Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Predicting Public Engagement with Tourism Videos Across Unseen YouTube Channels Using Text-Informed Ensembles

Hongzhou Yuan[1]
Zhongkai University of Agriculture and Engineering
Jianye Liang[2]
valencia88222@163.com

Zhongkai University of Agriculture and Engineering

Article

Keywords: tourism video, engagement prediction, cultural context, channel generalization, short video, model audit

Posted Date: November 2nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11154644/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Predicting Public Engagement with Tourism Videos Across Unseen YouTube Channels Using Text-Informed Ensembles

Hongzhou Yuan[1], Jianye Liang[2]*

[1] School of Management, Guangzhou College of Commerce, Guangzhou 511363, China. ORCID: 0009-0009-2378-776X.
[2] Zhongkai University of Agriculture and Engineering, Guangzhou 510225, China. ORCID: 0009-0005-0572-1362.
Corresponding author: Jianye Liang; Email address: valencia88222@163.com

Keywords: tourism video; engagement prediction; cultural context; channel generalization; short video; model audit

Abstract

Public engagement counters for tourism videos reflect both content and exposure-related conditions. We analysed 4,231 public YouTube tourism videos, including 3,271 no longer than 60 s, and modelled contemporaneous views, likes and comments. A leakage-controlled neural branch combined train-fitted text, thumbnail and structured inputs; its predictions were blended with a validation-selected XGBoost metadata expert using validation-only convex weights. Across five randomized channel-grouped splits, the hybrid achieved mean R^2 values of 0.538, 0.516 and 0.495 for views, likes and comments, improving on XGBoost by 0.078, 0.063 and 0.064. Corresponding strict-short-video R^2 values were 0.529, 0.481 and 0.504. Removing text from training and evaluation reduced rebuilt-hybrid R^2 by 0.057–0.072 across outcomes. Cultural supervision and diffusion did not improve prediction, while removing the sparsely available visual branch improved the neural branch. Chronological transport was substantially weaker, particularly for comments. The results support a text-informed ensemble for channel-held-out contemporaneous engagement prediction and delimit the conditions under which its performance transfers.

Introduction

Culture and tourism are closely linked through tangible heritage, intangible practices, creativity and everyday ways of life [1]. Social media has become a major setting in which destination

information is searched, produced and circulated [2,3], while travellers' decisions to share tourism experiences depend on both individual motivations and platform-mediated social conditions [4]. Destination-marketing organizations use these platforms unevenly [5], and post-level evidence shows that sender and visual-content characteristics are associated with observable interaction [6]. Recent destination-video studies have extended this line of work by extracting visual, textual and acoustic characteristics at scale and relating their configurations to engagement [7–9]. These studies establish the value of automated content analysis, but they do not remove the influence of channel scale, publication timing or platform exposure.

Engagement is also multidimensional: users may consume, contribute to or create platform content [10], and brand-post responses vary with content strategy [11]. Predictive work has consequently moved beyond single-modality models. AMPS uses intra- and inter-modal attention for YouTube Shorts popularity prediction [12], whereas EnTube integrates titles, audio, thumbnails, video content and tags for three-class engagement prediction [13]. A 90,000-video benchmark instead targets early watch-based engagement [14]; the VQualA challenge formalized related short-video evaluation [15], and recent large-multimodal-model ensembles combine frames, text and audio [16]. These task definitions are not interchangeable with public counter snapshots, but they demonstrate the need to state the outcome and observation time precisely.

The closest studies therefore answer complementary rather than identical questions. Tourism-focused analyses relate multimodal feature configurations to engagement in destination-specific Douyin, US-national-park or projected-destination-image samples [7–9]. AMPS and EnTube formulate popularity or engagement as classification and emphasize full video content or pre-publication features [12,13], whereas SnapUGC and VQualA use watch-based outcomes for newly published short videos [14–16]. Those outcomes, sampling frames and evaluation targets

are not directly comparable with contemporaneous YouTube counters. The present study is distinguished by continuous prediction of three public counters, channel-level exclusion from training, validation-only ensemble weighting, a strict-duration rerun, and explicit reporting of geographic, chronological, missing-modality and component failures; it does not claim to supersede content-only early-engagement forecasting.

More generally, multimodal machine learning distinguishes representation, alignment, fusion and co-learning problems [17]. Cross-modal transformers model unaligned sequences [18], building on scaled dot-product attention [19]. YouTube systems research further shows that recommendation exposure [20], early view trajectories [21], time-varying popularity growth [22] and cross-platform early signals [23] all influence observed popularity. Public counters are therefore contemporaneous outcomes of content, channel history, exposure and elapsed time rather than prospective responses to content alone [24].

Cultural context remains substantively important because heritage, local identity, ritual, craft, food and language cues can shape destination meaning. However, detecting a cultural label is not equivalent to learning a distinct cultural representation, and neither establishes a causal effect on engagement. Diffusion cross-attention has been proposed as an inductive bias for disentanglement in image reconstruction [25], while denoising diffusion models supply the underlying noise-prediction framework [26]. Causal-representation research defines substantially stronger counterfactual and intervention-based requirements [27,28]. The current observational YouTube design does not meet those requirements.

Evaluation must also respect the deployment unit. Deep models can exploit shortcuts [29], and random cross-validation is unreliable when observations are hierarchically or temporally dependent [30,31]. Distribution-shift benchmarks [32] and uncertainty studies [33] show that in-

distribution accuracy does not guarantee transportability or calibrated confidence. Split-conformal methods offer finite-sample marginal coverage under exchangeability [34], but their guarantees require care when exchangeability is violated [35]. We therefore use channel-grouped splits, describe variability across randomized splits rather than treating them as independent population replications, and audit country and temporal boundaries separately.

We therefore ask four evidence-bounded questions. First, how accurately can contemporaneous public engagement counters be predicted for channels absent from training? Second, does the text-informed neural branch complement a validation-selected structured expert, and is that complementarity reduced when text is removed throughout training and evaluation? Third, do reviewed cultural-context labels contain separable information, and does cultural supervision improve prediction? Fourth, do the visual and diffusion-inspired components provide reproducible gains? We contribute a documented public-platform data contract, a provenance-aware cultural annotation workflow, leakage-controlled channel-grouped evaluation, strict short-video and geographic audits, and full reporting of positive, null and contradictory component evidence.

Results

Dataset and evaluation partitions

The leakage-controlled evidence sequence is summarized in Fig. 1. The processed table contained 4,231 unique videos from 3,141 channels. Of these, 3,271 met the strict duration threshold of 60 s, 3,215 were published from 2023 through 2025, and 2,677 met both conditions. The outcomes were contemporaneous API snapshots of log(1 + views), log(1 + likes) and log(1 + comments), not future engagement measured after a fixed post-publication horizon. Public

share counts were unavailable from the Data API, while favoriteCount is deprecated and fixed at zero [36]; neither was used as an outcome.

