Predicting YouTube Video Popularity Using Multi-Aspect Content and Engagement Features

Qiutong SHEN † and Masaomi KIMURA ‡

† Data Science/Engineering Lab, Shibaura Institute of Technology, Tokyo, Japan
†† University Malaysia Kelantan, Malaysia
E-mail: † ma24006@sic.shibaura-it.ac.jp, ‡ masaomi@sic.shibaura-it.ac.jp

Abstract Understanding which factors influence video popularity is important for online video platforms. This study analyzes YouTube video view using a dataset of trending videos collected from ten countries. After basic data cleaning and consolidation, multi aspect features are constructed to describe video content and user engagement, including features extracted from titles and tags, together with temporal and video level attributes. Video popularity is examined from different perspectives, and view prediction is treated as a regression task to study how content related, and time related factors are associated with viewer attraction. Due to the long-tailed distribution of view, the data is divided into low view and high view subsets, which are modeled using different prediction models. Videos with low view are modeled using a multilayer perceptron based on structured features. Videos with higher view are modeled using a collaborative tag aware graph neural network, which captures relationships between videos through shared tags and additional contextual information such as category and country. Experimental results show that the proposed approach improves prediction accuracy compared to baseline models and helps identify factors associated with video popularity on YouTube.

Keyword Social Media, Video, View Prediction, YouTube

1. Introduction

Video popularity is an important topic for online video platforms, as views reflect user attention and content visibility. On platforms such as YouTube, a large number of videos compete for limited attention, while only a small portion of videos achieve very high view. As a result, predicting video popularity and understanding which factors influence views remain challenging problems. Video popularity is affected by multiple factors related to content, timing, and user interaction, and the influence of individual features on views can vary across features.

Existing studies have explored video popularity prediction using different features and models, but prediction accuracy remains limited, and video information is often not deeply analyzed from multiple perspectives. In particular, textual information and timing related factors are not always fully considered, and the uneven distribution of video popularity is often not explicitly addressed. In this study, video popularity is analyzed using multi aspect features, and a distribution aware modeling strategy is adopted to better handle videos with different popularity levels.

2. Related Work

Video popularity prediction has been widely studied as a supervised learning problem using content, metadata, and engagement features. Early YouTube-focused studies relied on structured features such as views, comments, ratings and categories, and applied regression or classification models to predict popularity or popularity levels [1, 2, 3]. While these works demonstrated the usefulness of engagement and temporal features, they typically employed a single global model and did not explicitly address the highly long-tailed distribution of video views, leading to biased performance dominated by low-view samples.

To handle heterogeneity in popularity, several studies proposed two-stage or level-based prediction frameworks. Ouyang et al. first predicted future popularity levels and then applied specialized regressors conditioned on level transitions, showing improved accuracy over single model baselines [4]. Related classification approaches also grouped videos into popularity categories before prediction [2]. These methods share our motivation to account for popularity heterogeneity, but their regime definitions are typically heuristic, and routing decisions are optimized for level prediction rather than end-to-end view regression. In contrast, our work defines popularity regimes using data-driven knee-point detection and selects routing thresholds by directly optimizing regression performance.

Another line of research emphasizes temporal dynamics of popularity. SMTPD introduced a large-scale benchmark for temporal popularity prediction and showed

that early popularity trajectories are highly predictive of future views[5]. Knowledge-graph-based temporal models further combined sequential modelling with relational reasoning to predict popularity evolution [6]. While effective in settings with rich temporal signals, these approaches rely on historical trajectories. Our work instead focuses on a static or early-stage setting, where popularity must be inferred from content, engagement, and relational structure without long observation windows.

Recent studies explored richer semantic representations through deep and multimodal models. Some approaches incorporated visual cues from video covers together with textual features, demonstrating that visual–textual fusion can improve popularity prediction [7]. Others employed heavy multimodal pipelines or retrieval augmented models to capture cross-video semantic similarity [8,9]. Very recent work further leveraged large language models to predict popularity and provide explanations, often using newly collected large-scale datasets [10]. While these methods enhance semantic expressiveness, they typically apply a single complex model to all videos. Our approach instead adopts lightweight textual semantics and emphasizes regime-dependent modeling.

