranking finds 16. This is a hypothetical allocation of monitoring effort across the whole cohort, not a chronological daily intervention. Points at other capacities are descriptive sensitivity checks.

The daily-batch diagnostic is less favourable and more operationally realistic. Across 52 dates with eligible new videos, rounding up 20 percent within each daily queue selects 60 videos, or 28.2 percent of the full cohort. The selected model finds 29 positives, compared with 24 for views. Their precisions are 48.3% and 40.0%. A nominal percentage does not therefore map to a fixed total workload in a sparse daily queue. These untested daily-policy differences are descriptive and are not presented as a second confirmed intervention benefit.

3.7 Coverage and specification sensitivity

The first-appearance conclusion changes with collection assumptions. Requiring at least 100 regional videos on every follow-up date leaves 184 final videos and preserves a point gain over prevalence. Requiring at least 150 on every date leaves only 66 and makes the selected forest score worse than prevalence. Figure 10 shows the direction reversal directly. This can reflect altered video composition, reduced fitting support, collection artefacts or ordinary variation. The archive does not identify which explanation dominates. The strict subset is therefore evidence that apparent robustness should not be overstated.

Table 6. First appearance sensitivity comparisons

Variant	Later videos	Model Brier	Prior Brier	Reduction
At least 2 appearances in 7 days	213	0.2051	0.2244	0.0192
At least 4 appearances in 7 days	213	0.0977	0.1111	0.0134
Enabled counters and no removal flag	206	0.1984	0.2211	0.0228
All 7 dates have at least 100 regional videos	184	0.2112	0.2362	0.0250
All 7 dates have at least 150 regional videos	66	0.2278	0.2171	-0.0107
Exclude 5 largest development channels	182	0.1972	0.2222	0.0250
14 day opening washout	213	0.2022	0.2239	0.0217
First record duplicate resolution	213	0.2000	0.2239	0.0239

Each sensitivity refits the selected forest parameters without a final-cohort parameter search. History retains the primary at-least-three-date definition and original matured source population. Different outcomes or subsets change base rates and cannot be compared through raw scores alone.

