ClipFarm: A Multimodal Framework for Latent Attribute Discovery and Generation in Short-Form Content

Foster Deighton
Western University
fdeighto@uwo.ca

Luke Jang
Western University
hjang55@uwo.ca

Cole Book
Western University
cbook5@uwo.ca

Liam Ma
Western University
lma287@uwo.ca

Pranav Chopra
Western University
pchopr2@uwo.ca

***Abstract*—Short-form video platforms exhibit heavy-tailed engagement distributions and extreme outcome variance, making virality prediction difficult under temporal constraints. We present a deterministic multimodal framework for fixed-horizon virality prediction that constructs temporally consistent labels, extracts modality-specific embeddings using pretrained encoders, and applies reproducible fusion strategies designed for rolling retraining pipelines.**

We evaluate multiple supervised architectures, including a metadata-conditioned gated fusion network that dynamically weights modality contributions. The system supports idempotent daily updates and stable representation alignment across retraining cycles. Finally, we demonstrate how the learned representation space enables downstream tasks such as generative video ranking and content exploration.

I. Introduction

Short-form video platforms such as YouTube Shorts, TikTok, and Instagram Reels have introduced highly compressed content formats with large variability in engagement outcomes. A small fraction of videos accumulate the majority of views, producing heavy-tailed distributions that complicate supervised prediction.

Many existing virality prediction approaches rely on unimodal signals such as text metadata or early engagement statistics and are often trained on temporally inconsistent targets. In contrast, short-form videos inherently combine multiple information channels, including visual motion, audio structure, spoken language, and contextual metadata. Effectively modeling these signals while maintaining temporally consistent targets remains a central challenge.

This work presents a deterministic multimodal framework for fixed-horizon virality prediction built around three design principles:

1) Fixed-horizon labeling to reduce temporal target noise,
2) Reproducible modality-specific representation extraction,
3) Stable fusion mechanisms suitable for rolling retraining.

Based on these principles, we develop a system that supports reproducible multimodal learning and stable daily retraining pipelines.

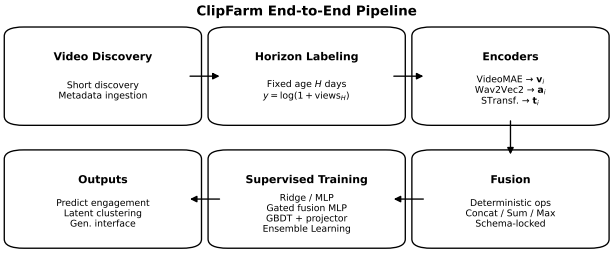

Fig. 1. High-level system architecture from horizon labeling through multimodal fusion and supervised training.

Our contributions are summarized as follows:

- A fixed-horizon labeling framework for constructing temporally consistent virality targets.
- A micro-batch multimodal embedding pipeline that incrementally extracts visual, audio, and textual representations.
- A schema-locked fusion layer supporting multiple deterministic multimodal fusion strategies.
- A metadata-conditioned gated regression model that dynamically weights modality contributions.
- A framework connecting multimodal engagement prediction to downstream generative video ranking and exploration.

II. Background and Related Work

A. Virality Prediction

Prior work on engagement prediction has explored early-trajectory modeling, metadata-based regressors, and text-only semantic analysis. However, many approaches rely on temporally inconsistent targets or incorporate post-publication statistics, leading to label leakage. Fixed-horizon labeling addresses this issue by measuring outcomes at a consistent age, reducing target variance and improving comparability across samples.

B. Multimodal Representation Learning

Recent advances in pretrained models have enabled strong modality-specific encoders for visual (e.g., VideoMAE), audio (e.g., wav2vec2), and textual (e.g., Sentence-Transformer) signals. Multimodal learning methods typically combine these representations through early fusion, late fusion, or learned cross-modal attention mechanisms.

