

A MultiModal Autoregressive Transformer Framework for Social Popularity Prediction

Mingzhu Wang¹, Dong Zhou^{1*}, Zeliang Zhou¹

¹School of Information Science and Technology, Guangdong University of Foreign Studies,
Guangzhou 510006, China
dongzhou @gdufs.edu.cn

Abstract

Understanding and predicting the temporal popularity of social media content is critical for digital marketing and recommendation systems. Existing methods often rely on static feature concatenation, struggling to capture complex multimodal interactions and strict sequential dependencies in long-term popularity evolution, frequently underestimating viral content. To address these limitations, we propose a MultiModal Autoregressive Transformer Framework (MMATF). Our framework extracts and refines key representations from video covers, texts, category tags, and numerical statistics. A dynamic modality interaction module adaptively fuses these inputs to capture evolving content dynamics, establishing a solid foundation for temporal relevance. Furthermore, an autoregressive temporal encoding module strictly prevents future information leakage. Extensive experiments on SMTPD and ICIP datasets demonstrate MMATF significantly outperforms advanced baselines across 7-, 14-, and 30-day horizons. Specifically, compared to the strongest baseline, our model reduces Mean Absolute Error by 3.8% and 4.3%, and improves Spearman's Rank Correlation by 0.3% and 0.4% for 30-day predictions on SMTPD and ICIP respectively, proving consistent long-term forecasting accuracy.

1 Introduction

Rapid social network growth has generated massive User Generated Content (UGC). Social information has been extensively exploited to improve personalized search and recommendation by capturing user interests and behavioral preferences (Zhou et al., 2016). Accurate Social Media Popularity Prediction (SMPP) remains vital for intelligent recommendation (Troudi et al., 2021), digital marketing, content optimization (Xu et al., 2026), and public opinion monitoring (Cheng et al., 2024). Early

research highlights the critical role of early popularity (Cao et al., 2020). Recently, multimodal frameworks (e.g., HyFea (Lai et al., 2020), MMF (Lin and Lee, 2024)) and advanced variants (Wang et al., 2023) have improved heterogeneous feature integration.

However, translating multimodal UGC spanning visual, textual, and metadata domains into stable predictions remains challenging (Xu et al., 2025a). Existing methods often rely on simple concatenation or static weighting, losing deep semantics and inducing cross-modal inconsistency (Hsu et al., 2024). Recent studies in other cross-modal retrieval scenarios have further demonstrated the potential of hierarchical alignment mechanisms for bridging heterogeneous semantic spaces (Wang et al., 2025a). While foundation models like BLIP (Li et al., 2022) and CLIP (Radford et al., 2021) improve vision-language understanding, orchestrating modality fusion for popularity forecasting requires deeper exploration (Xu et al., 2020; Hou and Li, 2025).

Crucially, capturing temporal dynamics behind popularity evolution remains a bottleneck. Popularity is a dynamic process driven by user behaviors, social trends, and diffusion mechanisms (Xi et al., 2025). Existing methods typically extract statistical spatio-temporal correlations rather than explicitly modeling sequential causality. Without autoregressive constraints, predictions risk being confounded by spurious correlations, yielding unstable forecasting (Jing et al., 2025). Multi-scale decomposition methods often struggle to capture intricate interactions without explicit constraints (Cheng et al., 2023). Consequently, models remain vulnerable to prediction drift (Lee et al., 2025; Lin and Lee, 2024). Furthermore, over-reliance on environment-specific features restricts generalization in cold-start or missing modality scenarios (Cheng et al., 2025a), while heavy fine-tuning limits interpretability (Cheng et al., 2025b).

*Corresponding author.

To address these limitations, we propose the MultiModal Autoregressive Transformer Framework (MMATF). Beyond semantic alignment, MMATF explicitly models temporal progression via strict autoregressive masking. Embedding an autoregressive mask into the temporal attention network restricts prediction paths to historical states, mitigating information leakage and spurious correlations. Coupled with an adaptive multimodal interaction mechanism, the framework bridges heterogeneous streams while preserving sequence boundaries, improving prediction accuracy and long-term stability.

Our core contributions are summarized as follows:

- We design an attention-driven extraction and dynamic interaction module that adaptively fuses visual, textual, categorical, and numerical features into a noise-reduced joint embedding space.
- We develop an autoregressive temporal encoding module that embeds a unidirectional mask into self-attention layers, preventing information leakage and shielding predictions from spurious correlations.
- We integrate extraction, fusion, and strict sequential modeling into a unified architecture. Extensive evaluations on the SMTPD and ICIP datasets confirm that MMATF significantly outperforms state-of-the-art baselines across three standard metrics over multiple horizons.

2 Related Work

Early popularity prediction research relied on statistical models and social network topologies to model propagation dynamics (Rizoiu et al., 2017) or extract structural features (Wu et al., 2023). While previous works confirmed the critical role of early popularity (Szabó and Huberman, 2010) and clustered temporal evolutionary patterns (Cheung and Lam, 2022), these methods depend heavily on early statistical features, struggling with the rich multimodal content and long-term dynamics inherent in modern social media. Exploring macro and micro dynamic features remains a persistent focus (Wang et al., 2025b).

Modern social media’s multimodal nature prompts integrating visual, textual, and metadata features using pre-trained models like ALBEF (Xie et al., 2025; Zhou et al., 2023), ResNet, and BERT (Beniwal and Saraswat, 2024). However,

early methods typically employed simple concatenation or static weighting (Xie et al., 2023; Kang et al., 2019), limiting their effectiveness (Chen and Guestrin, 2016). Later research introduced self-attention (Lin et al., 2022) to adaptively allocate weights based on semantics. Nevertheless, maintaining computational efficiency and strict temporal logic under large-scale inputs remains challenging (Wu et al., 2022). Constrained by early fusion, many models focus on single-point predictions rather than refined sequential evolution modeling.

Forecasting long-term dynamics requires handling distant historical dependencies and maintaining autoregressive masking. Early datasets limited to specific domains (Mazloom et al., 2016; Sanjo and Katsurai, 2017) or lacking temporal alignment (Wu et al., 2019) only supported static point prediction. The emergence of large-scale, temporally aligned multimodal benchmarks like SMTPD enables accurate long-term popularity prediction, imposing stricter temporal modeling requirements.

Despite advancements in sequence modeling via temporal convolutional networks (Wan et al., 2019), LSTM architectures (Sangwan and Bhatnagar, 2025), and Temporal Fusion Transformers (Bi et al., 2026), existing approaches exhibit critical shortcomings. They often rely on static fusion, failing to capture dynamic feature interactions over time. Furthermore, standard bidirectional architectures (Larkin and Hystad, 2022) and standard Transformers lack rigorous temporal constraints. In noisy social media environments, they often implicitly utilize future information or fit spurious spatio-temporal correlations, leading to severe prediction drift amidst volatile content trends.

Unlike these approaches, our proposed MMATF introduces strict autoregressive masking into the temporal modeling process. By fundamentally restricting sequence dependencies, our framework successfully resolves the vulnerabilities of future data leakage and spurious correlations.