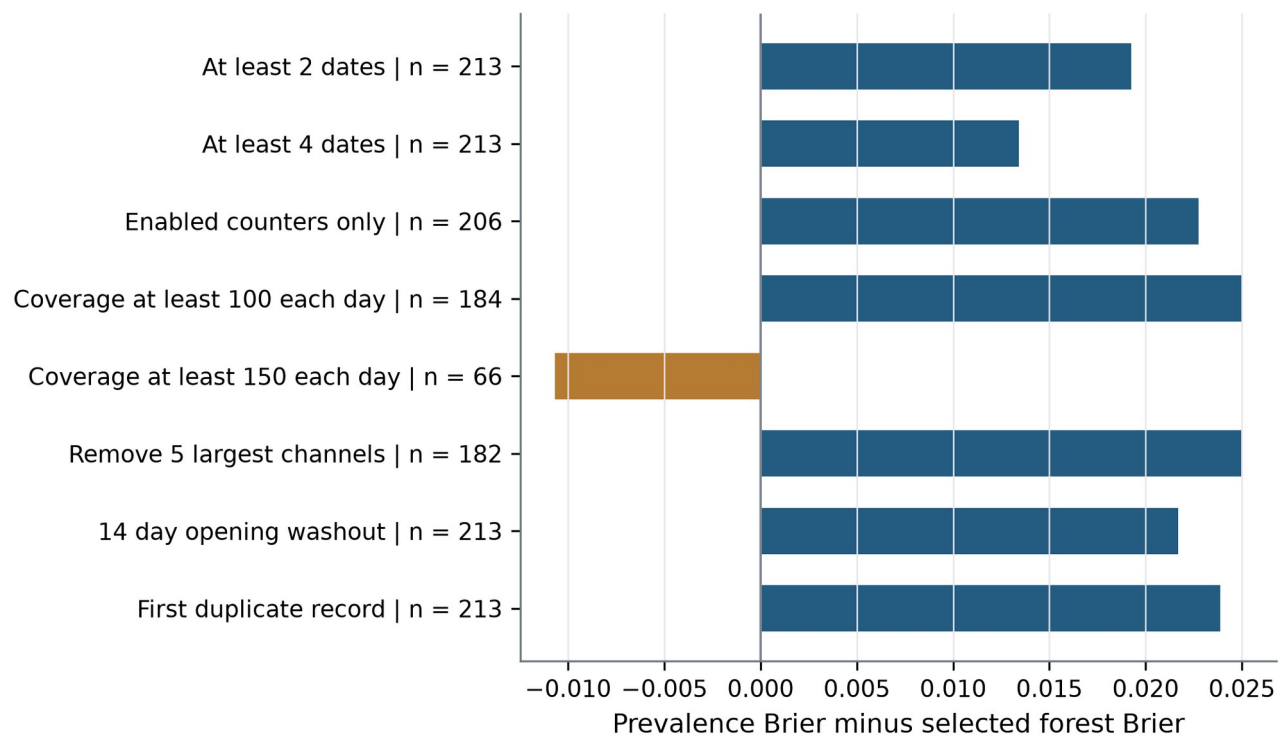


Figure 10. Sensitivity to outcome and collection assumptions. First-appearance sensitivities refit the selected forest specification without retuning its parameters. Positive bars favour the learned model. Each bar compares scores within its own restricted cohort or outcome definition, not across different base rates. The strictest coverage subset contains 66 final videos and reverses the score advantage. No confidence interval is implied by these descriptive bars.

Changing the primary threshold to at least two or at least four recorded dates changes the task prevalence substantially. The stricter four-date target can have higher AUC while still predicting a much rarer event, so that AUC increase is not evidence of a better practical forecasting system. Removing the five largest

development channels, extending the opening washout to fourteen days and resolving duplicates by the first source row assess different concerns. They are reported without selecting whichever restriction happens to favour the model. Counter-enabled and no-removal subsets also reduce concern about unusual metadata, but the main predictive specifications already exclude removal status.

For daily data, 895 video pairs have consecutive recorded dates. Views decrease in six pairs, likes in eight and comments in 33. These recorded corrections can arise from measurement or moderation processes and are not interpreted as negative consumption. Excluding negative counter changes leaves 485 final landmarks from the same 228 videos and retains a modest forest gain over prevalence. Alternative duplicate resolution changes the daily score only minimally at the reported precision. The remaining uncertainty concerns unobserved collection and platform processes, rather than an exact duplicate-resolution choice.

4 Discussion

4 1 What the evidence supports

The findings support a limited but useful prediction claim. Public counters and matured channel history contain some information about later representation among videos already observed in a historical regional archive. The information is stronger than a single global probability for several comparisons, with positive paired intervals for the selected forest against prevalence and the simple views-and-age model. The first-appearance logistic comparator is less certain, and the daily forest also provides a modest gain. Strict coverage selection can reverse the first-forest advantage, preventing a universal forecasting claim. A monitor can therefore use these scores as structured evidence when organising a review queue, while retaining alternative simple rankings and uncertainty rather than treating a predicted probability as an instruction.

The analysis also shows why an interaction total can be misleading. Views, likes and comments do not share a denominator or interpretation, and a conditional model assigns like and comment density different directions. This result concerns a selected cohort after visibility has occurred. It does not say that likes reduce popularity, that comments cause trending, or that a platform algorithm rewards a particular engagement ratio. The practical lesson is to retain separate measures and inspect their context. Aggregating them into a single impressive engagement number can discard useful information and obscure the limits of inference.

A second substantive lesson is that prediction time defines what can be learned. At first appearance, there is no exact preceding-day archive observation for the new entry. At later landmarks, lifecycle and recorded growth become available but repeated observations alter the evaluation unit. The study handles these as different tasks rather than filling the missing first-day trajectory with invented values. It further shows that an apparently sensible momentum feature set has only a small incremental contribution in fixed-parameter removal checks. A model with more inputs is not automatically a better explanation.

4 2 A practical monitoring workflow

A feasible use is a music publicity analyst preparing a daily note. The analyst records the new regional snapshot, distinguishes unavailable counters from zero counters, and computes the initial probability using only available first-snapshot information and completed earlier channel outcomes. The probability is shown

alongside views, interaction rates and archive coverage. It is then used to order a limited review queue. A human review can examine whether a high-ranked clip has a meaningful announcement, a growing discussion, or an unusual measurement pattern. Those interpretations are external follow-up work, not conclusions supplied by the numerical classifier.