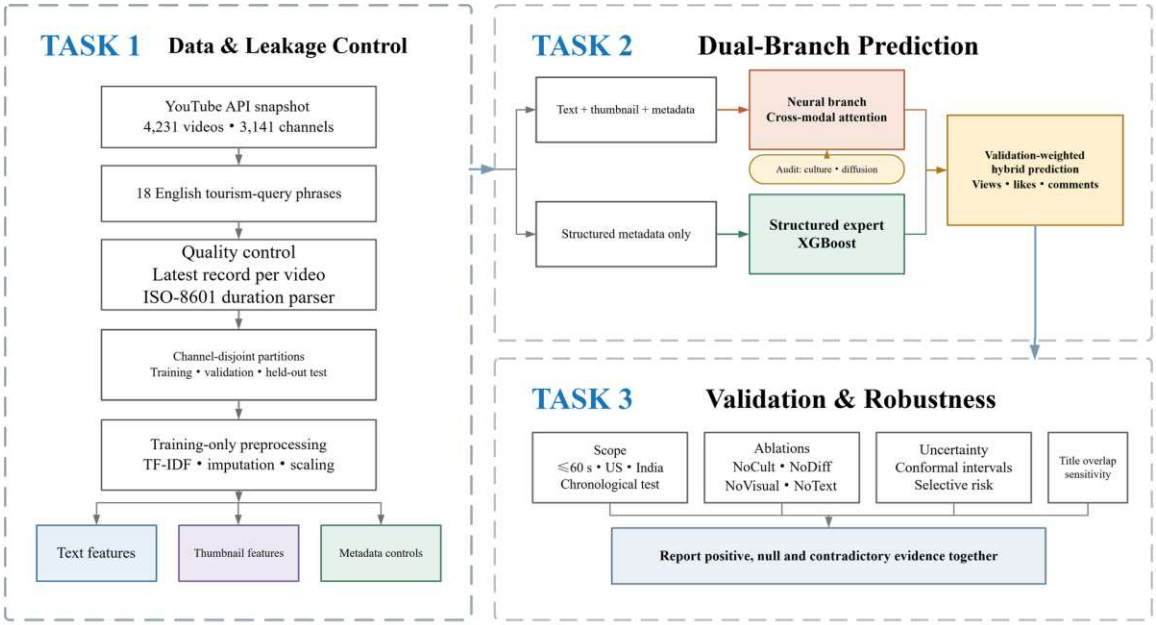


Figure 1. Task-structured research framework and leakage-controlled evidence sequence. Task 1 covers the public API snapshot, quality control, channel-disjoint partitioning and training-only preprocessing. Task 2 combines a multimodal neural branch with an independently selected structured expert through validation-only weighting. Task 3 groups scope, component, uncertainty and title-overlap audits. Test outcomes are reserved for final evaluation, and outputs are contemporaneous public-counter predictions rather than future forecasts or causal effects.

Video duration had a median of 29 s (interquartile range 16–60 s). Most channels contributed only one record (87.1%), which makes channel-disjoint evaluation important even though within-channel replication was limited. Description, tag and channel-country fields were missing for 35.3%, 60.8% and 30.4% of records, respectively; usable frozen thumbnail embeddings were available for 19.9% (Table 1; Fig. 2). The outcome distributions remained strongly right-skewed after log transformation, especially because 37.4% of videos had zero public comments at collection.

Table 1. Dataset characteristics.

Characteristic	Value

Characteristic	Value
Videos	4,231
Unique channels	3,141
Duration, median (IQR)	29 s (16–60)
Videos no longer than 60 s	3,271 (77.3%)
Single-video channels	2,737 (87.1%)
Missing description	1,493 (35.3%)
Missing tags	2,573 (60.8%)
Missing channel country	1,285 (30.4%)
Usable thumbnail embedding	841 (19.9%)

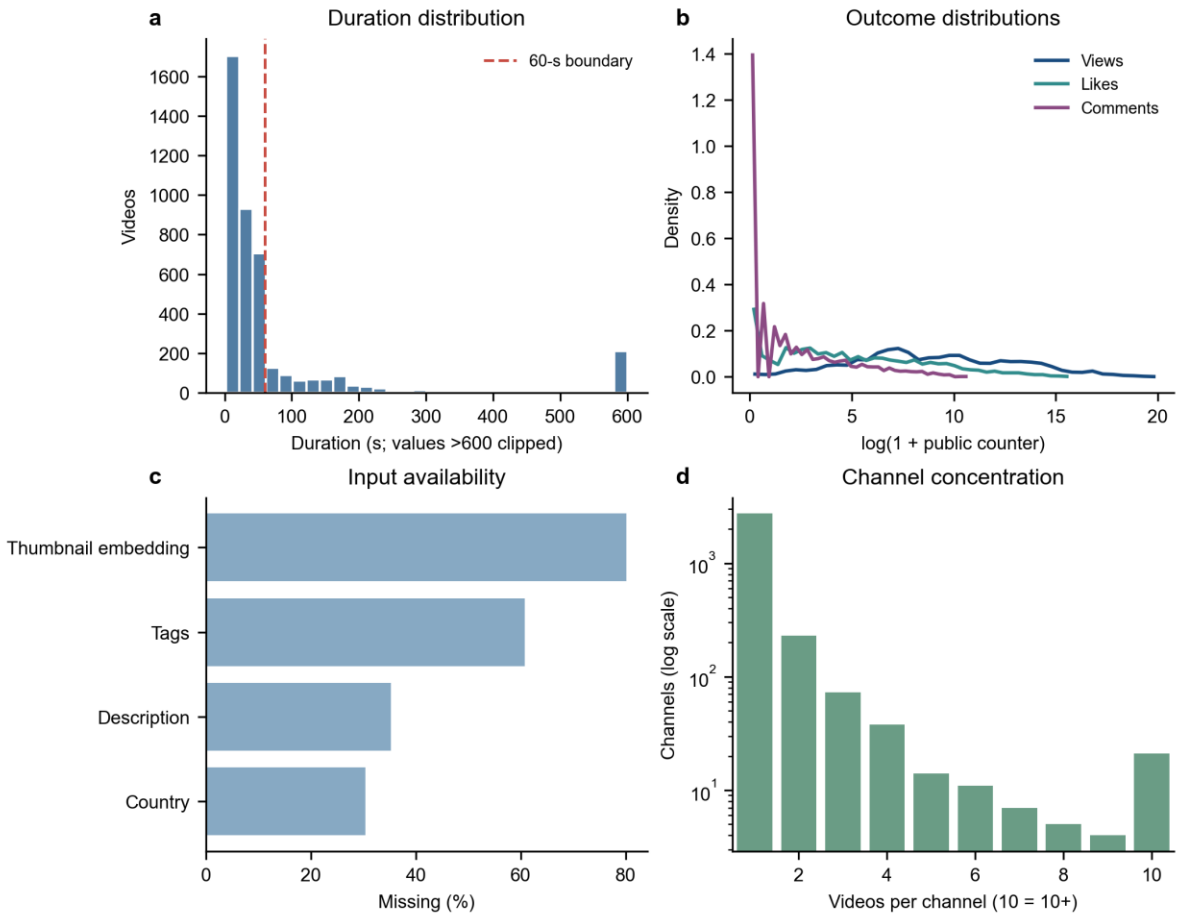


Figure 2. Dataset profile. a, Video-duration distribution, with values above 600 s clipped for display and the 60-s short-video boundary marked. **b**, Distributions of the three log-transformed public counters. **c**, Missingness of major input fields and thumbnail embeddings. **d**, Videos per channel; channels contributing ten or more videos are grouped at ten. This profile motivates channel-grouped evaluation and explicit missingness audits.