3 Methodology

3.1 Problem Definition

Let the input be defined as a multimodal content set $X = \{X_v, X_t, X_c, X_n\}$, where the four distinct modalities are formally defined as follows. The visual features X_v are extracted solely from the video cover image. The textual features X_t are derived from the concatenation of the video title,

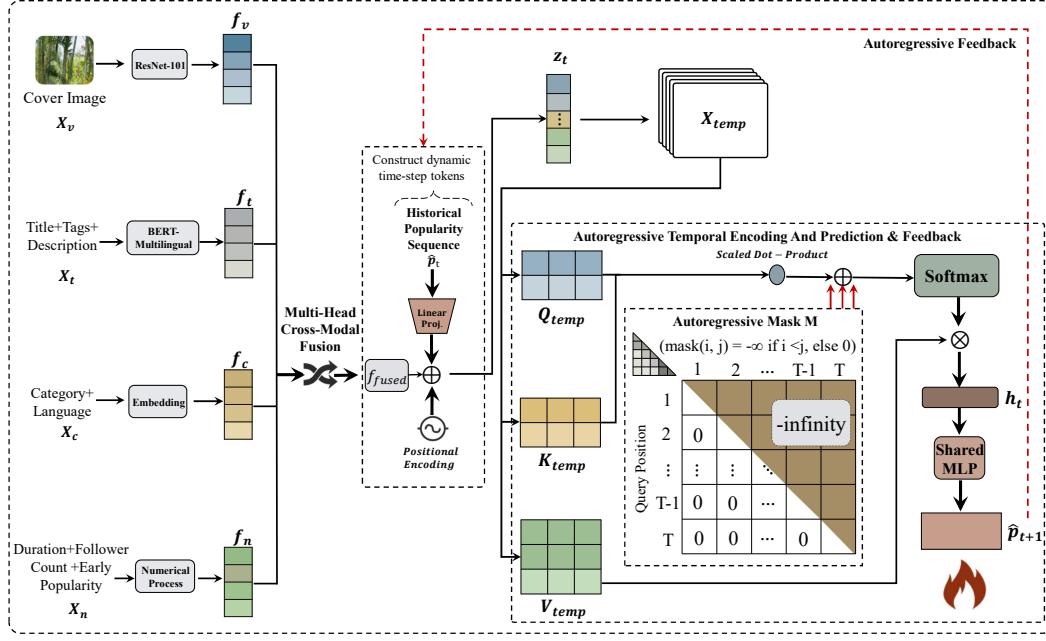


Figure 1: The overall architecture of our proposed MMATF.

description, and tags. The categorical features X_c comprise the discrete category and language classifications of the content. The numerical features X_n are continuous variables representing the video duration, follower count, and early popularity metrics (extracted strictly from the first 24 hours after upload, before the prediction window starts).

The primary objective is to output a T -day future popularity prediction sequence, denoted as $\hat{P} = \{\hat{p}_t \mid t = 2, \dots, T + 1\}$. Let $P = \{p_t \mid t = 1, \dots, T\}$ represent the corresponding ground truth sequence. The model aims to minimize the discrepancy between P and $\hat{P}$ while strictly preserving autoregressive masking, ensuring that the prediction $\hat{p}_t$ at any given time step t depends exclusively on historical and present information up to time t .

3.2 System Architecture

The overall architecture of the proposed MMATF is illustrated in Figure 1. It is designed to map raw, multi-source heterogeneous inputs into a final chronological popularity sequence $\hat{P} \in \mathbb{R}^T$ through a complete non-linear transformation. The framework logically progresses through three primary stages.

- **Feature Extraction** independently learns representations and extracts key features for each specific modality through dedicated neural pathways.

- **Multimodal Fusion** takes these extracted modality feature vectors and feeds them into a Multi-Head Attention module. This step facilitates deep, dynamic feature interaction and aggregation across the heterogeneous inputs, producing a single, static fused vector $f_{fused} \in \mathbb{R}^{d_{model}}$.
- **Temporal Modeling** bridges the static multimodal context with dynamic time-series forecasting. At each time step t , the aggregated multimodal representation f_{fused} is concatenated with the corresponding historical popularity state. To explicitly provide the model with chronological order awareness, distinct Positional Encodings (PE) are added to this combined feature, resulting in the step-specific joint token z_t . These sequential tokens z_t are then stacked across T steps to construct the final temporal sequence matrix X_{temp} . Finally, X_{temp} is directly fed as input into an **Autoregressive Transformer Encoder**, which employs autoregressive masking within its self-attention layers to enforce strict sequential dependencies and prevent future information leakage.

The target prediction sequence over T days is generated through a weight-shared output MLP. Throughout the entire architecture, residual connections and Layer Normalization are extensively deployed within the core computational blocks to

ensure gradient-friendly training and maintain stable information flow across the deep network.

3.3 Core Components

3.3.1 Multimodal Key Feature Extraction Module

This module aims to extract the core representations of the four modalities, reducing the dimensionality of high-dimensional heterogeneous inputs and uniformly mapping them into compact, dense feature vectors $f_v, f_t, f_c, f_n \in \mathbb{R}^{d_{model}}$ to facilitate subsequent fusion.

- **Visual Path.** A pre-trained ResNet-101 processes the video cover images to extract high-level visual semantics, which are subsequently projected to a unified dimension through a MLP:

$$f_v = \text{MLP}(\text{ResNet}(X_v)) \quad (1)$$

- **Textual Path.** We employ a pre-trained BERT-Multilingual model to encode the textual metadata. Specifically, the [CLS] token representation is extracted and projected through a MLP to capture the global semantic context:

$$f_t = \text{MLP}(\text{BERT}(X_t)_{[\text{CLS}]}) \quad (2)$$

- **Categorical Path.** Discrete category and language features are mapped into dense continuous spaces via distinct Embedding layers. They are subsequently processed through separate parallel MLPs (MLP_{cat} and MLP_{lang}) and integrated via element-wise multiplication. This multiplication serves as an explicit feature-crossing operation, allowing the model to capture the non-linear semantic interactions between the content category and the language context. Finally, the crossed features are passed through a final MLP:

$$f_c = \text{MLP}_f(\text{MLP}_{cat}(\text{Embed}_{cat}(X_c)) \odot \text{MLP}_{lang}(\text{Embed}_{lang}(X_c))) \quad (3)$$

- **Numerical Path.** Continuous numerical features undergo a numerical processing step (e.g., logarithmic transformation and standardization) before being mapped to the target dimension through a vector MLP:

$$f_n = \text{MLP}(\text{Process}(X_n)) \quad (4)$$

3.3.2 Dynamic Modality Interaction Module

To capture complex cross-modal correlations as explicitly depicted in our architecture, the four distilled modality features ($f_v, f_t, f_c, f_n \in \mathbb{R}^{d_{model}}$) are routed into the **Multi-Head Cross-Modal Fusion** block. Specifically, these vectors are stacked into a modality sequence matrix $X_{mod} \in \mathbb{R}^{S \times d_{model}}$ (where sequence length $S = 4$). Through linear projections to yield the Query (Q_{mod}), Key (K_{mod}), and Value (V_{mod}) matrices, the fusion process employs a Multi-Head Attention mechanism (Lin et al., 2022), mathematically formulated as:

$$\text{Attention}(Q_{mod}, K_{mod}, V_{mod}) = \text{softmax}\left(\frac{Q_{mod}K_{mod}^T}{\sqrt{d_k}}\right)V_{mod} \quad (5)$$

Following the attention computation, average pooling aggregates the output across the sequence dimension, compressing it into the unified dense vector f_{fused} . As illustrated in the framework, this fused representation f_{fused} ensures dimensional compatibility and serves as the foundational base for the subsequent dynamic time-step token construction, where it is integrated with the historical popularity sequence $\hat{p}_t$ and positional encoding.