On the following observed day, the analyst can refresh the estimate with the daily landmark model. A missing regional snapshot should trigger an unavailable status, not an assumed exit. Counter decreases should remain visible as recorded changes, and rates should not be assigned a behavioural meaning when feedback is disabled. Each prediction can retain its timestamp, information fields, version and eventual archival outcome. This makes the workflow auditable and supports monitoring whether its calibration degrades. It does not require downloading individual comments or examining private viewer information.

The pooled capacity exercise should be understood as a demonstration of prioritisation, while the daily-batch result exposes workload constraints. Even a nominal twenty-percent policy may select more than twenty percent overall because a day with one new video still produces one selection when rounded upward. An actual monitoring operation could set a fixed daily number, accumulate a weekly queue, or reserve capacity for unusual clips independent of their score. Those are different policies with different objectives. The evidence here does not choose an optimal staffing level or show how an intervention changes a video's outcome.

4 3 Channel transfer and unequal exposure

The channel-history design avoids an identity lookup table, but it still uses prior observed performance associated with a name. That can favour frequently represented channels with more informative history and can give new channels only a global prior. The later test therefore includes a channel-overlap diagnostic rather than claiming that disjoint video identifiers guarantee disjoint production environments. There are 98 unseen-channel primary videos and 99 unseen-channel daily videos. The selected first forest has somewhat weaker point Brier on unseen channels; the daily forest has lower point Brier on the unseen-channel subset. Differences in prevalence and cohort composition prevent attributing these contrasts solely to channel familiarity.

The subgroup patterns are descriptive. Channel names are not stable account identifiers, the subsets have limited counts and their outcome distributions may differ. Consequently, the analysis does not establish a fairness result or a universal benefit for major producers. Removing the largest development channels checks whether those records alone explain the overall result, but it is not a substitute for a prospectively separated channel evaluation. A promotion analyst should avoid equating a cold-start channel probability with artistic quality, commercial potential or an assessment of the artist.

4 4 Limits of the historical archive

The most consequential limitation is outcome measurement. Missing collection dates are excluded explicitly, but partial recorded dates remain. An absent record can reflect collection coverage or ranking processes that this file does not distinguish. First appearance can also be left-censored before the archive start, and the opening washout only reduces that concern. The thresholded seven-day outcome counts reappearances and cannot support a statement that a video stayed continuously on a list for a specified duration. A rank trajectory, watch-time outcome or complete survival endpoint would require different observations.

Predictor availability is a related constraint. Counters are snapshot measurements with limited timing resolution; publication age is day-level; channel names can change; and the removal flag is excluded because its timing was not verified. Ratios depend on global cumulative views, not regional independent viewers. The analysis does not use comment language or sentiment, music genre, release schedules, advertising exposure, recommender impressions or the identity of individual users. This keeps the investigation compact and reproducible, but it also limits an explanation of why a particular clip appears again.

The later cohorts cover only a small portion of 2018, and this retrospective study does not validate a current production product. YouTube announced a transition away from its broad Trending page toward other discovery surfaces and category charts in 2025 [13]. The current system and counter definitions cannot simply be assumed identical to the 2017 to 2018 archive. Historical prediction skill is therefore a research and portfolio result, not evidence that these coefficients or probabilities can be deployed unchanged against present YouTube Charts.

Statistical uncertainty also remains broader than a single interval. Whole-channel resampling preserves within-name dependence, while date-block resampling addresses calendar dependence separately. Neither scheme includes model-search uncertainty or establishes independence among channels on the same day. Multiple specifications, capacity fractions and sensitivities create opportunities for favourable chance findings. This manuscript retains the development selection, states which diagnostics follow result inspection, and reports failures such as the strict-coverage reversal. It does not describe every positive comparison as a separately confirmed discovery.

5 Conclusion

Continued archive visibility can be studied with modest public-data models when prediction times, missing dates and video identity are handled explicitly. In the later first-appearance cohort, a development-selected channel-history forest reaches AUC 0.687 and a pooled monitoring precision of 62.8% at 43 selections, with positive paired channel intervals against simple comparators within this cohort. These conditional results do not establish performance on a future platform or a causal promotion effect. For next-day landmarks, a daily forest gives a modest clearer probability-score improvement. Like and comment density supply different conditional information, and stricter coverage can reverse the first-model advantage. The resulting contribution is a reproducible framework for separating first observation, later updating and limited monitoring capacity, with conclusions confined to the historical archive that was actually analysed.