Graph-based methods introduce relational structure into popularity prediction by modeling inter-video dependencies. GraphInf used graph convolutional networks to capture influence among videos in short-video networks [11], while collaborative tag-aware graph neural networks showed that tag-mediated message passing effectively captures implicit similarity in long-tail recommendation settings [12]. These works motivate our CTGNN design, but existing graph popularity models generally apply graph inference uniformly to all items. In contrast, we restrict graph modeling to the high-view regime, where relational neighbourhoods are sufficiently dense, and fuse multiple graph-derived and direct feature representations through a gating mechanism.

Interpretability-oriented studies further analysed feature contributions in popularity prediction. Xie et al. highlighted the dominant role of engagement-related signals in deep learning models for YouTube viewership prediction [13]. Our permutation feature importance analysis for low-view videos is consistent with these findings and provides additional insight by linking feature dominance to regime-specific data characteristics. More broadly, practical lessons from large-scale click prediction systems emphasize robust modeling under extreme imbalance and the importance of strong feature baselines, while advances in weakly supervised consistency learning provide methodological background for structure aware modeling under limited supervision [14].

Overall, prior work has explored feature-based, temporal, multimodal, and graph-based approaches to video popularity prediction. Our contribution differs by explicitly modeling the long-tailed distribution through a distribution aware routed framework that aligns model choice with data structure: a MLP for low-view videos and a collaborative tag aware graph neural network for high-view videos, with data-driven regime definition and prediction-oriented routing.

3. Methodology

3.1 Overview

We propose a routed prediction framework to address the long-tailed nature of video view. Given an input video, a routing classifier estimates the probability that the sample belongs to a high-view regime. The sample is then routed to one of two specialized regressors: (i) an MLP predictor for lower-view videos using structured features, and (ii) a collaborative tag aware graph neural network (CTGNN) for higher-view videos that leverages relational signals induced by shared tags and contextual nodes. The full model is trained and evaluated under a consistent high/low definition and routing policy, with selection of thresholds and hyperparameters reported in the experimental section.

3.2 Feature Representation

Each video is represented by multi-aspect features capturing engagement, video metadata, textual semantics, and external attention.

Structured features include engagement statistics (e.g., likes, dislikes, comment count), video metadata (e.g., duration, title length, tag count), and time related attributes derived from publishing and trending timestamps (e.g., interval from publish to trending). A binary weekend indicator is incorporated only in the MLP branch to provide a lightweight temporal cue for low-view prediction, while the graph model focuses on relational/content signals and contextual nodes; in practice, such coarse temporal indicators tend to be weakly discriminative for the high-view regime and can be partially absorbed by category/country context.

To capture semantic information beyond scalar features, we include (i) dense textual embeddings derived from title content (e.g., word2vec-based embeddings) and (ii) category representations. Country information is used

explicitly for the routing and MLP predictors (e.g., via one-hot encoding) and is also modelled in CTGNN through a dedicated country node type.

To quantify external media attention related to video content, we construct a keyword set from video titles and use these keywords to retrieve the number of related news mentions within a fixed time window around the publish date. The maximum mention count among the retained keywords is used as a compact proxy feature for external attention. The concrete window length and data source used for mention retrieval are described in the experimental setup.

Titles may appear in multiple languages. For keyword extraction, we prioritize the original title when its language is supported; otherwise, we fall back to an English translated title to ensure consistency for downstream processing.

We apply YAKE to extract candidate key phrases from each title. Stop words are constructed as the union of language-specific stopword lists across supported languages. Extracted phrases are normalized by lowercasing and punctuation removal, and phrases that are empty, too short, or stopwords are discarded. YAKE is configured to extract up to trigrams, reflecting the observation that many salient title concepts are multi-word expressions rather than isolated tokens.

