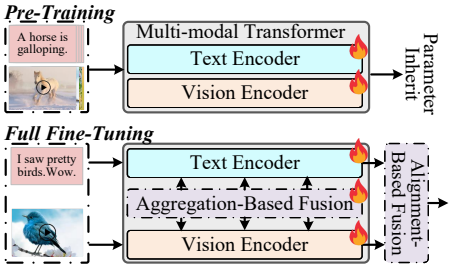
# In-context Prompt-augmented Micro-video Popularity Prediction

**Zhangtao Cheng, Jiao Li, Jian Lang, Ting Zhong, Fan Zhou***

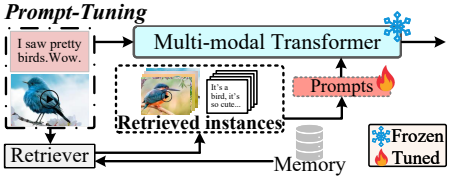
University of Electronic Science and Technology of China, Chengdu, Sichuan, China
zhangtao.cheng@outlook.com, jiao_li@std.uestc.edu.cn, jian_lang@std.uestc.edu.cn,
zhongting@uestc.edu.cn, fan.zhou@uestc.edu.cn

## Abstract

Micro-video popularity prediction (MVPP) plays a crucial role in various downstream applications. Recently, multi-modal methods that integrate multiple modalities to predict the popularity have exhibited impressive performance. However, these methods face several unresolved issues: (1) *limited contextual information* and (2) *incomplete modal semantics*. Incorporating relevant videos and performing full fine-tuning on pre-trained models typically achieves powerful capabilities in addressing these issues. However, this paradigm is not optimal due to its weak transferability and scarce downstream data. Inspired by prompt learning, we propose ICPF, a novel In-Context Prompt-augmented Framework to enhance popularity prediction. ICPF maintains a model-agnostic design, facilitating seamless integration with various multimodal fusion models. Specifically, the multi-branch retriever first retrieves similar modal content through within-modality similarities. Next, in-context prompt generator extracts semantic prior features from retrieved videos and generates in-context prompts, enriching pre-trained models with valuable contextual knowledge. Finally, knowledge-augmented predictor captures complementary features including modal semantics and popularity information. Extensive experiments conducted on three real-world datasets demonstrate the superiority of ICPF compared to 14 competitive baselines.




(a) Multimodal fusion methods.




(b) Our method (ICPF).

Figure 1: Motivation of our work. (a) Previous multi-modal fusion methods rely on full fine-tuning for prediction. (b) Our model first retrieves similar videos as in-context prompts, enriching multimodal transformers with contextual knowledge to enhance popularity prediction.

## Introduction

Micro-video popularity prediction (MVPP) aims to assess the popularity level of a given micro-video, which holds significant implications for product marketing, platform management, and government operations. For example, marketers can leverage popularity predictions to make informed decisions about which videos are most likely to resonate with their target audiences. Additionally, evaluating video content can assist governments in identifying potential public opinion crises (Tatar et al. 2014). Given its importance, substantial research efforts have been devoted to addressing MVPP across various real-world applications, such as online advertising (Xin et al. 2021; Hu et al. 2024), and social network analysis (Zhou et al. 2021; Cheng et al. 2023, 2024b).

With the success of deep learning, multimodal fusion strategies have shown significant potential for automatic popularity assessment (Zhou et al. 2021). Recently, researchers (Zhang et al. 2022; Du et al. 2023) have employed multimodal transformers (e.g., CLIP (Radford et al. 2021)) to directly map micro-videos to their popularity through end-to-end neural integration of multiple modalities. Existing multimodal fusion methods for the MVPP task are generally categorized into alignment-based fusion (Xie, Zhu, and Chen 2023) and aggregation-based fusion (Cheung and Lam 2022; Zhong et al. 2024) (see Fig. 1(a)). These approaches have become the dominant trend in the field of MVPP, as joint training on multimodal data significantly enhances micro-video analysis.