C. Fusion Stability in Production Systems

In continuously updated datasets, representation stability becomes critical. If fusion weights or projection layers are reinitialized across training cycles, identical content may map to different embedding coordinates, reducing comparability over time. Deterministic fusion strategies provide a reproducible alternative that supports stable representation alignment in rolling retraining pipelines.

This work emphasizes representation stability alongside predictive performance.

III. Problem Formulation

Each short-form video is represented by a canonical identifier v_i. For every video we observe:

- Raw visual stream X_i^{video}
- Raw audio waveform X_i^{audio}
- Text metadata and transcript X_i^{text}

We define a fixed-horizon engagement target measured H days after publication:

$$y_i = \log(1 + \text{views}_{i,H})$$

where $\text{views}_{i,H}$ denotes the total view count observed exactly H days after the video is published.

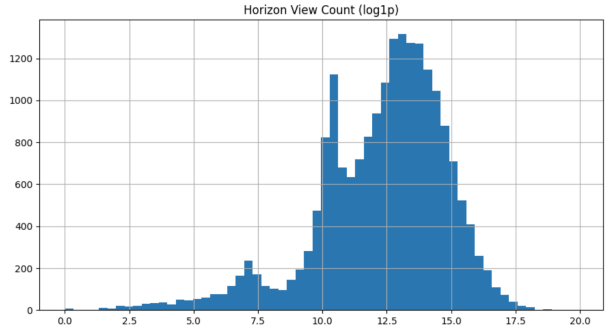

Fig. 2. Distribution of the fixed-horizon engagement target after $\log(1 + x)$ transformation.

Our objective is to learn a predictive function

$$f : (X_i^{video}, X_i^{audio}, X_i^{text}, M_i) \to y_i$$

where M_i represents structured metadata features.

The learning setup must satisfy the following constraints:

1) Labels are temporally consistent.
2) Features do not include post-horizon statistics that could introduce leakage.
3) Representations remain stable under rolling retraining.

IV. Data and Horizon Labeling

Raw view counts vary depending on when they are measured. To reduce target noise, we construct fixed-horizon labels measured at predefined ages (e.g., 7-day and 30-day post-publication).

During each collection cycle, candidate videos are discovered within a narrow publish-time window centered at $(t - H)$ days. The view count observed at collection time becomes the horizon label $\text{views}_{i,H}$, ensuring that all samples are evaluated at a consistent age.

To support daily ingestion without redundant processing, each video is assigned a canonical identifier derived from its URL. A deterministic source hash governs delta processing so that only new or modified items propagate through downstream stages. This design enables idempotent micro-batch updates during repeated collection cycles.

A. Text

Textual information is collected from both platform metadata and automatically generated transcripts. Metadata fields include video titles and descriptions, while transcripts capture the spoken content within each video. Audio tracks are extracted and transcribed using an automatic speech recognition system. All text sources are cleaned and standardized to form a unified textual representation for each video.

B. Audio

Audio signals are extracted directly from the video files. The resulting waveforms capture speech patterns, acoustic structure, and background sound characteristics that may influence viewer engagement.

C. Video

The visual modality is derived from the raw video stream. To preserve temporal structure, frames are sampled uniformly across the video duration rather than selecting a single representative frame. This sampling strategy captures motion dynamics, pacing, and visual transitions that are common in short-form content.

D. Metadata and Early Engagement Signals

In addition to content-based modalities, structured metadata and early engagement signals are collected where available. Metadata include attributes such as video duration and posting context, while early engagement signals reflect initial audience response measured within a short observation window. These features are stored in tabular form and later incorporated alongside learned embeddings during supervised modeling.

V. Multimodal Representation

Each modality is embedded independently using pretrained encoders to produce fixed-dimensional representations for visual, audio, and textual signals.

A. Video Embeddings

Videos are decoded and uniformly sampled to $N = 16$ frames. The sampled frames are processed using a VideoMAE encoder, and the final CLS token representation is extracted:

$$\mathbf{v}_i \in \mathbb{R}^{768}$$

This embedding captures visual structure, motion patterns, and scene composition across the sampled frames.