In addition to per-title extraction, we construct a corpus TF-IDF vocabulary over all unique titles. TF-IDF is computed using an n-gram range of 1–3 to capture both single word keywords and multi-word entities (e.g., person names, events, products). This 1–3 gram design is intentional: unigram-only vocabularies often fragment important phrases, whereas trigrams allow the model to preserve semantically coherent expressions that better align with how titles convey topics.

To improve robustness against noisy or overly generic terms, the final keyword set for each title is defined as the intersection between (i) the filtered YAKE keywords extracted for that title and (ii) a global set of high-importance TF-IDF n-grams. This design keeps keywords that are both salient locally (YAKE) and informative globally (TF-IDF), reducing sensitivity to either method's failure modes.

3.3 Proposed Model

Our proposed model consists of three components: a routing classifier, a low view regressor based on a multilayer perceptron, and a collaborative tag aware graph neural network for high view prediction, as illustrated in Figure 3.3. The routing classifier outputs a probability p_{high} indicating whether a video belongs to the high view regime. During inference, p_{high} is compared with a routing threshold to determine which prediction branch is activated. The boundary between high and low view regimes is determined in a data-driven manner via knee-point detection on the training view distribution, while the routing threshold is selected based on validation performance.

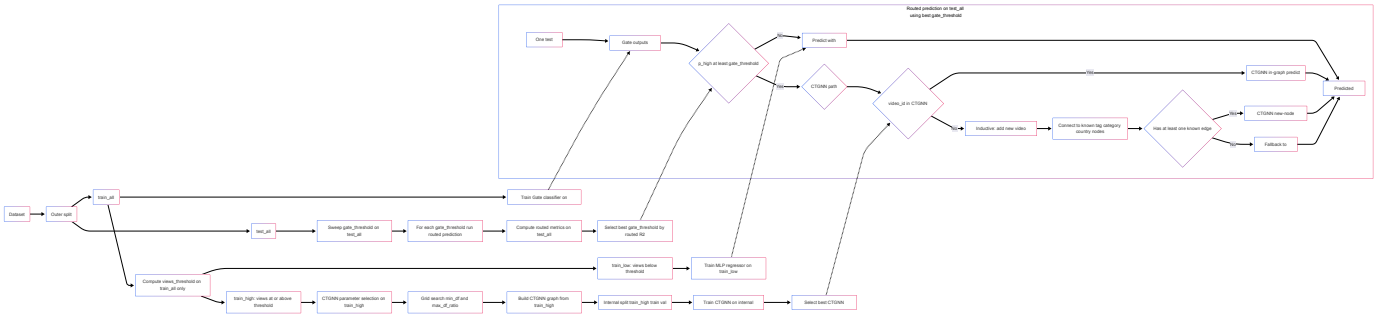


Table 3.3 Overall Workflow of the Gate-Based Routing Framework

3.3.1 Low-view Regressor (MLP)

For lower-view videos, we use an MLP regressor trained on standardized structured features. Figure 3.3.1 summarizes the architecture: repeated Linear → BatchNorm → LeakyReLU → Dropout blocks followed by a final linear output layer. Both inputs and targets are standardized on the training low split, and predictions are inverse transformed to the original view scale for reporting.

Figure3.3.1 Flowchart of the MLP Regressor

3.3.2 High-view Regressor (CTGNN)

CTGNN operates on a heterogeneous graph with four node types:

- video nodes: one node per unique video,
- tag nodes: representing tags associated with videos,
- category nodes: representing the video category,
- country nodes: representing the market/context where the video is observed.

Edges are created from each video to its associated tag, category, and country nodes, together with reverse edges to enable bidirectional message passing.

To mitigate noise from extremely rare tags and reduce the dominance of overly frequent tags, we filter tags using document frequency (DF) constraints computed over the set of videos used to build the graph. Let N be the number of videos and $DF(t)$ the number of videos containing tag t. A tag is retained if:

$$\min\_df \leq DF(t) \leq [\max\_df\_ratio \cdot N]$$

The lower bound min_df removes extremely rare tags that are unlikely to yield reliable collaborative signals, while the upper bound max_df_ratio suppresses overly frequent tags that act as hubs and introduce indiscriminate connectivity.