3.3.3 Mechanism of Autoregressive Temporal Encoding Module

To capture the chronological progression of video popularity over the T -day horizon, this module introduces an autoregressive temporal encoding mechanism to explicitly model the unidirectional flow of time.

At each step t , the static representation f_{fused} is concatenated with the projected previous popularity state $\hat{p}_{t-1}$ (using ground truth p_{t-1} during training, and autoregressive predictions during inference) and augmented with positional encodings to construct sequential tokens $X_{temp} \in \mathbb{R}^{T \times d_{model}}$.

To strictly prevent future information leakage within the Transformer self-attention layers, an autoregressive mask $M \in \mathbb{R}^{T \times T}$ is introduced, defined as:

$$M_{i,j} = \begin{cases} 0, & \text{if } j \leq i \\ -\infty, & \text{if } j > i \end{cases} \quad (6)$$

Linear projections of X_{temp} yield the temporal Query (Q_{temp}), Key (K_{temp}), and Value (V_{temp})

matrices. The masked attention calculation is formulated as:

$$\begin{aligned} & \text{Attention}_{\text{autoreg}}(Q_{\text{temp}}, K_{\text{temp}}, V_{\text{temp}}) \\ &= \text{Softmax} \left(\frac{Q_{\text{temp}} K_{\text{temp}}^T}{\sqrt{d_k}} + M \right) V_{\text{temp}} \end{aligned} \quad (7)$$

By adding the mask matrix M , the upper triangular region ($j > i$) is penalized to $-\infty$, ensuring that the hidden state h_t at step t dynamically aggregates historical contexts up to time t without accessing future sequences.

3.3.4 Prediction Output and Loss Function

A weight-shared output MLP maps temporally encoded hidden states h_t to popularity predictions $\hat{p}_t$. Sharing weights across all T sequence steps reduces the parameter space, mitigates overfitting, and ensures consistent gradient propagation.

To address the non-differentiable $\arg \max$, we use a soft-argmax approximation ($\tau = 0.1$) during back-propagation. The overall loss $\mathcal{L}$ integrates Smooth L1 Loss (SL) on predictions and their first (Δ) and second-order (Δ^2) temporal derivatives, a soft-argmax peak alignment term, and total variation (TV) regularization (LR):

$$\begin{aligned} \mathcal{L} = & \text{SL}(\hat{P}, P) + \lambda_1 \text{SL}(\Delta \hat{P}, \Delta P) \\ & + \lambda_2 \text{SL}(\Delta^2 \hat{P}, \Delta^2 P) + \epsilon LR \\ & + \alpha |\text{softargmax}(\hat{P}) - \text{softargmax}(P)| \end{aligned} \quad (8)$$

where $\lambda_1, \lambda_2, \alpha, \epsilon > 0$ are cosine-annealed scalars, and the TV regularization term LR is defined as:

$$LR = \sum_{t=1}^T |\Delta \hat{p}_{t+1}| + \sum_{t=1}^T |\Delta^2 \hat{p}_{t+1}| \quad (9)$$

Here, $\hat{P}$ and P are the predicted and ground truth sequences. The finite differences $\Delta \hat{p}_{t+1}$ and $\Delta^2 \hat{p}_{t+1}$ capture trend dynamics, while the soft-argmax term minimizes peak position discrepancies.

4 Experiments and Results

4.1 Experimental Setup

We implement MMATF using PyTorch on an NVIDIA GPU. Experimental evaluations utilize SMTPD (Xu et al., 2025c) and ICIP (Xu et al., 2025b) datasets. SMTPD comprises 282K samples across 15 categories, covering over 90 languages

and 30 days of daily popularity data. ICIP contains 21K images with associated textual metadata, category labels, and 30-day engagement metrics. Both datasets are partitioned using 5-fold cross-validation.

Models are trained for 40 epochs using the AdamW optimizer with a batch size of 64. Pre-trained ResNet-101 and BERT-Multilingual backbones remain frozen. Only projection layers and the Transformer are trained with an initial learning rate of 1×10^{-2} , dynamically adjusted via PyTorch's ReduceLROnPlateau scheduler. Additional architecture tuning results for the number of attention heads and Transformer layers are reported in Appendix C.

Training and inference strategies adopt distinct input constructions balancing efficiency with strict real-world constraints. During training, the full ground truth sequence $P = \{p_1, \dots, p_T\}$ is fed simultaneously. Utilizing a $T \times T$ autoregressive mask enables parallel processing while preventing future information leakage. Conversely, inference operates in a step-by-step autoregressive mode (Algorithm 1). The sequence is dynamically generated by iteratively feeding each predicted state $\hat{p}_{t+1}$ as the subsequent input token, accurately reflecting real-world forecasting constraints.

To ensure fair evaluation, we report Mean Absolute Error (MAE), symmetric Mean Absolute Percentage Error (sMAPE), and Spearman's Rank Correlation (SRC) across 7-, 14-, and 30-day horizons. All results are averaged over 5-fold cross-validation.

MAE measures average prediction error magnitude ($\downarrow$ lower is better):

$$MAE_d = \frac{1}{n} \sum_{i=1}^n |\hat{P}_{d,i} - P_{d,i}| \quad (10)$$

sMAPE calculates symmetric percentage errors to mitigate growth asymmetry ($\downarrow$ lower is better):

$$\text{sMAPE}_d = \frac{100}{n} \sum_{i=1}^n \frac{|\hat{P}_{d,i} - P_{d,i}|}{|\hat{P}_{d,i}| + |P_{d,i}| + \epsilon} \quad (11)$$

SRC assesses ranking consistency between predicted and actual trends ($\uparrow$ higher is better), where r denotes the rank:

$$\text{SRC}_d = \frac{\sum_{i=1}^n x_i y_i}{\sqrt{\sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i^2}} \quad (12)$$

where $x_i = r_{\hat{P}_{d,i}} - \bar{r}_{\hat{P}_d}$ and $y_i = r_{P_{d,i}} - \bar{r}_{P_d}$ denote the mean-centered ranks of the predicted and actual values respectively.