B. Audio Embeddings

Audio waveforms are resampled to 16 kHz mono and processed using wav2vec2. Temporal features produced by the encoder are mean-pooled to obtain a fixed-length representation:

$$\mathbf{a}_i \in \mathbb{R}^{768}$$

These embeddings capture speech characteristics, rhythm, and background acoustic structure.

C. Text Embeddings

Textual information is derived from platform metadata and automatically generated transcripts. Titles, descriptions, and transcripts are encoded using a SentenceTransformer model. The resulting representations are concatenated and normalized to produce:

$$\mathbf{t}_i \in \mathbb{R}^{768}$$

D. Feature Normalization

All modality vectors are L2-normalized prior to fusion to ensure consistent scaling across encoders.

E. Multimodal Fusion

Fusion combines modality embeddings while preserving representation stability across rolling retraining cycles. Deterministic fusion operators are used to prevent coordinate drift between training updates.

Let $\mathbf{v}_i$, $\mathbf{a}_i$, and $\mathbf{t}_i$ denote video, audio, and text embeddings respectively.

1) Concatenation:

$$\mathbf{f}_i = [\mathbf{v}_i \| \mathbf{a}_i \| \mathbf{t}_i \| m_i]$$

where $m_i \in \{0, 1\}$ indicates text availability.

2) Sum Pooling: Embeddings are zero-padded to equal dimensionality and summed elementwise.

3) Max Pooling: Embeddings are zero-padded and combined using an elementwise maximum operator.

Fusion configurations are schema-locked to ensure consistent feature dimensionality across retraining cycles.

VI. Training Architectures

We evaluate four supervised regression model families under the fixed-horizon labeling framework. All models are trained to predict

$$y_i = \log(1 + \text{views}_{i,H})$$

where $H \in \{7, 30\}$ denotes the prediction horizon in days.

A. Concat-MLP

The `concat_mlp` model consumes the fused multimodal vector $\mathbf{f}_i$ together with structured metadata features. Numeric metadata features are standardized, while low-cardinality categorical attributes are embedded and concatenated.

The final input representation is

$$\mathbf{x}_i = [\mathbf{f}_i \| \mathbf{m}_i^{num} \| \mathbf{m}_i^{cat}]$$

A multilayer perceptron with hidden dimensions $[1024, 512, 256]$, GELU activations, and dropout regularization outputs a scalar prediction in log space.

B. Gated Fusion MLP

To explicitly model modality reliability, we implement a metadata-conditioned gated architecture.

Each modality embedding is first passed through a modality-specific transformation:

$$\mathbf{z}_i^{video}, \mathbf{z}_i^{audio}, \mathbf{z}_i^{text}$$

A gating network conditioned on metadata context produces softmax-normalized weights

$$\boldsymbol{\alpha}_i = \text{softmax}(g(\mathbf{m}_i))$$

The fused representation is then computed as

$$\mathbf{h}_i = \sum_{k \in \{video, audio, text\}} \alpha_{i,k} \mathbf{z}_i^k$$

This representation is passed to a prediction head with hidden layers $[256, 128]$.

The gating mechanism enables dynamic weighting of modality contributions and provides robustness when textual information is unavailable.

C. Ridge Regression

A linear baseline combines fused embeddings with standardized numeric metadata and one-hot encoded categorical features. The regularization coefficient is selected via cross-validation.

D. Gradient Boosted Trees with Learned Projection

To reduce the dimensionality of the fused multimodal representation, we first train a neural projection network mapping $\mathbf{f}_i$ into a lower-dimensional latent space. These projected features are then used as input to a histogram-based gradient boosting regressor.

This hybrid approach combines nonlinear representation learning with the robustness and interpretability of tree-based models.