CTGNN derives multiple source-wise representations for each video through heterogeneous message passing over tag, category, and country nodes, inspired by collaborative tag aware graph learning that leverages tags to capture implicit similarity among items [12].

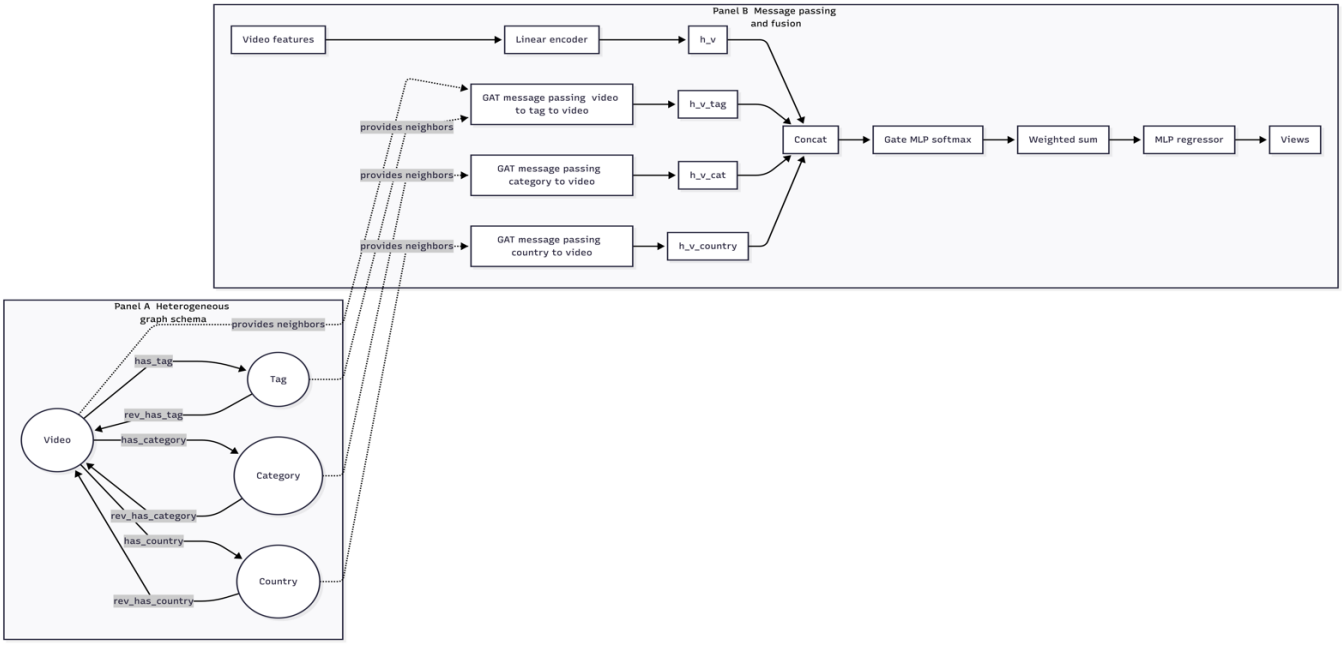


Figure 3.3.2 CTGNN Architecture and Message Passing Flow

As illustrated in Figure 3.3.2, collaborative information is propagated by performing message passing from video nodes to tag nodes and subsequently from tag nodes back to video nodes. This mechanism enables videos sharing common tags to exchange information indirectly. Category and country nodes supply contextual features, which are fused through a gating network for prediction.

CTGNN produces four video representations from different information sources: the tag collaborative representation h_t, the category contextual representation h_c, the country contextual representation h_{co}, and the direct feature representation h_v obtained from the video encoder. Since the reliability of these sources can vary across videos (e.g., some videos have sparse tag connections while others benefit strongly from tag-mediated neighbors), we adopt a learnable gating mechanism to adaptively weight them on

a per-video basis.