Five randomized channel-grouped splits were generated with seeds 20260807–20260811. Within every split, channels were mutually exclusive across training, validation and test partitions. Each test split contained 472 channels and 599–711 videos. Across the five splits, 2,386 unique videos appeared in a test partition; mean pairwise Jaccard overlap was 0.083 for both test videos and test channels. These splits are repeated partitions of one query-based dataset, not independent population samples. All vocabulary fitting, imputation, feature scaling and outcome standardization used training rows only.

Model fitting was confined to the training channels, early stopping and ensemble selection were confined to validation channels, and all reported test metrics were computed only after those choices had been fixed.

Channel-held-out engagement prediction

The validation-weighted hybrid obtained mean R^2 values of 0.538 for views, 0.516 for likes and 0.495 for comments (Table 2; Fig. 3a). Corresponding mean Spearman correlations were 0.740, 0.701 and 0.652. The standalone neural branch obtained mean R^2 values of 0.368, 0.350 and 0.335, whereas the validation-selected XGBoost metadata expert obtained 0.460, 0.453 and 0.431. Table 2 reports descriptive 95% t intervals across the five randomized splits; these intervals summarize split sensitivity and are not population-sampling confidence intervals.

Table 2. Hybrid performance across five randomized channel-grouped splits.

Outcome	MAE, mean (95% split interval)	RMSE, mean (95% split interval)	R^2, mean (95% split interval)	Spearman, mean (95% split interval)
Log views	2.076 (2.021–2.132)	2.596 (2.522–2.670)	0.538 (0.485–0.592)	0.740 (0.708–0.773)
Log likes	1.982 (1.894–2.070)	2.503 (2.365–2.642)	0.516 (0.500–0.532)	0.701 (0.693–0.708)
Log comments	1.289 (1.218–1.360)	1.687 (1.601–1.772)	0.495 (0.444–0.546)	0.652 (0.621–0.684)

As an additional low-capacity benchmark, validation-selected Ridge models achieved mean R^2 values of 0.333, 0.311 and 0.274 using text alone and 0.341, 0.320 and 0.292 using text plus metadata for views, likes and comments. Metadata-only Ridge values were 0.092, 0.094 and 0.080. These results show that text was predictive without the proposed architecture, while the stronger nonlinear expert and validation-weighted ensemble provided the highest overall accuracy.

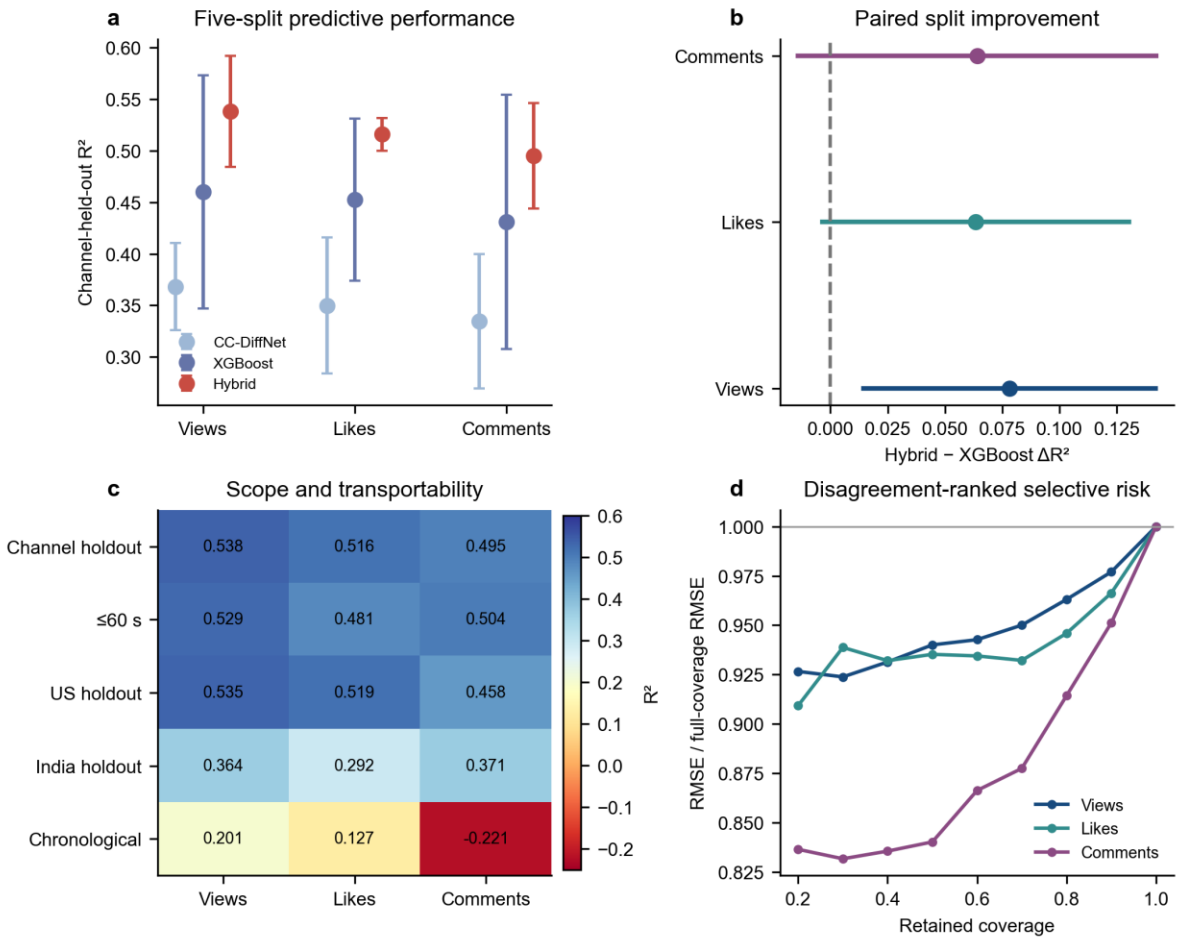


Figure 3. Predictive performance, scope and selective risk. a, Mean channel-held-out R^2 across five randomized channel-grouped splits; error bars denote descriptive 95% t intervals. **b**, Split-paired hybrid-minus-XGBoost R^2 differences and descriptive intervals. **c**, Hybrid R^2 in the complete sample, strict ≤60-s subset, country-held-out audits and chronological test. **d**, Mean disagreement-ranked risk–coverage curves across five randomized splits, normalized to full-coverage RMSE.