Despite significant progress, existing methods still face substantial challenges: **(1) Limited contextual information.** Prior works primarily rely on the limited modal cues available within individual micro-videos. However, the distribution of followers varies significantly among source users on social platforms, leading to substantial differences

*Corresponding author.

in social feedback for identical micro-videos depending on the viewing user group (Zhou et al. 2021). Consequently, modeling individual videos in isolation fails to provide an effective solution to the MVPP task. **(2) Incomplete modal semantics.** Cross-modal gaps between different modalities often render alignment-based fusion methods suboptimal, as the limited information exchange enforced by alignment loss is insufficient for robust representation learning. Additionally, aggregation-based fusion methods frequently overlook intra-modal propagation, leading to an imbalance between inter-modal knowledge sharing and intra-modal information processing. Thus, existing multimodal fusion approaches struggle to comprehensively capture and integrate semantic information from diverse modalities in the MVPP task.

To address these challenges, we draw inspiration from in-context learning (Brown et al. 2020) in the language domain, where language models perform diverse tasks and extract essential cues by leveraging a few task-specific input-label pairs as prompts. This motivates us to search similar micro-videos and capture valuable contextual information to mitigate the aforementioned challenges. Intuitively, for a given target micro-video, appending the social feedback from previously posted similar videos provides user-specific context and rich modal semantics, thereby enhancing popularity prediction. Considering that multimodal fusion methods typically rely on pre-trained multimodal transformers, we propose incorporating relevant contextual knowledge from similar instances into the fine-tuning process on task-oriented training datasets. While full fine-tuning is effective, it is computationally inefficient and imposes significant demands on parameter storage (Jia et al. 2021). Furthermore, fine-tuning multimodal transformers on small datasets often leads to instability (Mosbach, Andriushchenko, and Klakow 2021). Inspired by recent advances in prompt tuning from the language domain, we focus on improving adaptation efficiency for micro-video analysis by freezing the parameters of multimodal transformers and fine-tuning only the learnable prompt parameters (see Fig. 1(b)).

To this end, we propose ICPF, a novel **I**n-**C**ontext **P**rompt-augmented **F**ramework for enhancing MVPP. Technically, we reformulate popularity prediction in a principled prompt-and-predict manner. The core idea is to leverage valuable contextual knowledge from similar videos and use it as in-context prompts to fine-tune multimodal transformers, thereby improving prediction accuracy. Fundamentally, ICPF follows a model-agnostic design, enabling seamless integration with various multimodal fusion models. The framework consists of three key modules: a multi-branch retriever, an in-context prompt generator, and a knowledge-augmented predictor.

Specifically, the multi-branch retriever first disentangles modalities into uni-modality branches and identifies similar modal content through within-modality similarities, thereby alleviating cross-modal gaps. Second, in-context prompt generator perceives the semantic correlations between the target and retrieved videos and produces in-context prompts tailored for different inputs. The generated prompts enables an adjusting effect of incomplete information by using rich modal semantics from similar videos, thereby enhancing the MVPP task. The adaptive prompts are then integrated into multimodal transformers to achieve more accurate representations of videos. Finally, knowledge-augmented predictor adaptively selects the most suitable and crucial cues, extracting complementary fused features that incorporate multimodal semantics and popularity information of retrieved instances. Following are our main contributions:

- We propose ICPF, pioneering a in-context prompt-augmented framework that captures semantic correlations between the target and retrieved micro-videos, generating dynamic prompts to enhance popularity prediction.
- We design three key components including multi-branch retriever, in-context prompt generator, and knowledge-augmented predictor. These components facilitate the integration of expressive contextual semantics to enhance multimodal transformers in a model-agnostic manner.
- Extensive experiments conducted on three datasets demonstrate the effectiveness of ICPF in the MVPP task. Compared with 14 competitive baselines, our ICPF achieves 22%, 9%, and 8% improvements in terms of nMSE, MAE, and SRC on the TikTok dataset, respectively. The source codes and datasets are available at https://github.com/Jolieresearch/ICPF.