Concretely, we first concatenate the four representations and feed them to a small gating network $g(\cdot)$ that outputs a 4-dimensional weight vector:

$$\alpha = softmax(g([h_t; h_c; h_{co}; h_v])), \alpha \in \mathbb{R}^4$$

The function $g(\cdot)$ acts as a source-wise gating mechanism that adaptively determines the relative importance of different information sources for each video. By taking the concatenated representations as input, it produces fusion weights that allow the model to emphasize reliable collaborative or contextual signals while attenuating noisy or weak ones. Unlike node attention used during message passing, this gating mechanism operates at the representation level and controls how multiple source-specific embeddings are combined for final prediction.

Here, α_i indicates the relative importance assigned to the source for the current video. The softmax operation ensures $\alpha_i \geq 0$ and $\sum_{i=1}^{4} \alpha_i = 1$, making the fusion a convex combination and keeping the weights interpretable. The fused representation is then computed as:

$$h_{fuse} = \alpha_1 h_t + \alpha_2 h_c + \alpha_3 h_{co} + \alpha_4 h_v$$

Finally, a lightweight MLP is adopted as a regression head to map the fused representation h_{fuse} to the predicted view count. As CTGNN focuses on representation learning over the heterogeneous graph, the MLP acts as a standard readout module that converts the graph aware embedding into a scalar output. Combined with the gating mechanism, this design allows the model to emphasize relational and contextual signals when they are informative, while naturally falling back to the direct feature pathway when graph evidence is weak or unavailable.

4. Experiment

4.1 Dataset and Preprocessing

The dataset is based on the Trending YouTube Video Statistics collection from Kaggle, which provides daily statistics for trending YouTube videos. Trending videos collected between November 14, 2017 and June 14, 2018 are used in this study. Data from ten countries are included, namely Canada, Germany, France, the United Kingdom, India, Japan, South Korea, Mexico, Russia, and the United States. Each record contains basic video metadata such as a unique video identifier (video_id), title, publish time, trending date, category, and engagement statistics.

Since the same video has appear multiple times across different trending dates in the dataset, the raw dataset contains duplicate records for a single video. In this study, a collaborative tag aware graph neural network is constructed using video_id as the unique node identifier. Without proper deduplication, records from different trending dates would be merged into the same node, causing features, labels, and edges from different time points to be incorrectly combined. To ensure semantic consistency in graph construction, duplicate videos with the same video_id are removed by retaining only the record with the most recent trending date. Videos with missing or incomplete information are also excluded.

After data cleaning, a set of multi-aspect features is constructed to describe different properties of each video. These features include information related to video duration and the market where the video trended, as well as temporal characteristics derived from publish time and trending date, such as the number of days from publishing to trending, whether the video was published on a weekend, and the day of the week of publication. Text based features are extracted from video titles and descriptions, including sentiment scores, title length measured by word count, and the number of tags. For videos whose titles are not in English, titles are translated into English to ensure consistency in text analysis.

We extract compact keywords from each video title using YAKE, and further filter them using global TF-IDF statistics.

YAKE extraction. For each row, we select the original title if its language is supported; otherwise, we use the English translated title. We extract up to 3 keywords per video using YAKE with n=3 (allowing up to trigrams), returning the top candidates after stopword and punctuation filtering.

Global TF-IDF vocabulary. To obtain a global set of salient phrases, we compute TF-IDF on de-duplicated titles using $ngram\_range =$ (1,3). We then rank terms by their average TF-IDF across documents and select the top-N terms. The final keyword set for each video is the intersection between its YAKE keywords and this global top-N vocabulary, plus any global top N terms that appear in the title.

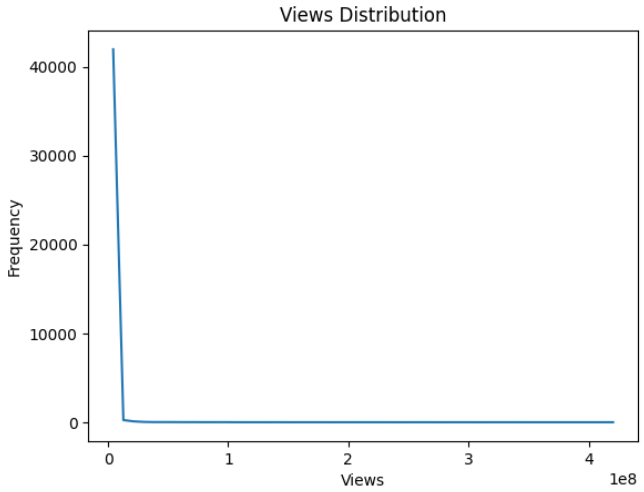
We sweep N and measure how much the global vocabulary covers the extracted YAKE keywords. We select N at the point where the marginal increase in coverage becomes small (coverage gain < 0.01 between consecutive candidates), and use a fixed N (e.g., 30k) for stable downstream processing.