Record-level predictive fit and structured-expert importance

The prespecified seed-20260807 test partition contained 665 videos from channels absent from training and validation. In this split, hybrid R^2 was 0.556 for views, 0.529 for likes and 0.494 for comments, with RMSE values of 2.563, 2.479 and 1.750, respectively (Fig. 4a–c). The binned observed-versus-predicted densities show broad alignment with the identity line, together with compression at the upper tails; the model therefore captured substantial rank and level information but did not reproduce the most extreme engagement values.

Grouped permutation analysis of the independently selected XGBoost expert provided a descriptive check on the structured controls (Fig. 4d). Jointly permuting channel-subscriber features in the held-out test set increased RMSE by 55.5%, 57.3% and 54.5% for views, likes and comments. Duration increased view and like RMSE by 10.3% and 5.6% but comment RMSE by only 0.8%; video age produced increases of 5.2%, 5.2% and 6.7%. These outcome-specific values are means over 20 deterministic permutations. They quantify predictive dependence within the fitted expert and should not be interpreted as causal effects or as test-set model selection.

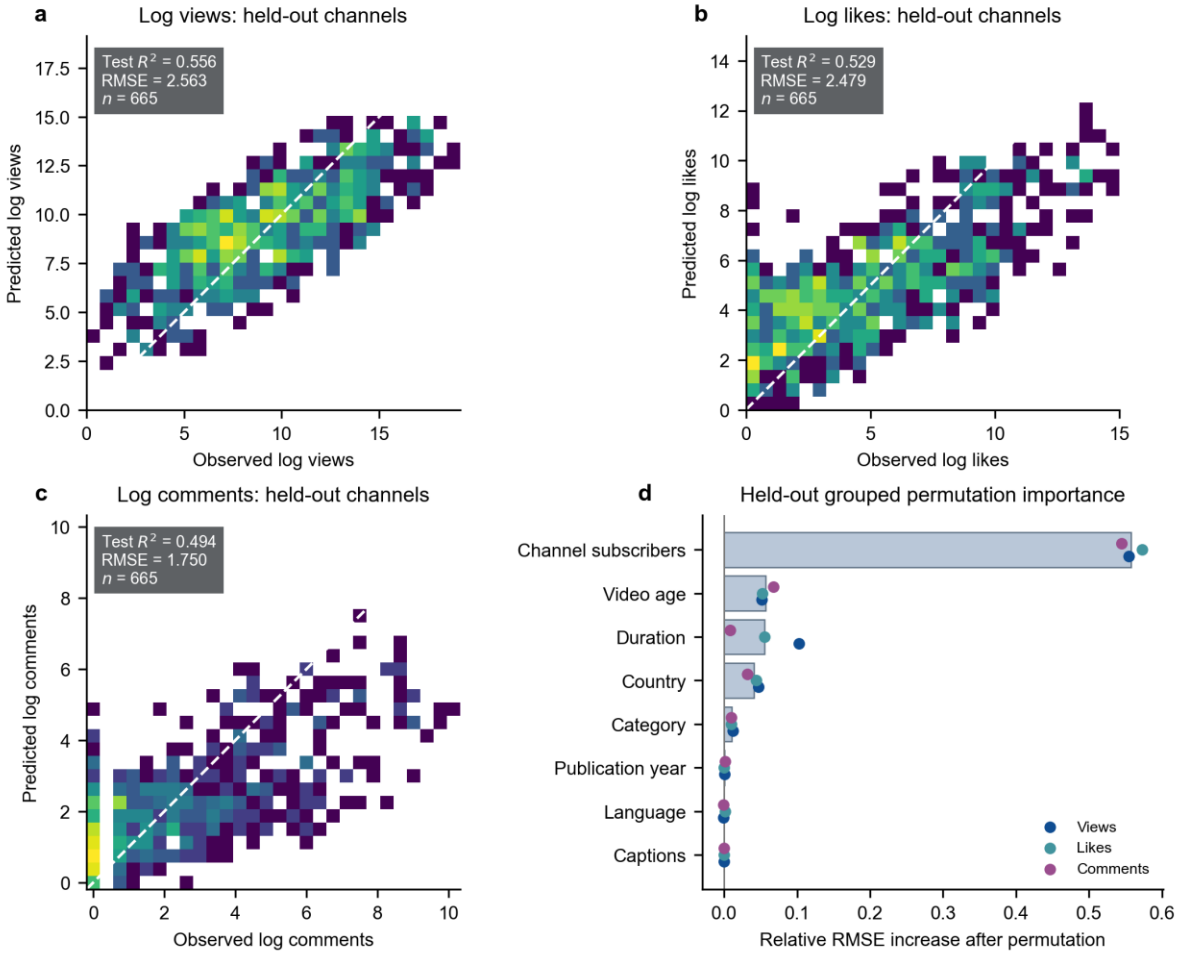


Figure 4. Held-out predictive fit and structured-expert importance. a–c, Aggregate two-dimensional densities of observed and hybrid-predicted log engagement in the prespecified seed-20260807 channel-held-out test split; dashed lines denote identity, and annotations report test R^2, RMSE and sample size. **d**, Grouped permutation importance for the selected XGBoost expert.

Dots show the mean relative RMSE increase over 20 deterministic joint permutations for each outcome; bars show the arithmetic mean across outcomes. These post hoc test-set diagnostics did not alter model or ensemble selection.

Paired comparison with the structured expert

Relative to XGBoost, the mean hybrid R^2 difference was +0.078 for views, +0.063 for likes and +0.064 for comments (Fig. 3b). Descriptive 95% t intervals across randomized splits were +0.014 to +0.142, −0.003 to +0.130 and −0.014 to +0.142, respectively. The hybrid exceeded XGBoost in five of five splits for views and four of five for both likes and comments. Because the split seeds are dependent re-partitions of one dataset, we treat these comparisons as stability summaries rather than independent-replication hypothesis tests.

Validation-selected neural weights averaged 0.410 for views, 0.384 for likes and 0.440 for comments. Thus, the structured expert retained the larger mean weight, while the neural branch contributed complementary residual information. The gain supports the validation-weighted ensemble but does not show that every neural component was beneficial.

Strict short-video and country-held-out validation

The frozen pipeline was rerun on all 3,271 videos with correctly parsed duration ≤60 s, representing 2,424 channels; no hyperparameter was retuned on short-video test outcomes. Across five randomized channel-grouped splits, the hybrid achieved mean R^2 values of 0.529 (95% split interval 0.483–0.575) for views, 0.481 (0.390–0.572) for likes and 0.504 (0.417–0.590) for comments. These results reproduce the ensemble pattern in a duration-defined short-video subset.

Country-held-out analyses were restricted to records with known channel country and used one prespecified seed. When all US channels (n=761) were held out, hybrid R^2 was 0.535, 0.519 and 0.458 for views, likes and comments. Holding out all Indian channels (n=956) reduced R^2 to 0.364, 0.292 and 0.371. The lower Indian performance indicates heterogeneous geographic

transport and argues against treating channel-held-out accuracy as universal cross-country validity (Fig. 3c). These are two targeted domain audits, not estimates of country effects.