Algorithm 1 MMATF Autoregressive Inference Process

```

1: Input: Multimodal set  $X = \{X_v, X_t, X_c, X_n\}$ , Sequence length  $T$ , Initial historical popularity  $p_1$ 
2: Output: Predicted popularity sequence  $\hat{P} = \{\hat{p}_2, \dots, \hat{p}_{T+1}\}$ 
3: Extract modality features  $f_v, f_t, f_c, f_n \in \mathbb{R}^{d_{model}}$  via respective networks
4:  $f_{fused} \leftarrow \text{MultiHeadAttention}(f_v, f_t, f_c, f_n)$ 
5: Initialize target sequence state  $\hat{p}_1 \leftarrow p_1$ 
6: for time step  $t = 1$  to  $T$  do
7:    $z_t \leftarrow \text{LinearProj}(\text{Concat}(f_{fused}, \text{LinearProj}(\hat{p}_t))) + \text{PE}$ 
8:   Form current temporal sequence matrix  $X_{temp} \leftarrow [z_1, \dots, z_t]$ 
9:   Construct autoregressive mask  $M \in \mathbb{R}^{t \times t}$  where  $M_{i,j} = -\infty$  if  $j > i$  else 0
10:   $Q_{temp}, K_{temp}, V_{temp} \leftarrow X_{temp}$ 
11:   $h_t \leftarrow \text{Softmax}\left(\frac{Q_{temp}K_{temp}^T}{\sqrt{d_k}} + M\right)V_{temp}$ 
12:   $\hat{p}_{t+1} \leftarrow \text{SharedMLP}(h_t)$ 
13: end for
14: return  $\hat{P} = \{\hat{p}_2, \dots, \hat{p}_{T+1}\}$ 

```

4.2 Baseline Methods

To comprehensively assess our model’s performance, we incorporate several baseline methods for comparison. All baselines were re-implemented and hyper-parameter tuned under identical training settings (5-fold cross-validation, same optimizer and scheduler) to ensure a fair comparison.

- **MLP** (Xu et al., 2025c): A stacked fully connected network that captures static feature correlations but fails to model temporal sequential dependencies.
- **LSTM** (Xu et al., 2025c): A recurrent network utilizing gating mechanisms to capture long- and short-term temporal dependencies in popularity sequences.
- **MTD** (Wu et al., 2016): Decouples popularity into user-item and time-sensitive contexts via multi-scale dual temporal features, optimizing via matrix norm error minimization.
- **HMMVED** (Xie et al., 2023): A hierarchical framework that learns a coarse-grained popularity prior and aligns it with a content posterior via a multimodal variational information bottleneck.
- **MASSL** (Zhang et al., 2022): A multimodal VAE regression model that employs a Product of

Experts mechanism for cross-modal fusion and missing modality adaptation.

- **CBAN** (Cheung and Lam, 2022): Leverages a cross-modal bipolar attention mechanism to capture bidirectional semantic relationships between ViT and BERT features.
- **HyFea** (Lai et al., 2020): Extracts a diverse feature set (visual, textual, spatio-temporal, and user profiles) and utilizes CatBoost for popularity prediction.
- **TMALL** (Chen et al., 2016): A transductive learning approach that unifies social, visual, acoustic, and textual modalities into a shared optimal latent common space.

4.3 Performance Comparison

In this section, we compare our model with various baseline methods. To ensure the reliability of the results, all models are evaluated using 5-fold cross-validation. The average performance scores are presented in Table 1, and the evaluation metrics comprise MAE, SRC, and sMAPE.

- Compared to the best baseline, TMALL, our framework delivers consistent performance gains across all prediction windows. On the SMTPD dataset for short-term 7-day predictions, MMATF reduces the Mean Absolute Error by 3.6% and improves Spearman’s Rank Correlation by 0.2%. Notably, sMAPE fell by 4.7 percentage points. When the horizon extends to 14 days, the model achieves a 5.2% MAE reduction, a 0.2% SRC enhancement, and a 5.4 percentage point drop in sMAPE. For the 30-day forecasting, the framework maintains a robust 3.8% MAE reduction, a 0.3% SRC improvement, and a 5.5 percentage point lower sMAPE. The ICIP dataset reflects similar performance trends. At the 30-day mark, our model secures a 4.3% MAE decrease, a 0.4% SRC increase, and a 4.0 percentage point sMAPE reduction over TMALL.
- We observe an interesting dynamic among the baseline performances. Recent models generally show chronological improvements, advancing from HyFea to HMMVED and then to MASSL. However, the classic TMALL model demonstrates robust performance across all three metrics, surpassing even recent transformer-based architectures like CBAN. This result suggests that

Datasets	Method	Day 7	Day 14	Day 30
SMTPD	MLP	0.875 / 0.949 / 28.5%	0.997 / 0.940 / 33.2%	1.137 / 0.925 / 38.5%
	LSTM	0.681 / 0.969 / 22.4%	0.735 / 0.959 / 27.5%	0.782 / 0.958 / 32.4%
	HyFea	0.603 / 0.968 / 9.8%	0.682 / 0.957 / 12.1%	0.801 / 0.947 / 14.6%
	HMMVED	0.582 / 0.972 / 21.5%	0.661 / 0.963 / 24.8%	0.785 / 0.952 / 27.8%
	MTD	0.528 / 0.982 / 19.5%	0.605 / 0.973 / 23.2%	0.775 / 0.960 / 26.5%
	CBAN	0.525 / 0.982 / 18.8%	0.602 / 0.973 / 22.5%	0.771 / 0.960 / 25.8%
	MASSL	0.522 / 0.983 / 23.4%	0.599 / 0.974 / 26.8%	0.768 / 0.961 / 29.6%
	TMALL	0.499 / 0.985 / 13.2%	0.582 / 0.976 / 15.6%	0.758 / 0.962 / 18.0%
	MMATF	0.481 / 0.987 / 8.5%	0.552 / 0.978 / 10.2%	0.729 / 0.965 / 12.5%
ICIP	MLP	1.550 / 0.925 / 45.2%	1.620 / 0.910 / 50.1%	1.710 / 0.885 / 55.4%
	LSTM	1.408 / 0.931 / 38.5%	1.433 / 0.917 / 43.2%	1.498 / 0.897 / 48.6%
	HyFea	0.815 / 0.931 / 22.4%	0.875 / 0.908 / 25.6%	0.945 / 0.894 / 28.4%
	HMMVED	0.785 / 0.935 / 21.5%	0.850 / 0.912 / 24.2%	0.920 / 0.900 / 27.5%
	MTD	0.715 / 0.939 / 20.1%	0.780 / 0.918 / 22.8%	0.850 / 0.908 / 25.6%
	CBAN	0.710 / 0.940 / 19.5%	0.768 / 0.919 / 22.1%	0.838 / 0.909 / 24.8%
	MASSL	0.705 / 0.942 / 18.8%	0.760 / 0.921 / 21.5%	0.830 / 0.911 / 23.9%
	TMALL	0.675 / 0.950 / 17.5%	0.725 / 0.928 / 20.2%	0.782 / 0.920 / 22.5%
	MMATF	0.649 / 0.953 / 14.2%	0.698 / 0.931 / 16.8%	0.748 / 0.924 / 18.5%

Table 1: Performance comparison of eight models across SMTPD and ICIP datasets.