4.2 Training Protocol and Motivation

In preliminary experiments, we observed that using a single MLP regressor to predict video views often leads to unsatisfactory performance.

Due to the long-tailed nature of the view distribution as shown in Figure 4.2, the model is dominated by low-view samples, which results in biased predictions and systematic underestimation for highly popular videos.

Figure 4.2 Views Distribution



These observations motivate a regime-based modeling strategy, where videos with different popularity levels are handled by specialized models.

The gate model is trained on the full training set using standardized features, with a fixed internal train–validation split used for validation and performance reporting.

MLP is trained exclusively on the training low subset, allowing the model to specialize in predicting view for low popularity videos.

CTGNN is trained on the training high subset. Within this subset, 80% of the video nodes are used for parameter learning, while the remaining 20% are held out for validation.

4.3 High/Low Definition: views_threshold from Knee Point

To define high-view vs low-view regimes in a data-driven manner, we compute a threshold on the training view distribution using a knee-point method in log scale. The knee index is obtained by the maximum distance to the line connecting the endpoints of the normalized curve. Videos with $views \geq threshold$ are treated as high-view.

4.4 CTGNN Tag Filtering Parameter Selection (min_df, max_df_ratio)

We select min_df and max_df_ratio through a grid search. For each setting, we evaluate:

- validation R^2 / RMSE on training high internal validation nodes,
- graph structure health: (i) vocabulary size, (ii) fraction of videos with zero kept tags, and (iii) average kept tags per video.

We first filter settings that violate structure constraints (e.g., too many zero tag videos or too small tag vocabulary). Among feasible settings, we select a near best R^2 configuration (within a slack) and break ties using a composite score that rewards both predictive performance and healthy graph connectivity.

	min_df	max_df_ratio	val_r2	...	vocab_size	mean_kept_tags	score
4	20	0.05	0.934093	...	214	3.112689	2.592880
2	10	0.05	0.917212	...	539	5.115437	3.025861
5	20	0.10	0.838886	...	217	3.331196	2.522289
3	10	0.10	0.825588	...	542	5.333944	2.946466
0	5	0.05	0.819763	...	1453	7.774622	3.351136
1	5	0.10	-0.168826	...	1456	7.993129	2.368985

4.5 Routing Threshold Selection (gate_threshold)

We sweep the routing threshold $threshold_g$ on the outer test set over a predefined range and select the value that maximizes routed regression R^2. This choice reflects the goal of optimizing end-to-end prediction quality rather than pure classification accuracy.

gate_threshold	gate_acc	...	routed_r2	TN_FP_FN_TP
0.93	0.958643	...	0.930040	[7759, 146, 205, 377]
0.95	0.959350	...	0.929946	[7777, 128, 217, 365]
0.94	0.958996	...	0.929924	[7770, 135, 213, 369]
0.92	0.958053	...	0.929762	[7748, 157, 199, 383]
0.91	0.958525	...	0.929717	[7746, 159, 193, 389]
0.90	0.957936	...	0.929673	[7737, 168, 189, 393]
0.87	0.957111	...	0.929381	[7717, 188, 176, 406]
0.85	0.955815	...	0.929210	[7697, 208, 167, 415]
0.83	0.955108	...	0.928965	[7681, 224, 157, 425]
0.80	0.954519	...	0.928951	[7666, 239, 147, 435]
0.70	0.948627	...	0.928245	[7591, 314, 122, 460]

4.6 Evaluation

Metrics include RMSE, MAE, and R^2, reported overall and separately for true low and true high subsets. We additionally provide gate performance (accuracy, AUC, confusion matrix) and route counts to analyze model behavior.