Declarations

Frankie Gao is the sole named author. The stated affiliation is perspectiveopen.com and the corresponding contact is zcfbfga@ucl.ac.uk. The study uses public aggregate video metadata and does not collect private viewer records, individual comment text, downloaded videos, audio or thumbnails. No participant recruitment or intervention was performed, and no ethics approval number is asserted.

OpenAI Codex assisted with study implementation, code development, statistical checking, figure production and manuscript preparation. All numerical analyses, fitted models, resampling calculations and reported experiments were executed locally on the preserved dataset. No model-generated experimental observations or invented measurements were used. The sole named-author designation does not imply that the computational and writing assistance was absent.

The locally executed research package contains preserved source data and provenance, an explicit protocol, source-identity revisions, the frozen model configuration, development and later predictions, sensitivity and uncertainty tables, figure source files, an editable analysis workbook and an executed reproduction notebook. These working materials are retained privately by the author and publisher and are not deposited as public supplementary files. The tabular source is publicly available from the Kaggle dataset cited in reference [1], whose metadata labels it CC0 Public Domain. The study does not redistribute the underlying copyrighted audiovisual media. Readers may direct methodological and reproducibility enquiries to the corresponding author. No public code-repository deposit or journal DOI is claimed.

References

- [1] DataSnaek. Trending YouTube Video Statistics. Kaggle dataset, version 115; metadata last updated 3 June 2019. Korean CSV retrieved 2 October 2026; metadata verified 3 October 2026.
<https://www.kaggle.com/datasets/datasnaek/youtube-new>
- [2] Mitchell J. Trending YouTube Scraper. Collector repository and scraper.py; consulted 3 October 2026.
<https://github.com/mitchelljy/Trending-YouTube-Scraper>
- [3] Google. YouTube Data API v3 videos list. Documentation for chart and regionCode; consulted 3 October 2026.
<https://developers.google.com/youtube/v3/docs/videos/list>
- [4] Google. YouTube Data API v3 videos resource. Video statistics documentation; consulted 3 October 2026.
<https://developers.google.com/youtube/v3/docs/videos#statistics>
- [5] Szabo G, Huberman BA. Predicting the popularity of online content. arXiv preprint 0811.0405. 2008.
<https://doi.org/10.48550/arXiv.0811.0405>
- [6] Borghol Y, Ardon S, Carlsson N, Eager D, Mahanti A. The untold story of the clones: Content-agnostic factors that impact YouTube video popularity. arXiv version 1311.6526. 2013. <https://arxiv.org/abs/1311.6526>
- [7] Cerqueira V, Torgo L, Mozetic I. Evaluating time series forecasting models: An empirical study on performance estimation methods. Machine Learning. 2020;109:1997-2028. <https://doi.org/10.1007/s10994-020-05910-7>
- [8] Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. Patterns. 2023;4(9):100804. <https://doi.org/10.1016/j.patter.2023.100804>
- [9] Breiman L. Random forests. Machine Learning. 2001;45:5-32. <https://doi.org/10.1023/A:1010933404324>
- [10] Gneiting T, Raftery AE. Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association. 2007;102(477):359-378. <https://doi.org/10.1198/016214506000001437>
- [11] Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE. 2015;10(3):e0118432. <https://doi.org/10.1371/journal.pone.0118432>
- [12] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research. 2011;12:2825-2830. <https://jmlr.org/papers/v12/pedregosa11a.html>
- [13] YouTube. Changes to discovering trending content on YouTube. Official announcement, July 2025; consulted 3 October 2026. <https://support.google.com/youtube/thread/356702168?hl=en&msgid=356716700>

Appendix A Feature removal and channel diagnostics

The feature-removal experiments hold the chosen hyperparameters fixed and refit each reduced model on the development cohort. This avoids replacing a fitted feature with an arbitrary perturbation while leaving the estimator unchanged. It also makes clear that removals compare prediction specifications, not

interventions on engagement. The removal flag is absent from all full and reduced predictors. Figure 11 displays these comparisons and Figure 12 displays previously observed versus unseen channel-name subsets.