Cultural-label separability

The stratified annotation pool contained 1,000 videos. GLM-5.2 generated 998 usable machine-assisted labels that the authors report were manually reviewed. Fifty-five were retained as uncertain (U) and two remained unlabeled. On 943 definite labels, a channel-held-out TF–IDF logistic classifier achieved accuracy 0.832, F1 0.892, ROC-AUC 0.864 and PR-AUC 0.955. Cultural context was therefore detectable from public text, although this result does not establish a distinct latent mechanism.

Component and representation audit

Five-split component audits bounded the stronger architectural claim (Fig. 5a). Removing cultural supervision (NoCult) changed mean R^2 by only +0.0010, +0.0007 and +0.0007 for views, likes and comments relative to the complete neural branch. Removing the diffusion objective (NoDiff) changed R^2 by +0.0022, +0.0044 and +0.0081. Thus, neither component produced a measurable predictive gain. Removing the visual branch (NoVisual) increased mean R^2 from 0.368 to 0.390 for views, from 0.350 to 0.372 for likes and from 0.335 to 0.360 for comments. Only 841 of 4,231 videos (19.9%) had usable frozen visual embeddings, so this result applies to the present sparse-thumbnail design rather than to visual information in general. After validation-only recalibration with the unchanged XGBoost expert, the rebuilt no-text hybrid achieved mean R^2 values of 0.467, 0.457 and 0.438 for views, likes and comments. Its paired changes relative to the full hybrid were −0.072, −0.059 and −0.057, and it lost in all five paired splits and outcomes (Fig. 5b). Text therefore contributed reproducible predictive

information within this pipeline, although the observational design does not identify a causal effect of text.

The corrected-code latent probe likewise failed to validate the proposed partition. On 147 labeled test videos, Zc ROC-AUC was 0.582, compared with 0.626 for Zn; the bootstrap Zc−Zn difference was −0.044 (95% interval −0.153 to 0.067), and linear CKA between the representations was 0.900. The cultural information was therefore not preferentially isolated in Zc, and the two latent spaces remained strongly similar. These results do not establish causal or statistical disentanglement.

Controlled missingness and temporal boundary

Diffusion training did not improve missing-text robustness in the prespecified seed-20260807 stress test. At 30%, 50% and 70% deterministic test-text masking, NoDiff produced higher macro-average R^2 than the complete model (Fig. 5c). At 70% masking, both models had negative R^2 for every outcome. A corrected-code chronological hybrid test yielded R^2 values of 0.201, 0.127 and −0.221 for views, likes and comments. These single-split results show that channel-grouped performance does not imply robustness to severe modality loss or future distribution shift.

Prediction intervals and selective risk

Validation-residual split-conformal intervals with 90% nominal coverage achieved mean empirical coverage of 0.876 for views, 0.883 for likes and 0.901 for comments (Fig. 5d). Descriptive 95% split intervals were 0.853–0.899, 0.847–0.919 and 0.874–0.928, respectively. Coverage was therefore close to nominal for comments but below nominal for views and likes under channel shift. Model disagreement provided a modest selective ranking signal: retaining the 20% of test videos with the lowest absolute neural/XGBoost disagreement reduced mean

RMSE from 2.596 to 2.405 for views, from 2.503 to 2.276 for likes and from 1.687 to 1.411 for comments (Fig. 3d).

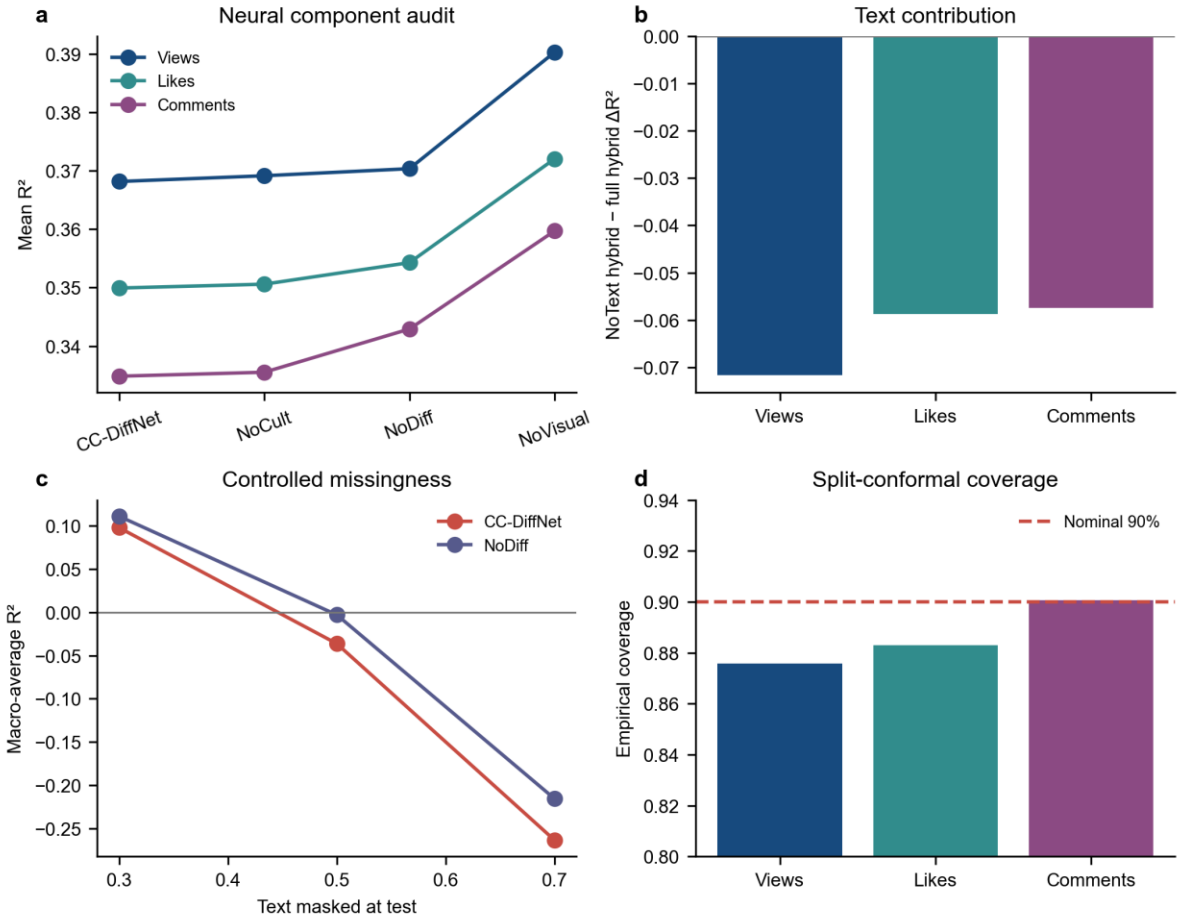


Figure 5. Component and robustness audits. a, Five-split mean R^2 for the complete neural branch and the NoCult, NoDiff and NoVisual ablations. **b**, paired loss in hybrid R^2 after removing text throughout training, validation and test. **c**, Macro-average R^2 after deterministic masking of 30%, 50% and 70% of test text, comparing the complete model with NoDiff. **d**, Mean empirical coverage of validation-residual split-conformal intervals at 90% nominal coverage.