On the outer test set, the MLP only model achieves strong performance on the True LOW subset with an RMSE of 1.95×10^5, an MAE of1.23×10^5, and an R^2 of 0.729.

For the True HIGH subset, the CTGNN only model achieves an R^2 of 0.972 with an RMSE of 5.99×10^6 and an MAE of 2.96×10^6, with full inductive coverage.

Using the routed model with the best gate threshold = 0.93, the gate achieves an accuracy of 0.959 and an AUC of 0.972. The overall routed regression attains an RMSE of 1.92×10^6, an MAE of 3.62×10^5 , and an R^2 of 0.930.

When evaluated by true popularity regimes, the routed model yields an RMSE of 4.16×10^5 ($R^2 = -0.227$) on True LOW videos and an RMSE of 6.00×10^6 ($R^2 = 0.926$) on True HIGH videos. In total, 7,964 samples are routed to

the MLP and 523 samples are handled by CTGNN via inductive inference.

4.7 Feature Importance

To analyze the contribution of individual input features to low view prediction, we compute single feature permutation importance for the MLP model on the outer test True LOW subset, with detailed results summarized in Table 4.7. A baseline root mean squared error is obtained using the original test features. For each feature independently, its values are randomly permuted across samples while all other feature columns are kept unchanged. Permutation based feature importance is adopted because it provides a model agnostic measure of feature contribution by quantifying the degradation in predictive performance when the information carried by a feature is disrupted, while keeping its marginal distribution unchanged. This permutation breaks the association between the selected feature and the target variable. The model is evaluated on the permuted data, and the increase in prediction error relative to the baseline is recorded. This procedure is repeated multiple times with different random permutations, and the mean and standard deviation of the resulting RMSE increase are reported as the feature's importance score.

Table 4.7 Feature Permutation Importance of the MLP on the True LOW Subset

feature	rmse_increase_mean	rmse_increase_std	baseline_rmse
likes	1.44E+05	8.26E+02	1.95E+05
dislikes	8.99E+04	7.40E+02	1.95E+05
tag_count	1.64E+04	1.18E+03	1.95E+05
Duration	1.47E+04	7.56E+02	1.95E+05
title_length	1.45E+04	1.06E+03	1.95E+05
comment_count	1.33E+04	2.00E+02	1.95E+05
sentiment_score_description	9.88E+03	2.19E+02	1.95E+05
sentiment_score	5.22E+03	2.64E+02	1.95E+05
interval	4.48E+03	3.33E+02	1.95E+05
is_weekend	2.06E+03	3.18E+02	1.95E+05
max_keyword_mentions	1.84E+03	4.10E+02	1.95E+05

Engagement related features, including likes, dislikes, and comment count, lead to the largest increases in RMSE when permuted, indicating that the MLP relies primarily on observable signals of audience interest for low-view prediction.

In this context, both positive and negative feedback are informative, as they reflect that viewers were sufficiently interested in the video to watch it and express an opinion. In contrast, content semantic features, such as word embeddings and sentiment scores, exhibit relatively lower importance, suggesting that early audience interest provides more reliable predictive cues than intrinsic content semantics in the low-view regime.

5. Conclusion

This paper investigates YouTube video popularity prediction using multi-aspect content and engagement features under a highly skewed view distribution. To address the heterogeneous characteristics of videos at different popularity levels, we propose a routed framework that dynamically selects between a feature-based regressor and a graph model.

Extensive experiments show that a feature-based MLP is sufficient for low-view videos, where engagement statistics and basic content attributes dominate prediction and relational structure is sparse or noisy. In contrast, high-view videos exhibit stronger and more consistent relational patterns, such as shared tags, categories, and countries, forming dense neighborhoods that cannot be effectively captured by independent feature modeling. For this regime, the collaborative tag aware graph neural network (CTGNN) becomes necessary to exploit cross-video dependencies through message passing on a heterogeneous graph.