## Related Work

**Micro-video Popularity Prediction**. Researchers address the MVPP task using various methods that can be broadly divided into two categories: **(1) Feature-engineering methods** focus on designing and integrating hand-crafted modal features from various aspects of micro-videos (e.g., video content, and user engagement) into well-defined machine learning models (Khosla, Das Sarma, and Hamid 2014; Lai, Zhang, and Zhang 2020; Hsu et al. 2023). However, these methods are constrained by restrictive assumptions and require tedious feature extraction. **(2) Multimodal fusion methods** focuses on designing automatic model architectures for effective data modality extraction and integration (Cheng et al. 2024a). These approaches excel at capturing cross-modal correlations, handling complex high-dimensional data, and learning relevant features without extensive manual engineering. CBAN (Cheung and Lam 2022) integrates positive and negative attention mechanisms. MMVED (Xie, Zhu, and Chen 2023) adopts a variational paradigm for multimodal encoding-decoding. MMRA (Zhong et al. 2024) employs a retrieval-augmented strategy.

However, existing works still face several unresolved issues, including limited contextual information and incomplete modal semantics, which result in suboptimal performance. To address these issues, we propose a novel in-context prompt-augmented framework (ICPF) that leverages rich contextual information from relevant instances to enhance popularity prediction.

**Prompt Learning**. Prompt learning (Liu et al. 2023) is an emerging technique that conditions frozen pre-trained models to perform downstream tasks by incorporating learnable prompt parameters, which are prepended to the input tokens to guide model predictions. Initially proposed in the language domain (Brown et al. 2020), it later gained traction in the image domain (Jia et al. 2022) due to its flexibility and

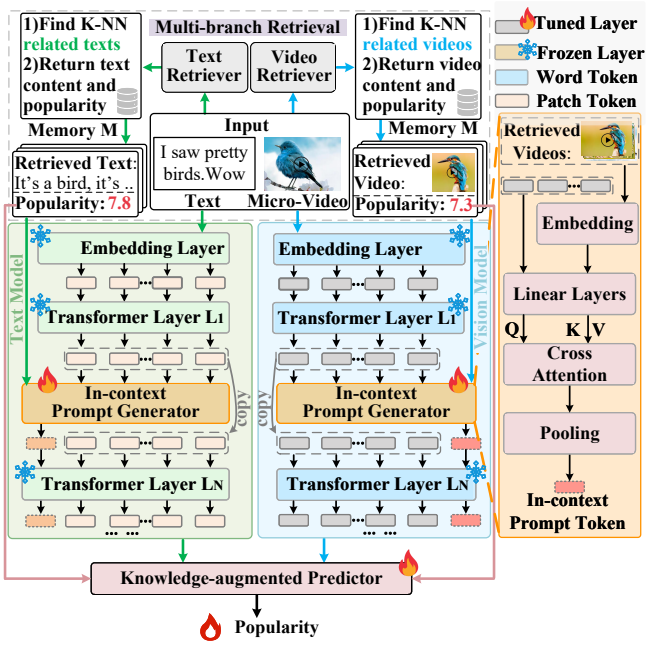



Figure 2: Overall framework of our ICPF.

excellent performance. Owing to its parameter-efficient nature, prompt learning has achieved significant success across various tasks, including image classification (Jia et al. 2022; Bahng et al. 2022) and graph-based tasks (Yu et al. 2024; Jiang et al. 2024). Following the success of prompt learning, recent works have extended its application to multimodal learning (Zhou et al. 2022). For example, CoOp (Zhou et al. 2022) was the first to introduce prompt tuning to vision-language models using learnable prompt template words. VPT (Jia et al. 2022) integrates learnable parameters directly into the vision encoder. MaPLe (Khattak et al. 2023) appends soft prompts to the hidden representations at each layer of both the text and image encoders, achieving notable improvements in few-shot image recognition. To the best of our knowledge, we are the first to introduce prompt learning into the field of MVPP to improve prediction accuracy.