Normalised-title duplicates between development and test partitions affected 0–20 test videos per split. Removing these records changed mean hybrid R^2 only from 0.538 to 0.538 for views, from 0.516 to 0.517 for likes and from 0.495 to 0.495 for comments. The main findings were therefore not explained by exact title reuse across partitions.

Discussion

The principal finding is residual complementarity between the validation-selected metadata expert and the neural branch. Structured controls, particularly video age and channel scale, formed the strongest individual predictor, but the neural predictions improved the convex hybrid in most randomized channel splits. The five-split NoVisual result indicates that this complementarity did not depend on the sparse thumbnail branch. By contrast, NoText substantially reduced both neural and rebuilt-hybrid R^2 in every split and outcome, supporting a reproducible text contribution within the frozen pipeline. This does not isolate individual words or establish a causal text effect; it may reflect text–metadata interactions, neural nonlinearities or model diversity. The pattern is consistent with tourism-video work showing that content configurations matter [7–9] and with engagement models that combine content and metadata [12–16]. It also cautions against evaluating a proposed neural model only against weak linear or single-modality baselines.

The uncertainty analysis adds an operational boundary. Split-conformal intervals were slightly under-covered after channel shift, consistent with the need to distinguish exchangeable calibration theory from shifted deployment [33–35]. Disagreement-based selection reduced average error most clearly for comments and less for views. These outputs may support cautious ranking or human review, but they are not deployment guarantees.

The strict duration analysis reproduced the ensemble advantage, strengthening the relevance to short-video communication. However, country-held-out performance was asymmetric: results were close to the complete-sample estimates for US channels but materially lower for Indian channels. This gap may reflect language, channel scale, destination mix, collection-query coverage or other unobserved platform processes. It is a transportability warning, not evidence of a country effect.

The cultural result is narrower. Cultural presence was separable from text, but removing cultural supervision did not reduce engagement prediction. The current latent probe also found nearly equal cultural ROC-AUC in Zc and Zn and high similarity between the spaces. Detectability should therefore not be equated with a task-useful or disentangled representation. The NoDiff audit likewise shows that an inductive bias transferred from image reconstruction [25] cannot be assumed to improve engagement prediction.

These findings alter the defensible contribution. The study supports a validation-weighted text-informed neural–metadata ensemble under channel shift, together with a reusable cultural-context and component-audit workflow. It does not support the original claim that cultural supervision, causal disentanglement, diffusion attention or the visual branch drove the improvement. Causal-representation claims require observable consequences of counterfactual assertions or explicit interventions [27,28], which are absent here.

Several limitations remain. The cultural labels were machine-assisted and single-adjudicated rather than independently double-coded, so Cohen's κ is not estimable. Two annotation rows remain missing. Only 841 videos had usable frozen thumbnail embeddings, and raw audiovisual streams were not analysed. Public counters mix content response with unobserved recommendation exposure and observation time, so the task is contemporaneous counter modelling rather than upload-time forecasting. The query-based sample is not representative of all tourism videos, and chronological performance was negative. Finally, the five randomized splits overlap and come from one dataset; their intervals quantify split sensitivity, not population-sampling uncertainty.

Methods

Data collection and outcome construction

Public video records were collected with the YouTube Data API v3 on 7 August 2026 between 07:25 and 11:28 UTC in three documented rounds comprising 18 unique English query phrases. The complete phrases and per-round retrieval counts are reported in Supplementary Methods. The API returned 4,694 metadata rows; retaining the most recent collection record for each video ID yielded 4,231 unique videos. No relevance filter based on outcomes or model performance was applied after retrieval. The retained fields included titles, descriptions, tags, publication timestamps, channel identifiers, public subscriber counts when available, duration, category, language, country, caption status and public interaction counters. The Analytics API exposes owner-authorized metrics such as shares, but these are not public Data API counters [36,37]. Each counter was observed at collection time rather than after a standardized post-publication horizon. All three outcomes were log-transformed to reduce heavy-tailed scale.

Annotation protocol

The codebook distinguished heritage, folk custom, performing arts, crafts and food, place identity, religious culture, language signals, mixed cases and uncertainty. Automated output included presence, category, cultural role, an evidence span and uncertainty. GLM-5.2, accessed through the SiliconFlow API, was used to generate the machine-assisted first-pass annotations from the documented codebook; no model output was treated as an author or as independent human coding. The authors report reviewing the generated labels. Two positive records lacking evidence spans were repaired from their supplied titles or descriptions, with changes recorded in a repair log and no category changes. The completed files do not contain two independent annotator label sets; the workflow is therefore described as machine-assisted annotation followed by single human adjudication, not as double-coded gold-standard annotation.

Generative-AI assistance disclosure

GLM-5.2 was used only for the machine-assisted annotation stage described above. OpenAI Codex was used to assist with analysis-code drafting, reproducibility checks and manuscript drafting and editing. The authors reviewed the generated labels, code, numerical outputs, references and manuscript text, made the scientific decisions and retain full responsibility for the integrity and accuracy of the work. No generative-AI system is listed as an author.

Features and leakage prevention

Titles, descriptions and tags were represented using TF–IDF [38]. Thumbnails were encoded with a frozen ImageNet ResNet-18 [39]. Usable embeddings were available for 841 videos (19.9%); all other records used an explicit availability mask and learned missing-image token. Structured variables comprised video age, channel subscriber count, duration, publication year/month/day-of-week, category, language, country and caption availability. ISO-8601 durations were parsed by full-string matching of day, hour, minute and second components; unit tests covered PT45S, PT1M30S, PT1H2M3S and P1DT2H, and invalid strings were treated as missing. For each split, the TF–IDF vocabulary, imputer, numeric scaler, category encoder, image scaler and outcome scaler were fitted only on training rows.

Multimodal branch

Text, image and metadata inputs were projected to 64-dimensional tokens. Multi-head cross-modal attention produced a fused representation. The exploratory CC-DiffNet architecture included a cultural-presence head, a cross-covariance penalty between nominal cultural and non-cultural latents, reconstruction loss and a Gaussian noise-prediction objective inspired by diffusion cross-attention [25,26]. These terms were architectural hypotheses, not evidence of causal identification. Three regression outputs were optimized jointly. AdamW, gradient clipping and validation-loss early stopping were used. The frozen configuration used learning rate

5×10−4, diffusion weight 0.005, cultural weight 0.02, covariance weight 0.1, adversarial weight 0, batch size 128 and a maximum of 100 epochs with patience 15.