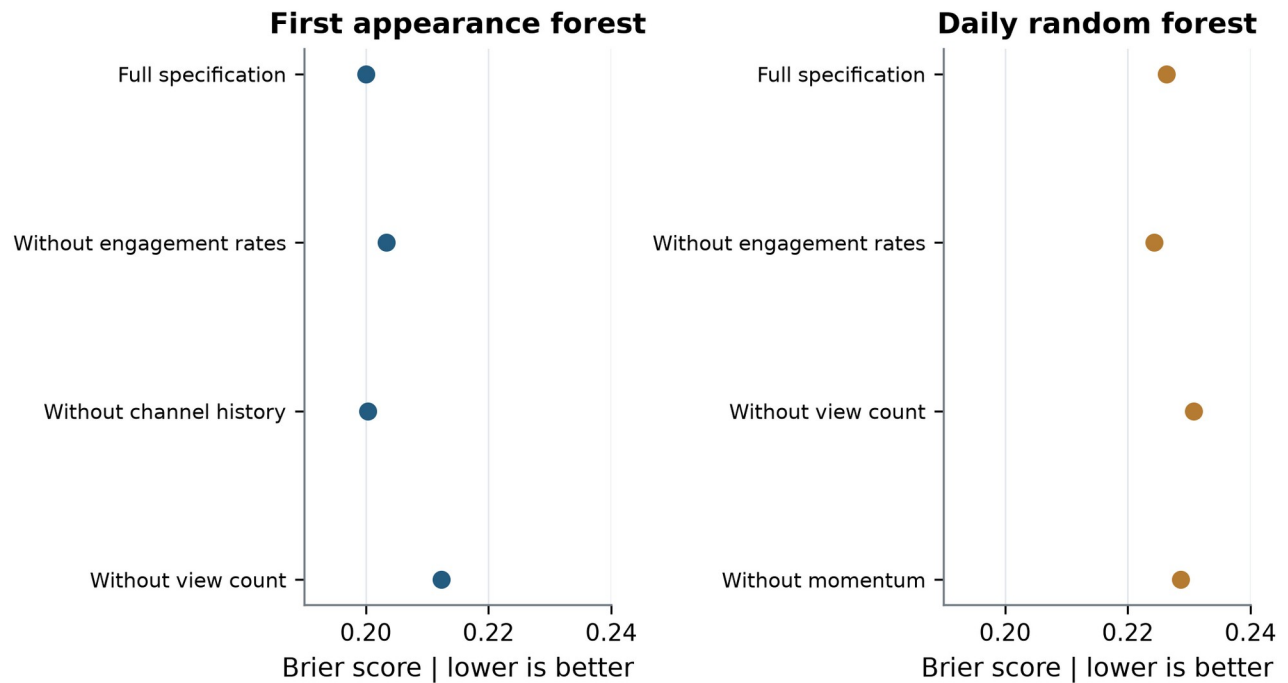


Figure 11. Fixed parameter refitted feature removals. Each reduced specification is refitted on development observations with the full model hyperparameters retained. The daily model improves slightly when static engagement rates are removed, and removing momentum changes its Brier score by only about 0.0022. The name RF momentum identifies its input set; it does not establish that momentum alone caused the observed gain. These are diagnostic comparisons without multiplicity-adjusted inference.