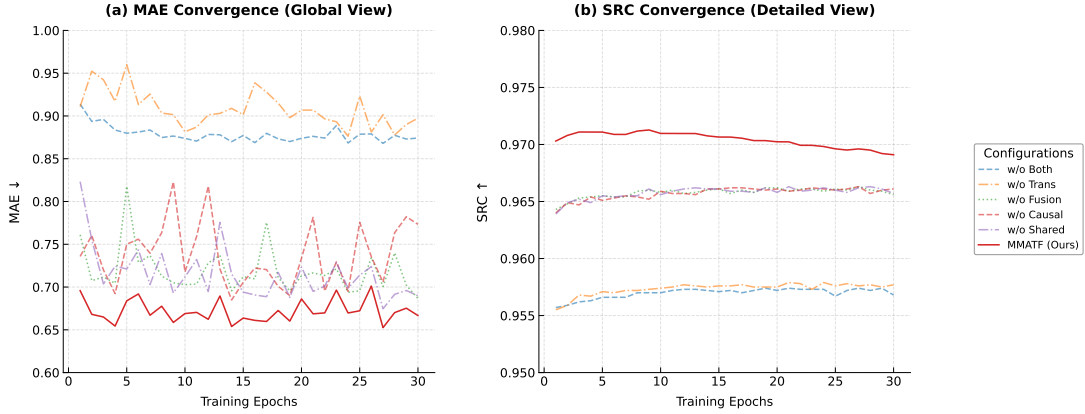


Figure 2: Learning curves of all configurations over training epochs in the ablation study.

TMALL’s latent common space design remains highly effective for multimodal fusion tasks. Furthermore, extending the prediction timeframe from 7 days to 30 days introduces cumulative uncertainty and temporal information loss, which impacts the performance of all models. Despite this natural degradation, MMATF exhibits strong long-term robustness. By utilizing autoregressive masking to properly model multimodal sequential dependencies, our framework mitigates information loss and sustains a clear metric advantage over all baselines in extended prediction scenarios.

4.4 Ablation Study

To further analyze the contribution of each component in MMATF, we conducted an ablation study on the SMTPD dataset under identical training configurations. Each variant was constructed by removing or replacing one specific component while keeping the remaining architecture unchanged. The

quantitative comparison results are presented in Table 2, and the corresponding training processes are visualized in Figure 2.

As shown in Figure 2, the complete MMATF model maintains a relatively stable optimization process throughout the training epochs, while the ablated variants exhibit different degrees of fluctuation during convergence. The MAE and SRC curves provide a detailed view of how different architectural designs affect the learning dynamics. Furthermore, the results in Table 2 demonstrate the performance changes caused by individual module removal, allowing us to analyze the specific contribution of each component to long-horizon popularity prediction.

- We first examined the influence of the dynamic modality interaction module by removing the multi-head attention based cross-modal interaction mechanism. The resulting variant suffers from the largest performance degradation, with

the MAE increasing by 6.45% and the SRC decreasing by 0.52%. This result highlights that directly combining visual, textual, categorical, and numerical representations is insufficient to fully capture the relationships among heterogeneous modalities. The dynamic interaction module provides an adaptive fusion process that helps the model extract complementary information from different sources.

- We further investigated the effectiveness of the autoregressive Transformer encoder by replacing it with a traditional LSTM-based temporal encoder. This modification increases the MAE by 4.25% and decreases the SRC by 0.31%. Compared with recurrent temporal modeling, the self-attention mechanism enables the model to establish dependencies among different time steps more flexibly, which is particularly important for predicting long-term popularity evolution.
- To evaluate the role of directional temporal modeling, we removed the autoregressive mask from the Transformer encoder. Without the temporal constraint, the MAE increases by 3.16%. This performance decline verifies that restricting attention to previous timestamps is necessary for maintaining the autoregressive prediction process and avoiding the influence of unavailable future information during forecasting.
- Finally, we analyzed the effect of the weight-sharing mechanism by removing the shared prediction head. The corresponding variant obtains a 1.78% increase in MAE. Although the performance change is relatively smaller than other components, the result shows that sharing parameters across different forecasting steps provides a unified prediction space and contributes to more consistent long-term trajectory modeling.

The ablation results on the SMTPD dataset demonstrate that the dynamic modality interaction and autoregressive temporal encoding modules have the largest impact on prediction performance, with MAE increases of 6.45% and 4.25% after removing them, respectively. These results verify the effectiveness of the proposed architectural components and their roles in improving long-horizon popularity prediction.

While the above experiments focus on the contribution of different model components, we further investigate the contribution of individual input modalities on the ICIP dataset. Specifically,

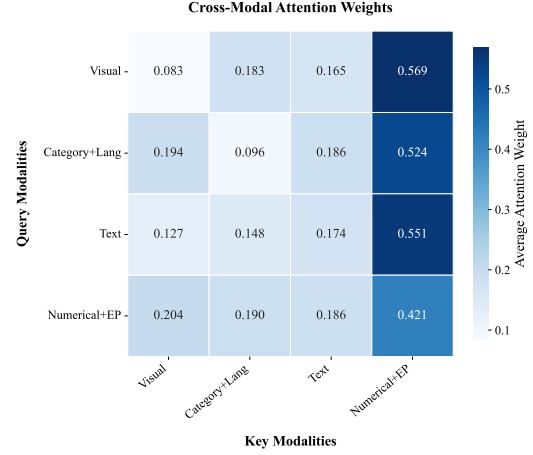


Figure 3: Visualization of cross-modal attention weights.

Configuration	MAE	SRC	Δ MAE	Δ SRC
w/o Dynamic Inter.	0.776	0.960	+6.45	-0.52
w/o Autoregressive Enc.	0.764	0.962	+4.25	-0.31
w/o Autoregressive Mask.	0.752	0.963	+3.16	-0.21
w/o Weight Sharing	0.742	0.964	+1.78	-0.10
Full Framework	0.729	0.965	—	—

Table 2: Ablation study of different components.

modality-level ablations are conducted under 7-day, 14-day, and 30-day forecasting settings by removing each modality group while keeping the remaining inputs and model components unchanged. The quantitative results are presented in Table 3.

As shown in Table 3, removing any modality group increases MAE and decreases SRC across all three forecasting horizons, showing that each modality provides useful information. Removing Numerical+EP causes the largest performance drop in both metrics. Among the semantic modalities, removing visual features has the greatest effect, followed by text and category-language information. The performance loss from removing semantic modalities also tends to increase with the forecasting horizon, especially for visual features. These results show that different modalities provide complementary information, while numerical statistics and early-popularity observations play the most important role across the three forecasting horizons.