Component ablations retained the same splits, hyperparameters and early-stopping rule. NoCult set the cultural-supervision weight to zero, NoDiff removed the noise-prediction objective, NoVisual zeroed image embeddings and availability masks, and NoText zeroed all TF–IDF inputs before construction of the training, validation and test loaders. For NoText, validation and test predictions were generated in separate deterministic runs for each of the five seeds, and the no-text hybrid was rebuilt using validation-only convex weights with the unchanged structured expert.

XGBoost expert and convex calibration

XGBoost [40] used the same training-only structured representation. Six candidate configurations varied tree number, depth and learning rate; one configuration was selected per outcome by minimum validation RMSE. The hybrid prediction was

$$\hat{y}_{hybrid} = w\,\hat{y}_{CC} + (1 - w)\,\hat{y}_{XGB}, \quad w \in \{0, 0.01, \ldots, 1\}. \quad (1)$$

The weight minimizing validation RMSE was applied once to the corresponding test set. Test outcomes were not used for tree-configuration or blend-weight selection.

Ridge regression supplied low-capacity text-only, metadata-only and text-plus-metadata baselines. For each outcome and split, the regularisation coefficient was selected from {0.1, 1, 10, 100} by validation RMSE and applied once to the test partition. The same channel groups and train-fitted feature builder were used as in the main analysis.

Predictive-fit and permutation-importance diagnostics

For the prespecified seed-20260807 split, hybrid predictions were aggregated into 28×28 observed-versus-predicted bins for each outcome before export as figure source data; video identifiers and row-level predictions were not exported. To assess structured-feature dependence,

all transformed columns belonging to one source feature group were jointly permuted in the XGBoost test matrix while other groups were held fixed. Relative importance was defined as (permuted RMSE − baseline RMSE) / baseline RMSE and averaged over 20 deterministic permutations. This test-outcome diagnostic was computed only after model selection and was not used for hyperparameter tuning or ensemble weighting. Because correlated predictors can share or substitute predictive information, these importances are descriptive and non-causal.

Scope and geographic validation

For strict short-video validation, ISO-8601 durations were converted to seconds and all records with duration ≤60 s were retained. The full frozen training, preprocessing, early-stopping, XGBoost-selection and validation-only blending pipeline was rerun with the current code for the same five seeds. Split construction remained channel grouped, and train, validation and test channel sets were mutually exclusive in all ten neural evaluation manifests (validation and test runs for five seeds).

For geographic validation, records without channel-country metadata were excluded. All channels assigned to the target country were placed in the test partition; channels from the remaining known countries were divided into channel-disjoint training and validation partitions. The US and India were selected before evaluation because they were the two largest known-country groups. These single-seed analyses assess transportability and were not pooled with the five-seed channel tests.

For duplicate-title sensitivity analysis, titles were Unicode-normalised, lower-cased, stripped of punctuation and collapsed to single spaces. Test records whose normalised title exactly matched any training or validation title in the same split were removed without refitting, and metrics were

recomputed on the retained test records. This diagnostic addresses literal title reuse; it does not detect semantic paraphrases.

Statistical analysis

MAE, RMSE, R^2 and Spearman correlation were computed separately for each outcome and randomized split. Videos were the prediction units, while channel grouping prevented a channel from appearing in more than one partition within a split. The five split seeds are dependent re-partitions of one query-based dataset rather than independent experimental replications. We therefore report split-wise values, arithmetic means, standard deviations, win counts and descriptive 95% t intervals across splits with four degrees of freedom. We do not interpret these intervals as population-sampling confidence intervals and do not use split-seed p values for confirmatory claims. Country, chronological, missingness and latent-probe analyses used the prespecified seed 20260807 and are reported as targeted descriptive or bootstrap audits.

For uncertainty analysis, the 90th-percentile absolute residual was estimated on each validation partition using the finite-sample split-conformal rank ceil((n+1)×0.90) and applied symmetrically on the log-outcome scale. Selective-risk curves ranked test videos by absolute disagreement between the neural branch and XGBoost expert; RMSE was then recomputed from 20% through 100% retained coverage. Neither calibration radii nor selection order used test outcomes. Current neural runs used Python 3.13.5, pandas 2.3.1, scikit-learn 1.7.1 and PyTorch 2.7.1 with CUDA 11.8; summary analyses used SciPy 1.16.1 and the metadata expert used XGBoost 3.0.5.

Ethics declaration

This study analysed publicly accessible YouTube video metadata and aggregate public engagement counters and did not recruit, intervene with or obtain private data from human

participants. Comment text and comment-author identifiers were not used in the reported models. AUTHOR_CONFIRMATION_NEEDED: Confirm the applicable institutional ethics-review or exemption wording before submission.

Funding

This work received no specific funding from any public, commercial or not-for-profit funding agency.

Data availability

Aggregate result tables and source data underlying all figures and supplementary tables, together with the data schema, query protocol, frozen seed definitions and analysis code, have been prepared for deposit in a persistent repository (AUTHOR_INPUT_NEEDED: insert repository name and DOI). Video-level YouTube API data—including video and channel identifiers, titles and descriptions, channel metadata, public counters linked to identifiers and comment-derived records—are not redistributed because they are third-party API Data subject to the YouTube API Services Terms of Service and Developer Policies, including credential-security and data refresh or deletion requirements. Raw audiovisual content was neither archived nor used in the reported models. Researchers may reproduce collection using the documented queries, their own credentials and current API responses, subject to the then-current YouTube terms; resulting sample sizes and public counters may differ because resources and engagement values change over time.

Code availability

All custom preprocessing, annotation-audit, model-training, baseline, calibration and statistical-summary scripts, together with aggregate result tables and figure source data, are included in the versioned reproducibility archive (AUTHOR_INPUT_NEEDED: insert the persistent repository

DOI or versioned URL after deposit). The archive includes a file-level SHA-256 manifest and explicitly excludes API credentials and video-level third-party platform data.

References

1. Richards, G. Cultural tourism: A review of recent research and trends. *J. Hosp. Tour. Manag.* **36**, 12–21 (2018). https://doi.org/10.1016/j.jhtm.2018.03.005

2. Leung, D., Law, R., van Hoof, H. & Buhalis, D. Social media in tourism and hospitality: A literature review. *J. Travel Tour. Mark.* **30**, 3–22 (2013). https://doi.org/10.1080/10548408.2013.750919

3. Xiang, Z. & Gretzel, U. Role of social media in online travel information search. *Tour. Manag.* **31**, 179–188 (2010). https://doi.org/10.1016/j.tourman.2009.02.016

4. Munar, A. M. & Jacobsen, J. K. S. Motivations for sharing tourism experiences through social media. *Tour. Manag.* **43**, 46–54 (2014). https://doi.org/10.1016/j.tourman.2014.01.012

5. Uşaklı, A., Koç, B. & Sönmez, S. How 'social' are destinations? Examining European DMO social media usage. *J. Destin. Mark. Manag.* **6**, 136–149 (2017). https://doi.org/10.1016/j.jdmm.2017.02.001