E. Optimization

Neural models are trained using the AdamW optimizer with early stopping based on validation RMSE. The training objective minimizes mean squared error in log space. Global random seeds are fixed across experiments to ensure reproducibility.

VII. Experimental Setup

A. Label Horizons

Experiments are conducted on fixed-horizon engagement targets measured at $H \in \{7, 30\}$ days after publication.

B. Data Splits

Two evaluation protocols are used.

Random Split. Videos are randomly partitioned into training (70%), validation (15%), and test (15%) sets using a fixed random seed. Duplicate video identifiers are removed prior to splitting.

Chronological Split. To approximate real-world deployment, videos are ordered by the `captured_at` timestamp. The earliest 70% of samples are used for training, the next 15% for validation, and the most recent 15% for testing. This protocol evaluates forward generalization under mild distribution shift.

C. Leakage Prevention

To avoid label leakage, post-horizon engagement statistics such as total view count, interaction counters, and derived rate metrics are excluded from feature construction. Only pre-horizon metadata and modality embeddings are used as predictors.

D. Evaluation Metrics

Model performance is evaluated using

- Root Mean Squared Error (RMSE) in log space
- Mean Absolute Error (MAE) in log space
- Coefficient of determination (R^2) in log space
- RMSE and MAE on inverse-transformed raw view counts

Slice metrics are additionally reported for text-present and text-missing subsets to evaluate robustness to missing modalities.

E. Fusion Strategy Comparison

To isolate representation effects, we compare concatenation, sum pooling, and max pooling fusion strategies under identical downstream training configurations.

VIII. Results

We evaluate model performance across multiple fusion strategies and prediction horizons using log-transformed engagement targets. Evaluation metrics include mean absolute error (MAE), root mean squared error (RMSE), and coefficient of determination (R^2) in log space.

Across all evaluated configurations, gradient boosted decision trees (GBDT) achieve the strongest predictive performance among individual model families, while stacked ensembles provide a modest additional improvement.

A. Model Family Performance

Table I summarizes average performance across all evaluated configurations.

GBDT achieves the lowest prediction error and highest explanatory power across the evaluation set. Neural architectures outperform the linear baseline but do not surpass the tree-based model. The strongest overall configuration uses concatenation fusion with the 7-day horizon target, achieving a test MAE of approximately 1.18 in log space.

TABLE I
Model performance on the virality prediction task using log-transformed engagement targets.

Model	Test MAE (log)	RMSE (log)	R^2 (log)
GBDT	1.18	1.50	0.56
Concat MLP	1.58	2.06	0.26
Gated Fusion MLP	1.68	2.14	0.21
Ridge	1.86	2.37	0.05

B. Ensemble Performance

In addition to individual model families, we evaluate a stacked ensemble combining the strongest base predictors. The ensemble aggregates predictions from the base regression models using a second-stage learner trained on validation outputs.

Predictions are produced in log-space using the transformed engagement target $\hat{y}_{\log} = \log(1+y)$. To interpret improvements in practical terms, predicted values are mapped back to raw view counts using the inverse transformation:

$$\hat{y}_{raw} = \begin{cases} 0, & \hat{y}_{\log} \leq -20 \\ e^{\hat{y}_{\log}} - 1, & -20 < \hat{y}_{\log} < 30 \\ e^{30} - 1, & \hat{y}_{\log} \geq 30 \end{cases}$$

Figure 3 compares the best-performing base model, deterministic fusion configuration, and the stacked ensemble across both prediction horizons.

Results show that stacking provides a modest improvement in predictive performance over the strongest individual model in log-space. However, when mapped back to the raw engagement scale, these improvements correspond to substantial differences in predicted view counts. For the 7-day horizon, the ensemble improves RMSE by approximately 42,780 views (1.62%). At the 30-day horizon, the improvement corresponds to roughly

130,973 views (1.98%). This highlights how small log-space error reductions translate into large practical gains when modeling heavy-tailed engagement distributions.