## Methodology

**Problem Definition**: Let $\mathcal{V} = \{V_1, \cdots, V_N\}$ represent the set of micro-videos available on online video platforms, where $N$ is the total number of videos. Specifically, each video $V_i$ is characterized by a unified triplet-wise format $V_i = (\boldsymbol{v}_i, \boldsymbol{t}_i, y_i)$ for video-text-popularity instances, where $\boldsymbol{v}_i$ represents a video, $\boldsymbol{t}_i$ is its corresponding textual descriptions, and $y_i$ denotes the ground truth popularity. The goal of MVPP is to predict the cumulative views $\hat{y}_i$ of a given video $V_i$ during a specific future period via leveraging all modal content that significantly contributes to forecasting its popularity trend post-release. Formally, the task can be expressed as: $\hat{y}_i = f_\phi(\boldsymbol{v}_i, \boldsymbol{t}_i)$, where $f_\phi(\cdot)$ denotes popularity model and $\phi$ represents the model parameters. In this work, we introduce multi-modal in-context prompts to enable popularity models to capture expressive contextual information from relevant instances to enhance popularity prediction. Specifically, we define the popularity prediction for a given video $V_i$ and the multi-modal context $\mathcal{C}$ as follows:

$$\hat{y}_i = f_\phi\{\boldsymbol{v}_i; \underbrace{\boldsymbol{v}_1^c \cdots \boldsymbol{v}_K^c}_{\text{Vision Context}}\}, \{\boldsymbol{t}_i; \underbrace{\boldsymbol{t}_1^c \cdots \boldsymbol{t}_K^c}_{\text{Text Context}}\}, \{\underbrace{y_1^c \cdots y_K^c}_{\text{Label Context}}\}), \quad (1)$$

where the context $C = \{(\boldsymbol{v}_k^c, \boldsymbol{t}_k^c, y_k^c)\}_{k=1}^K$ is the set of in-context prompts, and $K$ is the number of prompts.

Fig. 2 illustrates the key components of ICPF and their relationships. The following sections provide a detailed explanation of each component and their respective implementations.

### Multi-branch Retriever

In this section, we design a unified multi-branch retriever to identify similar micro-video instances for a given query within their respective modalities by calculating within-modality similarities. This retrieval strategy effectively mitigates the modal gap during retrieval and addresses modality inconsistencies in micro-videos.

**Memory Bank Construction**. To retrieve relevant video instances, we construct a micro-video memory bank denoted as $\mathcal{M} = \{(\boldsymbol{v}_i, \boldsymbol{t}_i, y_i)\}_{i=1}^{|\mathcal{M}|}$, where $|\mathcal{M}|$ represents the total number of videos stored in the bank $\mathcal{M}$. This memory bank comprises reference triplets in the form of $(\text{frames}, \text{text}, \text{popularity})$. Leveraging the naturally diverse information within $\mathcal{M}$, we can capture rich semantic cues from relevant videos to assist in predicting the popularity of the target video.

**Multi-branch Retrieval**. Due to the subjective nature of human behavior, textual and visual modalities in micro-videos often exhibit inconsistencies. To address this issue, we design a multi-branch retriever that effectively explores the unique characteristics of each modality and retrieves relevant instances by computing within-modality similarities. With this design, the retriever better mitigates heterogeneity and addresses modality inconsistencies.

Specifically, for video retrieval, we first extract $n$ key frames $\{f_i\}_{i=1}^n$ and project each frame $f_i$ into $N$ overlapping patches, denoted as $\mathcal{P}_i = \{p_1, p_2, \ldots, p_N\}$. All frame patches are then mapped into vision embeddings $\tilde{\mathbf{E}}_i^v \in \mathbb{R}^{N \times d_v}$ via a linear transformation. $d_v$ is the dimension of the patch tokens and $N$ represents the number of vision tokens. Subsequently, the vision embeddings are fed into a pre-trained vision encoder $\Psi_v$ (e.g., ViT (Dosovitskiy et al. 2021)) to generate the visual representations (i.e., the [CLS] token from the output of the last hidden state) denoted as $\Psi_v(\tilde{\mathbf{E}}_i^v) \in \mathbb{R}^{N \times d_v}$. Furthermore, we employ a non-parametric strategy that performs average pooling over all the frame-level visual representations to obtain the visual retrieval query vector $\mathbf{E}^v \in \mathbb{R}^{d_v}$ for the video. Finally, we compute the cosine similarity between the visual query vector $\mathbf{E}^v$ and the visual vectors stored in the memory bank $\mathcal{M}$ to retrieve the top-$K$ relevant instances:

$$\mathcal{S}^v = \underset{m \in \mathcal{M}}{\text{Top-}K}\left(\frac{\mathbf{E}^{v\top}\mathbf{E}^{m,v}}{\|\mathbf{E}^v\| \cdot \|\mathbf{E}^{m,v}\|}\right). \quad (2)$$