4.5 Modality Interaction Visualization

To analyze the feature interaction behavior of MMATF, we visualize the average cross-modal attention weights learned by the dynamic interaction module under the selected configuration.

As presented in Figure 3, the Numerical and Early Popularity features receive relatively larger

Configuration	Metric	Day 7	Day 14	Day 30
w/o Text	MAE	0.665 (+2.47%)	0.720 (+3.15%)	0.776 (+3.74%)
	SRC	0.945 (-0.84%)	0.923 (-0.86%)	0.916 (-0.87%)
w/o Visual	MAE	0.674 (+3.85%)	0.732 (+4.87%)	0.792 (+5.88%)
	SRC	0.942 (-1.15%)	0.918 (-1.40%)	0.909 (-1.62%)
w/o Category-Language	MAE	0.659 (+1.54%)	0.711 (+1.86%)	0.764 (+2.14%)
	SRC	0.949 (-0.42%)	0.926 (-0.54%)	0.918 (-0.65%)
w/o Numerical+EP	MAE	2.077 (+220.03%)	2.024 (+189.97%)	1.910 (+155.35%)
	SRC	0.496 (-47.95%)	0.503 (-45.97%)	0.527 (-42.97%)
Full Framework	MAE	0.649 (-)	0.698 (-)	0.748 (-)
	SRC	0.953 (-)	0.931 (-)	0.924 (-)

Table 3: Modality-level ablation on ICIP. Parentheses show relative changes from the Full Framework; EP denotes early popularity.

attention from the Visual, Text, and Category-Language modalities. Specifically, the attention weights from these three modalities to the Numerical and Early Popularity features are 0.569, 0.551, and 0.524, respectively. This result shows that historical popularity information provides important references when combining different types of content features. Meanwhile, the learned attention distributions reflect how MMATF dynamically adjusts the contribution of different modalities during multimodal fusion.

5 Conclusions

This paper proposes a Multimodal Autoregressive Transformer Framework (MMATF) for long-term social media popularity prediction. By adaptively fusing heterogeneous features via dynamic interaction, our framework overcomes the severe underestimation of viral content common in existing methods. Moreover, it employs an autoregressive temporal encoding module to strictly maintain sequential dependencies, preventing future data leakage while enabling precise modeling of fine-grained popularity evolution. Extensive experiments on SMTPD and ICIP datasets demonstrate MMATF’s superior training stability, faster convergence, and sustained forecasting accuracy. Future work includes integrating large language models for low-resource languages, designing lightweight variants to reduce computational overhead, and applying causal discovery to enhance interpretability.

Limitations

Despite its strong predictive accuracy, MMATF exhibits inherent execution characteristics that warrant further systematic refinement. The primary architectural constraint stems from the sequential nature of the autoregressive sequence construction during the inference stage, where predicting long-term popularity profiles requires step-by-step

chronological generation. While this causal design guarantees that predictions remain strictly immune to future data leakage and spurious spatio-temporal correlations, the iterative dependency inherently introduces a progressive computational latency compared to parallel non-autoregressive baselines as the forecasting horizon scales from 7 days to 30 days in Table 1.

Nevertheless, our empirical evaluations demonstrate that this computational investment is strongly justified, balancing tracking speed with long-term stability. As contextualized by the trade-offs in our architecture tuning benchmarks in Figure 3, this sequential constraint suppresses standard prediction drift to capture fine-grained viral trajectories where traditional parallel models suffer massive underestimation. The ablation results in Table 2 further confirm that eliminating this causal constraint degrades performance, proving that the localized latency is a necessary exchange for preserving text-popularity integrity.

To mitigate this latency obstacle in production environments, our future development focuses on establishing a hybrid lightweight architecture that employs truncated sequence history or selective multi-step block jumping. This strategy aims to unlock a substantial inference acceleration in cross-lingual streaming scenarios with an accuracy concession of less than a marginal threshold. Additionally, integrating explicit latent imputation paths will further fortify framework robustness against completely missing input modalities without overriding the frozen parameter efficiency of the underlying pre-trained backbones.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62376062), and the Ministry of Education of Humanities and Social Science Project (No. 23YJAZH220).

References

- Rohit Beniwal and Pavi Saraswat. 2024. [A hybrid BERT-CNN approach for depression detection on social media using multimodal data](#). *The Computer Journal*, 67(7):2453–2472.
- Jiayu Bi, Linjun Lu, and Poly Z. H. Sun. 2026. [MMSTG-Net: Multi-scale spatiotemporal graph network with cross-modal contrastive alignment for multimodal emotion recognition](#). *IEEE Journal of Biomedical and Health Informatics*.
- Qi Cao, Huawei Shen, Jinhua Gao, Bingzheng Wei, and Xueqi Cheng. 2020. [Popularity prediction on social platforms with coupled graph neural networks](#). In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pages 70–78.
- Jingyuan Chen, Xuemeng Song, Liqiang Nie, Xiang Wang, Hanwang Zhang, and Tat-Seng Chua. 2016. [Micro tells macro: Predicting the popularity of micro-videos via a transductive model](#). In *Proceedings of the 24th ACM International Conference on Multimedia*, pages 898–907.
- Tianqi Chen and Carlos Guestrin. 2016. [XGBoost: A scalable tree boosting system](#). In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794.
- Zhangtao Cheng, Jian Lang, Ting Zhong, and Fan Zhou. 2025a. [Seeing the unseen in micro-video popularity prediction: Self-correlation retrieval for missing modality generation](#). In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 142–152.
- Zhangtao Cheng, Jiao Li, Jian Lang, Ting Zhong, and Fan Zhou. 2025b. [In-context prompt-augmented micro-video popularity prediction](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 11527–11535.
- Zhangtao Cheng, Wenxue Ye, Leyuan Liu, Wenxin Tai, and Fan Zhou. 2023. [Enhancing information diffusion prediction with self-supervised disentangled user and cascade representations](#). In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 3808–3812.
- Zhangtao Cheng, Jienan Zhang, Xovee Xu, Goce Trajcevski, Ting Zhong, and Fan Zhou. 2024. [Retrieval-augmented hypergraph for multimodal social media popularity prediction](#). In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 445–455.
- Tsun Hin Cheung and Kin Man Lam. 2022. [Cross-modal bipolar attention for multimodal classification on social media](#). *Neurocomputing*, 514:1–12.
- Wenzheng Hou and Weixin Li. 2025. [Modality-aligned hierarchical attention network for multi-modal popularity prediction on social media](#). In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 14073–14078.
- Chih-Chung Hsu, Chia-Ming Lee, Yu-Fan Lin, Yi-Shiuan Chou, Chih-Yu Jian, and Chi-Han Tsai. 2024. [Revisiting vision-language features adaptation and inconsistency for social media popularity prediction](#). In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11464–11469.
- Xin Jing, Yichen Jing, Yuhuan Lu, Bangchao Deng, Xueqin Chen, and Dingqi Yang. 2025. [CasFT: Future trend modeling for information popularity prediction with dynamic cues-driven diffusion models](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 11906–11914.
- Peipei Kang, Zehang Lin, Shaohua Teng, Guipeng Zhang, Lingni Guo, and Wei Zhang. 2019. [CatBoost-based framework with additional user information for social media popularity prediction](#). In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2677–2681.
- Xin Lai, Yihong Zhang, and Wei Zhang. 2020. [HyFea: Winning solution to social media popularity prediction for multimedia grand challenge 2020](#). In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4565–4569.
- Andrew Larkin and Perry Hystad. 2022. [Integrating geospatial data and social media in bidirectional long-short term memory models to capture human nature interactions](#). *The Computer Journal*, 65(3):667–678.
- Chia-Ming Lee, Bo-Cheng Qiu, Cheng-Jun Kang, Yi-Hsuan Wu, Jun-Lin Chen, Yu-Fan Lin, Yi-Shiuan Chou, and Chih-Chung Hsu. 2025. [Anchoring trends: Mitigating social media popularity prediction drift via feature clustering and expansion](#). In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 14094–14100.
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. [BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation](#). In *Proceedings of the International Conference on Machine Learning*, pages 12888–12900.
- Hung-Hsiang Lin, Jiun-Da Lin, Jose Jaena Mari Ople, Jun-Cheng Chen, and Kai-Lung Hua. 2022. [Social media popularity prediction based on multi-modal self-attention mechanisms](#). *IEEE Access*, 10:4448–4455.
- Yu-Shi Lin and Anthony J. T. Lee. 2024. [MMF: Winning solution to social media popularity prediction challenge 2024](#). In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11445–11449.
- Masoud Mazloom, Robert Rietveld, Stevan Rudinac, Marcel Worring, and Willemijn van Dolen. 2016. [Multimodal popularity prediction of brand-related social media posts](#). In *Proceedings of the 24th ACM*