6. Blanco-Moreno, S., González-Fernández, A. M., Muñoz-Gallego, P. A. & Casaló, L. V. Understanding engagement with Instagram posts about tourism destinations. *J. Destin. Mark. Manag.* **34**, 100948 (2024). https://doi.org/10.1016/j.jdmm.2024.100948

7. Guo, S., Wang, D. & Deng, N. Deciphering the myths of engagement with destination short video: the perspective of modality. *Tour. Manag.* **113**, 105286 (2026). https://doi.org/10.1016/j.tourman.2025.105286

8. Deng, N., Niu, Y., Xiong, T., Yang, J. & Xu, X. What makes tourism short video engaging? A machine learning perspective. *J. Hosp. Tour. Res.* **50**, 832–854 (2026). https://doi.org/10.1177/10963480251385721

9. Tan, J. *et al.* Multimodal destination image and user engagement: a sequential research design. *Tour. Manag.* **111**, 105209 (2025). https://doi.org/10.1016/j.tourman.2025.105209

10. Dolan, R., Conduit, J., Frethey-Bentham, C., Fahy, J. & Goodman, S. Social media engagement behavior. *Eur. J. Mark.* **53**, 2213–2243 (2019). https://doi.org/10.1108/EJM-03-2017-0182

11. Tafesse, W. Content strategies and audience response on Facebook brand pages. *Mark. Intell. Plan.* **33**, 927–943 (2015). https://doi.org/10.1108/MIP-07-2014-0135

12. Cho, M., Jeong, D. & Park, E. AMPS: Predicting popularity of short-form videos using multi-modal attention mechanisms in social media marketing environments. *J. Retail. Consum. Serv.* **78**, 103778 (2024). https://doi.org/10.1016/j.jretconser.2024.103778

13. Le, T. *et al.* EnTube: Exploring key video features for advancing YouTube engagement. *Entertain. Comput.* **53**, 100934 (2025). https://doi.org/10.1016/j.entcom.2025.100934

14. Li, D. *et al.* Delving deep into engagement prediction of short videos. in *Computer Vision – ECCV 2024* 295–312 (Springer, 2025). https://doi.org/10.1007/978-3-031-72949-2_17

15. Li, D. *et al.* VQualA 2025 challenge on engagement prediction for short videos: Methods and results. *arXiv* 2509.02969 (2025). https://doi.org/10.48550/arXiv.2509.02969

16. Sun, W. *et al.* Engagement prediction of short videos with large multimodal models. *arXiv* 2508.02516 (2025). https://doi.org/10.48550/arXiv.2508.02516

17. Baltrušaitis, T., Ahuja, C. & Morency, L.-P. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **41**, 423–443 (2019). https://doi.org/10.1109/TPAMI.2018.2798607

18. Tsai, Y.-H. H. *et al.* Multimodal transformer for unaligned multimodal language sequences. in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* 6558–6569 (2019). https://doi.org/10.18653/v1/P19-1656

19. Vaswani, A. *et al.* Attention is all you need. in *Advances in Neural Information Processing Systems 30* 5998–6008 (2017). https://proceedings.neurips.cc/paper/7181-attention-is-all-you-need

20. Covington, P., Adams, J. & Sargin, E. Deep neural networks for YouTube recommendations. in *Proceedings of the 10th ACM Conference on Recommender Systems* 191–198 (2016). https://doi.org/10.1145/2959100.2959190

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

22. Figueiredo, F., Benevenuto, F. & Almeida, J. M. The tube over time: Characterizing popularity growth of YouTube videos. in *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining* 745–754 (2011). https://doi.org/10.1145/1935826.1935925

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

24. Borghol, Y., Ardon, S., Carlsson, N., Eager, D. & Mahanti, A. The untold story of the clones: Content-agnostic factors that impact YouTube video popularity. in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1186–1194 (2012). https://doi.org/10.1145/2339530.2339717

25. Yang, T., Lan, C., Lu, Y. & Zheng, N. Diffusion model with cross attention as an inductive bias for disentanglement. *Adv. Neural Inf. Process. Syst.* **37**, 82465–82492 (2024). https://doi.org/10.52202/079017-2622

26. Ho, J., Jain, A. & Abbeel, P. Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* **33**, 6840–6851 (2020). https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html

27. Wang, Y. & Jordan, M. I. Desiderata for representation learning: A causal perspective. *J. Mach. Learn. Res.* **25**, 1–65 (2024). https://www.jmlr.org/papers/v25/21-107.html

28. Li, X., Sun, S. & Feng, R. Causal representation learning via counterfactual intervention. *Proc. AAAI Conf. Artif. Intell.* **38**, 3234–3242 (2024). https://doi.org/10.1609/aaai.v38i4.28108

29. Geirhos, R. *et al.* Shortcut learning in deep neural networks. *Nat. Mach. Intell.* **2**, 665–673 (2020). https://doi.org/10.1038/s42256-020-00257-z

30. Roberts, D. R. *et al.* Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* **40**, 913–929 (2017). https://doi.org/10.1111/ecog.02881

31. Varoquaux, G. Cross-validation failure: Small sample sizes lead to large error bars. *NeuroImage* **180**, 68–77 (2018). https://doi.org/10.1016/j.neuroimage.2017.06.061

32. Koh, P. W. *et al.* WILDS: A benchmark of in-the-wild distribution shifts. in *Proceedings of the 38th International Conference on Machine Learning* 5637–5664 (PMLR, 2021). https://proceedings.mlr.press/v139/koh21a.html

33. Ovadia, Y. *et al.* Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. *Adv. Neural Inf. Process. Syst.* **32**, 13969–13980 (2019). https://proceedings.neurips.cc/paper/2019/hash/8558cb408c1d76621371888657d2eb1d-Abstract.html

34. Angelopoulos, A. N. & Bates, S. Conformal prediction: A gentle introduction. *Found. Trends Mach. Learn.* **16**, 494–591 (2023). https://doi.org/10.1561/2200000101

35. Barber, R. F., Candès, E. J., Ramdas, A. & Tibshirani, R. J. Conformal prediction beyond exchangeability. *Ann. Stat.* **51**, 816–845 (2023). https://doi.org/10.1214/23-AOS2276

36. Google for Developers. YouTube Data API v3: Videos resource. https://developers.google.com/youtube/v3/docs/videos (2026).

37. Google for Developers. YouTube Analytics and Reporting APIs: Metrics. https://developers.google.com/youtube/analytics/metrics (2026).

38. Salton, G. & Buckley, C. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manag.* **24**, 513–523 (1988). https://doi.org/10.1016/0306-4573(88)90021-0

39. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. in *2016 IEEE Conference on Computer Vision and Pattern Recognition* 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90

40. Chen, T. & Guestrin, C. XGBoost: A scalable tree boosting system. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (2016). https://doi.org/10.1145/2939672.2939785

Author contributions

Insert CRediT-aligned contributions for conceptualization, data curation, methodology, software, validation, visualization, writing and supervision.

Competing interests

The author(s) declare no competing interests. AUTHOR_CONFIRMATION_NEEDED.