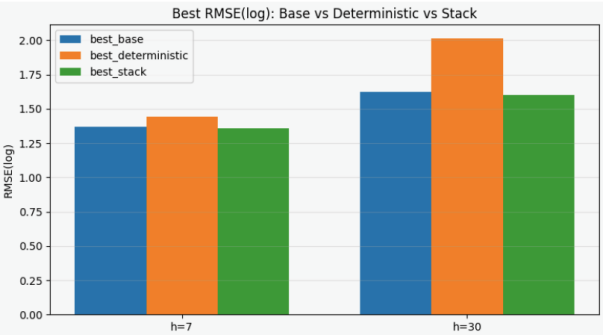

Fig. 3. Comparison of best RMSE(log) across modeling strategies.

C. Prediction Horizon Comparison

Prediction performance differs substantially between the 7-day and 30-day engagement horizons. Across all model families, the shorter horizon consistently yields lower prediction error and stronger model fit.

GBDT exhibits the smallest degradation between horizons, with MAE increasing by approximately 11% when moving from the 7-day to the 30-day target. Neural architectures experience substantially larger performance drops, with error increases exceeding 50%. Ridge regression shows the largest degradation, with error increasing by nearly 76%.

These results indicate that short-term engagement dynamics are substantially easier to model than longer-term outcomes, likely reflecting increased uncertainty in long-term content diffusion and platform recommendation behavior.

TABLE II
Prediction performance across engagement horizons.

Model	MAE (7-day)	MAE (30-day)	% Incr
GBDT	1.12	1.24	11%
Concat MLP	1.24	1.91	54%
Gated MLP	1.30	2.05	57%
Ridge	1.35	2.37	76%

D. Fusion Strategy Comparison

We compare three deterministic multimodal fusion strategies: concatenation, sum pooling, and max pooling. Across all strategies and prediction horizons, GBDT remains the best-performing model family.

For the 7-day horizon, concatenation fusion produces the lowest prediction error. Performance differences between fusion strategies are relatively small for tree-based models, suggesting that the GBDT learner adapts effectively to the high-dimensional feature spaces produced by different fusion operators.

Neural architectures exhibit greater sensitivity to fusion choice, particularly at the 30-day horizon where performance varies more widely across configurations. However, even under their strongest fusion settings, neural models do not outperform the tree-based approach.

E. Modality Robustness

To evaluate robustness to missing textual metadata, we compute performance slices for videos with and without transcript features. All models experience some degradation when text information is unavailable, but the magnitude of this effect varies across model families.

GBDT shows the strongest robustness, with minimal change in prediction error between text-present and text-missing subsets. Neural architectures exhibit substantially larger performance degradation, indicating greater reliance on textual embeddings.

TABLE III
Robustness to missing textual metadata.

Model	MAE (No Text)	MAE (Text Present)
GBDT	**1.195**	**1.178**
Concat MLP	1.993	1.552
Gated MLP	2.125	1.648
Ridge	2.091	1.848

IX. Representation Analysis and Latent Structure

Beyond predictive performance, we analyze the structure of the learned multimodal representation space to understand how videos are organized in the embedding domain. In particular, we investigate whether the fused representations capture latent attributes associated with content style and engagement behavior.

A. Embedding Space Structure

To analyze structure within the learned multimodal representation space, we perform unsupervised clustering over fused embeddings derived from visual, audio, and textual modalities.

Given embeddings $X \in \mathbb{R}^{N \times D}$, dimensionality is first reduced using principal component analysis (PCA). When $D > 50$, embeddings are projected onto the first 50 principal components in order to improve clustering stability while preserving the dominant variance directions.

Rather than fixing the number of clusters a priori, we evaluate candidate cluster counts across a predefined range and select the best configuration using clustering quality metrics. For each candidate K, K-Means clustering is fit across multiple random seeds and evaluated using a composite score that incorporates silhouette score, clustering stability across seeds, and penalties for small clusters. The selected configuration is then used to produce deterministic cluster assignments:

$$c_i = \text{KMeans}(X_i)$$

where $c_i \in \{1, \ldots, K\}$ denotes the cluster assignment for video i.