For text retrieval, we first tokenize the textual descriptions into a sequence of words and convert them into text embeddings $\tilde{\mathbf{E}}^t \in \mathbb{R}^{m \times d_t}$, where $m$ represents the number of word tokens and $d_t$ denotes the dimensionality of the text embeddings. Subsequently, the text embeddings are fed into a pre-trained text encoder $\Psi_t$ (e.g., AnglE (Li and Li 2023)) to generate the textual retrieval query vector $\mathbf{E}^t \in \mathbb{R}^{d_t}$ (i.e., [CLS] token from the output of the last hidden state). Following the retrieval process described in Eq. 2, we retrieve the top-$K$ most similar instances corresponding to the textual descriptions.

After the two-branch retrieval, we obtain the retrieved texts $\{t_i^c\}_{i=1}^K$ and videos $\{v_i^c\}_{i=1}^K$, which serve as contextual information to generate in-context prompts for guiding reasoning within pre-trained multimodal models.

## In-context Prompt Learning

In this section, we design dynamic in-context prompts that capture expressive contextual information from retrieved instances, thereby assisting pre-trained multimodal models in improving popularity prediction.

**Dynamic Prompt Learning.** Existing study (Dong et al. 2022) demonstrates that providing large language models with a few task-specific prompts significantly enhances model inference for language reasoning tasks. For instance, in machine translation, a structured prompt such as $\{\text{Example} : \textit{sea otter} \rightarrow \textit{loutre de mer}, \text{Query} : \textit{cheese} \rightarrow ?\}$ enables the model to learn from examples and translate "cheese" into "fromage". Inspired by this observation, we introduce in-context prompts to incorporate expressive contextual information derived from retrieved instances for prompt tuning the pre-trained model (i.e., CLIP (Radford et al. 2021)). Specifically, in the language branch, we inject the textual in-context prompts $\boldsymbol{P}_t^c$ (the details are provided in the next subsection) into the representations $\mathbf{h}_i^t \in \mathbb{R}^{m \times d_t}$ at the $i$-th layer of the language encoder, resulting in new learnable representations $\mathbf{h}_i^{t,p} = [\mathbf{P}_t^c, \mathbf{h}_i^t]$. Subsequently, the features $\mathbf{h}_i^{t,p}$ are passed to the $(i+1)$-th transformer layer, producing the corresponding features $\mathbf{h}_{i+1}^t$. Finally, after processing through the $N$-th layer, we obtain the final textual representations $\mathbf{H}^t \in \mathbb{R}^{m \times d_t}$, which incorporate rich contextual semantic information derived from text-related videos. To further reduce computational overhead, we strategically insert the dynamic prompts into a specific layer of the language encoder. Analogously, in the vision branch, we employ the same prompt-tuning process to capture video-wise contextual information, thereby obtaining the final visual representations $\mathbf{H}^v \in \mathbb{R}^{n \times d_v}$.

**In-context Prompt Generator.** To explicitly capture expressive contextual information, we design an in-context prompt generator that constructs text-wise and vision-wise dynamic prompts from the retrieved instances. For the text-wise prompts, we first tokenize and project the retrieved texts $\{t_i^c\}_{i=1}^K$ into text embeddings, denoted as $\mathbf{E}^{c,t} = \{\mathbf{E}_i^{c,t}\}_{i=1}^K \in \mathbb{R}^{K \times m \times d_t}$, where $m$ is the number of word tokens and $d_t$ denotes the dimension of the text embeddings. Subsequently, we use the $i$-th layer representations $\mathbf{h}_i^t$ from the language encoder as a query to interact with the retrieved text features $\mathbf{E}^{c,t}$ through a cross-attention mechanism. This design enables the capture of fine-grained modal semantic information from all retrieved texts and further facilitates the generation of high-quality text-wise prompts $\mathbf{P}_t^c \in \mathbb{R}^{m \times d_t}$. This process can be formulated as follows:

$$\mathbf{P}_t^c = \text{CrossAtt}(\mathbf{W}_t^Q \mathbf{h}_i^t, \mathbf{W}_t^K \mathbf{E}^{c,t}, \mathbf{W}_t^V \mathbf{E}^{c,t}), \quad (3)$$

$$\text{CrossAtt}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\right)\mathbf{V}, \quad (4)$$

where $\mathbf{W}_t^Q$, $\mathbf{W}_t^K$, and $\mathbf{W}_t^V$ are the projection matrices for the query, key, and value, respectively. Analogously, for the visual prompt, we first extract $n$ key frames from the $K$ retrieved videos and feed these frames into the vision encoder to obtain the retrieved visual representations $\mathbf{E}^{c,v} \in \mathbb{R}^{K \times n \times d_v}$. Next, we use the $i$-th layer's visual representations $\mathbf{h}_i^v$ from the vision encoder as the query to combine with the retrieved visual features $\mathbf{E}^{c,v}$ through a cross-attention, thereby generating the vision-wise prompts $\mathbf{P}_v^c \in \mathbb{R}^{n \times d_v}$.

## Knowledge-augmented Prediction

In this section, we utilize the cross-modal attention to effectively capture crucial information from both modalities and generates complementary representations for prediction. Additionally, we design a knowledge-augmented predictor that incorporates the popularity information from the retrieved instances to further enhance the prediction.

**Multimodal Fusion.** Specifically, we first project the textual representations $\mathbf{H}^t$ and visual representations $\mathbf{H}^v$ into the same dimension using linear projections, i.e., $\tilde{\mathbf{H}}^t \in \mathbb{R}^{m \times d}$ and $\tilde{\mathbf{H}}^v \in \mathbb{R}^{n \times d}$. Next, we treat the visual representations $\tilde{\mathbf{H}}^v$ as the query and the textual representations $\tilde{\mathbf{H}}^t$ as the key and value to perform cross-modal attention:

$$\mathbf{H}_t^v = \text{CrossAtt}(f_Q(\tilde{\mathbf{H}}^v), (f_K(\tilde{\mathbf{H}}^t), (f_V(\tilde{\mathbf{H}}^t)), \quad (5)$$

where $f_Q$, $f_K$, and $f_V$ represent the linear transformation functions applied to the query, key, and value, respectively. The visual-guided modal representations are denoted as $\mathbf{H}_t^v \in \mathbb{R}^{m \times d}$, where $m$ is the number of visual tokens and $d$ is the dimensionality of the feature space. Analogously, the textual representations $\tilde{\mathbf{H}}^t$ are employed as the query, while the visual representations $\tilde{\mathbf{H}}^v$ serve as both the key and value to derive the textual-guided modal representations $\mathbf{H}_v^t \in \mathbb{R}^{n \times d}$, where $n$ represents the number of textual tokens. Subsequently, we concatenate the text- and vision-guided representations, and utilize the average pooling strategy to obtain the modal fusion representation $\mathbf{H}_f = \text{pool}([\mathbf{H}_t^v, \mathbf{H}_v^t]) \in \mathbb{R}^d$.

**Label-augmented Prediction.** Given that similar videos often exhibit comparable popularity trends, we aim to leverage the valuable popularity-level information from the retrieved instances. Specifically, we first map the retrieved labels $\{y_i^c\}_{i=1}^K$ into their corresponding embeddings $\tilde{\mathbf{H}}_l = f_L(\{y_i^c\}_{i=1}^K) \in \mathbb{R}^{K \times d}$, where $f_L$ represents a fully connected layer. We then aggregate the $K$ label embeddings via

| Dataset | # Video | # User | # Train | # Val | # Test | V | T |
|---|---|---|---|---|---|---|---|
| MicroLens | 19,738 | 100,000 | 15,790 | 1,974 | 1,974 | 768 | 768 |
| NUS | 169,103 | 11,084 | 135,282 | 16,911 | 16,910 | 768 | 768 |
| TikTok | 21,846 | 155,684 | 17476 | 2185 | 2185 | 768 | 768 |