- International Conference on Multimedia*, pages 197–201.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. [Learning transferable visual models from natural language supervision](#). In *Proceedings of the International Conference on Machine Learning*, pages 8748–8763.
- Marian-Andrei Rizoioiu, Lexing Xie, Scott Sanner, Manuel Cebrián, Honglin Yu, and Pascal Van Hentenryck. 2017. [Expecting to be HIP: Hawkes intensity processes for social media popularity](#). In *Proceedings of the 26th International Conference on World Wide Web*, pages 735–744.
- Neeti Sangwan and Vishal Bhatnagar. 2025. [Multi-branch LSTM encoded latent features with CNN-LSTM for YouTube popularity prediction](#). *Scientific Reports*, 15(1):2508.
- Satoshi Sanjo and Marie Katsurai. 2017. [Recipe popularity prediction with deep visual-semantic fusion](#). In *Proceedings of the 2017 ACM Conference on Information and Knowledge Management*, pages 2279–2282.
- Gábor Szabó and Bernardo A. Huberman. 2010. [Predicting the popularity of online content](#). *Communications of the ACM*, 53(8):80–88.
- Abir Troudi, Leila Ghorbel, Corinne Amel Zayani, Salma Jamoussi, and Ikram Amous. 2021. [MDER: Multi-dimensional event recommendation in social media context](#). *The Computer Journal*, 64(3):369–382.
- Renzhuo Wan, Shuping Mei, Jun Wang, Min Liu, and Fan Yang. 2019. [Multivariate temporal convolutional network: A deep neural networks approach for multivariate time series forecasting](#). *Electronics*, 8(8):876.
- Chenxu Wang, Dong Zhou, Ting Liu, Jianghao Lin, Yongmei Zhou, and Aimin Yang. 2025a. [DiffTMR: Diffusion-based hierarchical alignment for text-molecule retrieval](#). In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 6280–6288.
- Jing Wang, Shuo Yang, Hui Zhao, and Yue Yang. 2023. [Social media popularity prediction with multimodal hierarchical fusion model](#). *Computer Speech & Language*, 80:101490.
- Yashen Wang, Dongyang Li, Jianshan Sun, Kun Yuan, Yinan Jiang, Yin Zhang, and Jie Cao. 2025b. [Combining macro and micro: feature-driven dynamic graph learning for social media popularity prediction](#). *World Wide Web*, 28(1):18.
- Bo Wu, Wen-Huang Cheng, Peiye Liu, Bei Liu, Zhaoyang Zeng, and Jiebo Luo. 2019. [SMP challenge: An overview of social media prediction challenge 2019](#). In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2667–2671.
- Bo Wu, Tao Mei, Wen-Huang Cheng, and Yongdong Zhang. 2016. [Unfolding temporal dynamics: Predicting social media popularity using multi-scale temporal decomposition](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, pages 272–278.
- Jianmin Wu, Liming Zhao, Dangwei Li, Chen-Wei Xie, Siyang Sun, and Yun Zheng. 2022. [Deeply exploit visual and language information for social media popularity prediction](#). In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 7045–7049.
- Leilei Wu, Lingling Yi, Xiao-Long Ren, and Linyuan Lü. 2023. [Predicting the popularity of information on social platforms without underlying network structure](#). *Entropy*, 25(6):916.
- Mingjun Xi, Pingshan Liu, and Seshu Yu. 2025. [Temporal information-aware multimodal learning network for user-generated video popularity prediction](#). *Multimedia Systems*, 31(6):454.
- Jiangnan Xie, Xiaolong Zheng, and Liang Zheng. 2025. [Prototype-aware multimodal alignment for open-vocabulary visual grounding](#). *arXiv preprint arXiv:2509.06291*.
- Jiayi Xie, Yaochen Zhu, and Zhenzhong Chen. 2023. [Micro-video popularity prediction via multimodal variational information bottleneck](#). *IEEE Transactions on Multimedia*, 25:24–37.
- Baoyu Xu, Lingjie Zou, Shuhuai Li, Xiaoyang Kang, and Lihua Zhang. 2026. [A dynamic edge caching method based on popularity prediction with an improved LSTNet network](#). *Cluster Computing*, 29(1):55.
- Kele Xu, Zhimin Lin, Jianqiao Zhao, Peicang Shi, Wei Deng, and Huaimin Wang. 2020. [Multimodal deep learning for social media popularity prediction with attention mechanism](#). In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4580–4584.
- Kele Xu, Qisheng Xu, Binli Luo, Han Zhou, Zengming Lin, Hui Geng, and Xianhan Tan. 2025a. [Higher-order vision-language fusion for video popularity prediction](#). In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 14086–14093.
- Xovee Xu, Yifan Zhang, Fan Zhou, and Jingkuan Song. 2025b. [Improving multimodal social media popularity prediction via selective retrieval knowledge augmentation](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 932–940.
- Yijie Xu, Bolun Zheng, Wei Zhu, Hangjia Pan, Yuchen Yao, Ning Xu, Anan Liu, Quan Zhang, and Chenggang Yan. 2025c. [SMTPD: A new benchmark for temporal prediction of social media popularity](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18847–18857.