By routing videos to different predictors according to their estimated popularity regime, the proposed model effectively matches model capacity with data structure. Further analysis confirms that engagement and attributes primarily drive low-popularity prediction, while relational and contextual information is crucial for modeling high-popularity videos. Overall, this work demonstrates that adaptive integration of content, engagement, and relational signals provides a principled and practical solution for large-scale popularity prediction.

References

[1] Mekouar, S., Zrira, N. & Bouyakhf, E.-H. Popularity prediction of videos in YouTube as case study: a regression analysis study. In: Proceedings of the 2nd International Conference on Big Data, Cloud and Applications. https://doi.org/10.1145/3090354.3090406 (2017).

[2] Li, Yuping, Kent X. Eng and Liqian Zhang. "YouTube Videos Prediction: Will this video be popular?" (2019).

[3] He, X., et al.: Practical lessons from predicting clicks on ads at Facebook. Proceedings of the Eighth International Workshop on Data Mining for Online Advertising pp. 1–9 (2014)

[4] S. Ouyang, C. Li, and X. Li, "A Peek Into the Future: Predicting the Popularity of Online Videos," IEEE Access, vol. 4, pp. 3026–3033, Jun. 2016, doi: 10.1109/ACCESS.2016.2580911.

[5] Y. Xu et al., "SMTPD: A New Benchmark for Temporal Prediction of Social Media Popularity," 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 2025, pp. 18847-18857, doi: 10.1109/CVPR52734.2025.01756.

[6] Liu, P., Yu, Z., Sun, Y., Xi, M. (2023). Video Popularity Prediction Based on Knowledge Graph and LSTM Network. In: Yu, Z., et al. Data Science. ICPCSEE 2023. Communications in Computer and Information Science, vol 1879. Springer, Singapore. https://doi.org/10.1007/978-981-99-5968-6_32

[7] Y. Tian and X. Wang, "Predicting video popularity based on video covers and titles using a multimodal large-scale model and pipeline parallelism," Applied and Computational Engineering, vol. 41, no. 1, pp. 182–189, Feb. 2024, doi: 10.54254/2755-2721/41/20230741.

[8] Haixu Liu, Wenning Wang, Haoxiang Zheng, Penghao Jiang, Qirui Wang, Ruiqing Yan, and Qiuzhuang Sun. 2025. Multi-Modal Video Feature Extraction for Popularity Prediction. arXiv:2501.01422 [cs] doi:10.48550/arXiv.2501.01422

[9] Zhong, T., Lang, J., Zhang, Y., Cheng, Z., Zhang, K., Zhou, F.: Predicting micro-video popularity via multi-modal retrieval augmentation. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 2579–2583 (2024)

[10] Pratik Kayal, Pascal Mettes, Nima Dehmamy, and Minsu Park. 2025. Large Language Models Are Natural Video Popularity Predictors. In Findings of the Association for Computational Linguistics: ACL 2025, pages 11432–11464, Vienna, Austria. Association for Computational Linguistics.

[11] Zhang, Yuchao, Pengmiao Li, Zhili Zhang, Chaorui Zhang, Wendong Wang, Yishuang Ning and Bo Lian. "GraphInf: A GCN-based Popularity Prediction System for Short Video Networks." International Conference on Web Services (2020).

[12] Z. Zhang, Y. Zhang, M. Dong, K. Ota, Y. Zhang, and Y. Ren, "Collaborative tag-aware graph neural network for long-tail service recommendation," IEEE Trans. Serv. Comput., vol. 17, no. 5, pp. 2124–2137, Sep./Oct. 2024, doi: 10.1109/TSC.2024.3349853.

[13] Xie, Jiaheng and Xinyu Liu. "Unbox the Black-Box: Predict and Interpret YouTube Viewership Using Deep Learning." Journal of Management Information Systems 40 (2020): 541 - 579.

[14] M. Xie, J. Xiao, and S.. Huang, "Label-aware global consistency for multi-label learning with single positive labels," in Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS*)*, New Orleans, LA, USA, 2022.