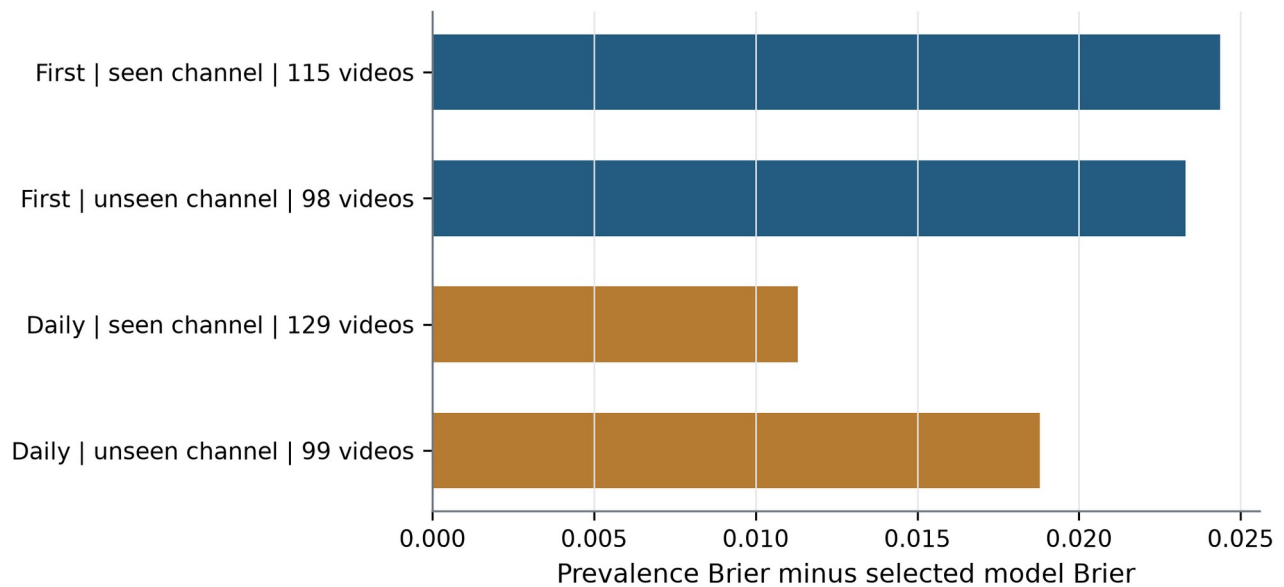


Figure 12. Channel overlap diagnostics. Channels are classified by whether their recorded name occurs in the respective fitting cohort. No video occurs in both fitting and final cohorts. First-appearance gains are larger on previously observed channels, whereas daily gains are larger on unseen channel names. Differences are descriptive subgroup comparisons, not demonstrated performance gaps; channel names are imperfect stable identifiers.

Table 7. Channel name overlap subsets in the later evaluation

Task and subset	Videos	Outcome rate	Selected Brier	Selected AUC
First Seen channel	115	33.0%	0.1969	0.704
First Unseen channel	98	34.7%	0.2037	0.677
Daily Seen channel	129	40.7%	0.2304	0.614
Daily Unseen channel	99	39.9%	0.2211	0.672

Subgroups recompute equal-video weights internally. Name overlap is assessed against the respective fitting cohort. These point comparisons do not prove a between-subgroup performance difference.

Appendix B Temporal blocks and forest repeat runs

Circular date-block results are presented alongside channel-cluster results, rather than substituted for them. The resampling unit changes which dependence structure is preserved. The 95 percent intervals in Table 8 remain conditional on the fitted models. Table 9 shows all five specified forest seeds, including seed 42 used in the manuscript; the most favourable later seed is not selected for the main result.

Table 8. Date block sensitivity against training prevalence

Task	Block days	Reduction	95 percent lower	95 percent upper
onset	7	0.0239	0.0034	0.0439
onset	14	0.0239	0.0031	0.0414
daily	7	0.0145	0.0071	0.0222

Task	Block days	Reduction	95 percent lower	95 percent upper
daily	14	0.0145	0.0075	0.0220

There are 2,000 circular calendar-block resamples per row. The original equal-video observation weights are retained within a resampled block.

Table 9. All random forest seed checks

Task and forest	Seed	Brier	ROC AUC
onset RF channel history	7	0.1993	0.695
onset RF channel history	11	0.2021	0.681
onset RF channel history	29	0.2003	0.694
onset RF channel history	42	0.2000	0.687
onset RF channel history	73	0.2009	0.682
daily RF momentum	7	0.2274	0.633
daily RF momentum	11	0.2268	0.637
daily RF momentum	29	0.2270	0.635
daily RF momentum	42	0.2264	0.640
daily RF momentum	73	0.2254	0.646

The first-appearance forest listed here is the best forest on development data, and is also the selected primary model after source-identity correction. Hyperparameters are held fixed across seeds.

Appendix C Reproducibility and data identity

The core sequence is data preparation, chronological model analysis, labelled additional diagnostics, independent numerical verification, figure generation and manuscript construction. The executed notebook refits the two selected specifications from the preserved processed data and compares every later predicted probability with the saved tables. It also recomputes the monitoring counts and produces visible analytical charts. The workbook provides a capacity input with formula-linked selections and explicitly separates source records, derived cohorts, predictions and validation results.

The raw Korean CSV SHA-256 digest is

4567450dce9c3b489c7a949d1e10eaf70c999d58355f537374f8b0189580ce77. The corrected configuration records this digest and the protocol digest. The local environment uses Python 3.12.14, NumPy 2.3.5, pandas 2.2.3 and scikit-learn 1.9.1 [12]. Independent checks reconstruct labels from source video-date sets, inspect every lag boundary and matured channel history, verify the absence of fitting/test video overlap, and recompute all later Brier scores, log losses and ROC AUCs using separate arithmetic. These checks validate the executed package; they do not establish prospective performance beyond the documented archive.