- Zhuoran Zhang, Shibiao Xu, Li Guo, and Wenke Lian. 2022. [Multi-modal variational auto-encoder model for micro-video popularity prediction](#). In *Proceedings of the 8th International Conference on Communication and Information Processing*, pages 9–16.
- Dong Zhou, Séamus Lawless, Xuan Wu, Wenyu Zhao, and Jianxun Liu. 2016. [Enhanced personalized search using social data](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 700–710.
- Jinyan Zhou, Xingang Wang, Ning Liu, Xiaoyu Liu, Jiandong Lv, Xiaomin Li, Hong Zhang, and Rui Cao. 2023. [Visual and linguistic double transformer fusion model for multimodal tweet classification](#). In *Proceedings of the International Joint Conference on Neural Networks*, pages 1–8.

A Qualitative Case Study

To further evaluate MMATF, we conduct a qualitative analysis on three representative YouTube videos across different languages exhibiting a classic high-popularity long-tailed decay over a 30-day window. The y -axis denotes log-normalized view counts to mitigate extreme scale variance.

As illustrated in Figure 4, most baseline models struggle to capture the true magnitude of viral dynamics. Traditional multimodal sequential models frequently underestimate popularity, likely due to suboptimal alignment between cross-lingual semantics and early numerical momentum. Similarly, the tree-based HyFea framework lacks the sequential inductive bias required to track fine-grained temporal evolution, resulting in flattened trajectories. In contrast, MMATF maintains a close alignment with the ground truth across all cases. This stability stems from its capacity for multimodal synergy and its autoregressive Transformer encoder, which tracks long-tail curves smoothly without future information leakage.

Specifically, in Case 1 and Case 3, MMATF accurately captures the initial peak magnitude and closely matches the subsequent stabilization trend, matching or outperforming the strongest baseline, TMALL. In Case 2, MMATF showcases notable stability, suggesting that its attention mechanism decisively interprets volatile multimodal signals where baselines may be confounded by scale. Overall, these qualitative trajectories corroborate our quantitative findings, supporting MMATF’s robustness in modeling long-tail lifecycles for cross-lingual social content.

B Detailed Ablation Configurations

To expand on the abbreviated notes presented in the main text, the comprehensive descriptions of the configurations utilized in our ablation studies are detailed as follows.

Component-Level Ablations:

- **w/o Dynamic Inter.:** Employs simple static feature concatenation instead of the dynamic multi-head modality interaction module.
- **w/o Autoregressive Enc.:** Replaces the autoregressive encoding mechanism entirely with a standard bidirectional LSTM architecture.
- **w/o Autoregressive Mask.:** Removes the causal mask, intentionally allowing future temporal information leakage during sequence modeling.

- **w/o Weight Sharing:** Removes the shared MLP layer, utilizing separate and unconstrained projection heads for each prediction step.

Modality-Level Ablations:

- **w/o Text:** Removes the textual path, relying only on visual, categorical, and numerical inputs.
- **w/o Visual:** Excludes the video cover image features extracted by ResNet-101.
- **w/o Category-Language:** Disables the discrete category and language classification embeddings.
- **w/o Numerical+EP:** Removes the continuous numerical features, including early-popularity (EP) observations collected within the first 24 hours.

C Architecture Tuning

We further investigate the influence of the number of attention heads and Transformer layers on the SMTPD dataset. Figure 5 presents the performance under different architectural configurations. Among the evaluated settings, the configuration with 8 attention heads and 4 Transformer layers achieves the most balanced performance, obtaining an MAE of 0.6387, a Loss of 0.3017, and an SRC of 0.9717.

Reducing the number of attention heads or Transformer layers generally leads to performance degradation, while further increasing model capacity does not yield consistent improvements. Therefore, we adopt 8 attention heads and 4 Transformer layers as the default configuration in the main experiments.

D Artifact Details

We utilize PyTorch (BSD-style license) for implementation. The SMTPD and ICIP datasets are used strictly for non-commercial academic research, consistent with their original intended purposes of advancing social media popularity prediction. The utilized SMTPD and ICIP datasets are public and were pre-processed by their original creators to anonymize users and mitigate personally identifying information (PII). We manually reviewed a random subset during data exploration and found no explicit PII or severely offensive content.

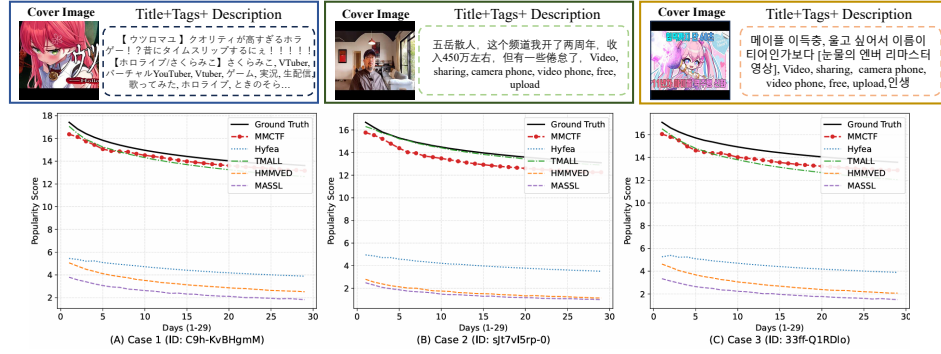


Figure 4: Qualitative case study comparing the prediction trajectories of our proposed MMATF against baseline models on three representative videos.

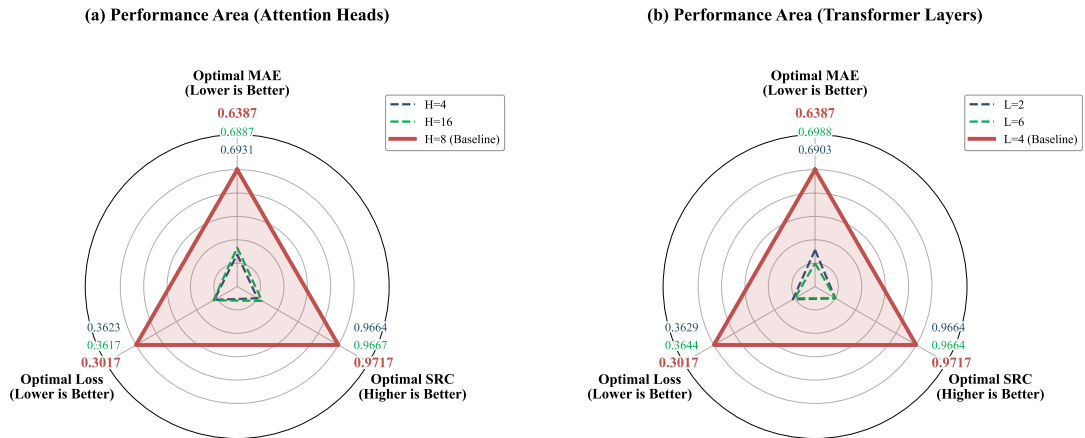


Figure 5: Architecture tuning results under different numbers of attention heads and Transformer layers. The left and right radar charts correspond to attention-head and Transformer-layer configurations, respectively.