B. Cluster-Level Engagement Analysis

To evaluate whether clusters correspond to distinct engagement regimes, we compute cluster-level engagement statistics:

$$\mu_k = \mathbb{E}[\log(1 + \text{views}_H) \mid c_i = k]$$

$$\sigma_k^2 = \text{Var}[\log(1 + \text{views}_H) \mid c_i = k]$$

Variation in μ_k across clusters indicates that different regions of the embedding space correspond to distinct engagement behaviors. Clusters with higher mean engagement represent regions of the representation space associated with stronger audience response.

Fig. 4. PCA visualization of fused multimodal embeddings with cluster assignments. The projection reveals structured groupings in the representation space corresponding to distinct video style regimes.

These differences suggest that the multimodal embeddings capture latent attributes related to video structure, pacing, narrative style, and audiovisual composition that influence downstream engagement outcomes. As a result, clustering provides an interpretable mechanism for identifying distinct content regimes within the dataset.

C. Generative Interface and Performance-Guided Ranking

The learned representation space also provides a natural interface between predictive modeling and generative content creation.

Given a generative model capable of producing candidate short-form videos, generated samples can be embedded using the same multimodal pipeline:

$$\mathbf{f}_{gen} = \Phi(\text{generated video})$$

The trained predictor can then estimate expected engagement:

$$\hat{y}_{gen} = f(\mathbf{f}_{gen}, M_{gen})$$

This enables performance-guided ranking among generated candidates. Generated videos that fall into embedding regions associated with higher predicted engagement can be prioritized for selection.

In addition, cluster-level interpretations can be translated into structured prompts describing common characteristics of videos within each cluster. These prompts can be used with modern video generation systems to produce new candidate videos aligned with specific latent content attributes.

While a fully closed-loop optimization system is beyond the scope of this work, the framework establishes the representational and predictive components necessary for future reinforcement learning or adaptive generative pipelines.

X. Limitations

Several limitations remain.

First, although chronological evaluation is included, the system is not explicitly optimized for severe distribution shift arising from platform policy changes, evolving recommendation algorithms, or shifts in creator behavior.

Second, pretrained encoders are frozen rather than jointly fine-tuned. While this improves representation stability under rolling retraining, it may limit the model's ability to capture deeper cross-modal interactions.

Third, clustering relies on K-Means, which assumes approximately spherical cluster structure. More flexible clustering methods may capture richer structure in the embedding space.

Finally, the dataset is platform-specific and subject to API sampling bias, which may limit generalization across different social media ecosystems or content formats.

XI. Conclusion

We present a deterministic multimodal framework for fixed-horizon virality prediction in short-form video platforms. By combining temporally consistent horizon labeling, pretrained modality-specific encoders, and schema-locked fusion strategies, the system supports stable representation learning and reproducible model training for continuously updated datasets.

Across multiple model families, gradient boosted decision trees achieve the strongest predictive performance, demonstrating the effectiveness of tree-based learners for high-dimensional multimodal feature spaces and heavy-tailed engagement targets.

Beyond prediction, analysis of the learned representation space reveals coherent latent structure within short-form video content. Unsupervised clustering over fused embeddings identifies distinct engagement regimes, providing interpretable summaries of content styles.

Embedding generated videos into the same representation space enables performance-guided ranking of candidate content, establishing a bridge between multimodal representation learning, predictive modeling, and generative content exploration.

References

[1] H. Feichtenhofer, Y. Li, K. He, and C. Xie, "Masked Autoencoders as Spatiotemporal Learners," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[2] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[3] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

[4] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2016.

[5] T. Baltrušaitis, C. Ahuja, and L. Morency, "Multimodal Machine Learning: A Survey and Taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.