Table 1: Statistics of three datasets.

pooling to obtain the final label representation $\mathbf{H}_l \in \mathbb{R}^d$. Subsequently, we concatenate the label-enhanced representation $\mathbf{H}_l$ with the modal fusion representations $\mathbf{H}_f$, and pass the resulting vector through a multi-layer perceptron (MLP) to predict the popularity $\hat{y} = \mathrm{MLP}([\mathbf{H}_f, \mathbf{H}_l])$. During training, we freeze all parameters in the pre-trained language-vision model and optimize only the prompt generator and knowledge-aware predictor using mean squared error (MSE) loss.

# Experiments

We now present our experimental results to validate the efficacy of ICPF and address the following research questions:

- **RQ1:** How does the performance of ICPF compare to all strong baselines across three different datasets?
- **RQ2:** How do the crucial components within our ICPF contribute to the overall performance?
- **RQ3:** How does the model performance change when adjusting the key hyperparameters of ICPF?
- **RQ4:** How does the integration of in-context prompts enhance the performance of our ICPF model?

## Experimental Settings

- *Dataset*. To analyze the effectiveness of our ICPF, we select three real-world micro-video datasets: MicroLens (Ni et al. 2023), NUS (Chen et al. 2016), and TikTok (https://www.tiktok.com/), from various online video platforms. The detailed statistics of three datasets are summarized in Table 1. Each dataset is randomly divided into training, validation, and test sets in a ratio of 8:1:1. $\mathbf{V}$, $\mathbf{T}$ are the dimensions of visual and textual features, respectively.

- *Baseline*. To evaluate the model's superiority, we compare ICPF with 14 strong baselines classified into three groups: *(1) Feature-engineering*: SVR (Khosla, Das Sarma, and Hamid 2014), HyFea (Lai, Zhang, and Zhang 2020), and MFTM (Hsu et al. 2023). *(2) Multimodal fusion methods*: CLSTM (Ghosh et al. 2016), TMALL (Chen et al. 2016), MASSL (Zhang et al. 2022), CBAN (Cheung and Lam 2022), HMMVE (Xie, Zhu, and Chen 2023), JAB (Weissburg, Kumar, and Dhillon 2022), MQMC (Du et al. 2023), and MMRA (Zhong et al. 2024). *(3) Prompt-tuning methods*: CLIP$^*$ (Radford et al. 2021), BLIP$^*$ (Li et al. 2022), and MaPLe (Khattak et al. 2023). Here, $*$ indicates that performing static prompt tuning on pre-trained models.

- *Metrics.* Following prior works (Chen et al. 2016; Zhong et al. 2024), we utilize three widely used metrics: normalized Mean Squared Error (**nMSE**), Mean Absolute Error (**MAE**) and Spearman's Rank Correlation (**SRC**).

- *Model implementation.* During retrieval, we utilize ViT-B/32 CLIP (Radford et al. 2021) as the image encoder and AnglE (Li and Li 2023) as the text encoder. Furthermore, We utilize the AdamW optimizer (Loshchilov and Hutter 2017) with a learning rate of $1 \times 10^{-4}$ for optimizing the parameters. The model is trained for 30 epochs with a batch size of 64 and tested with a batch size of 256.

## Main Results (RQ1)

We compare ICPF with 14 strong baselines to validate its effectiveness. The results are presented in Table 2. Based on these results, we make the following observations:

**(O1)**: Our ICPF consistently outperforms all competitive baselines across all datasets and metrics. Additionally, we retrain both ICPF and the best-performing baselines five times to assess the p-value. Notably, our ICPF achieves improvements of over 22%, 9%, and 8% in terms of nMSE, MAE, and SRC, respectively, on the TikTok dataset. These findings validate the effectiveness of our design in extracting contextual information to guide pre-trained models for enhancing the micro-video analysis through in-context prompts. Compared to all strong baselines, our ICPF generates more accurate micro-video popularity, attributed to its adaptive in-context prompts. Specifically, the multi-branch retriever enables the effective retrieval of relevant multimodal content, thereby enhancing the efficacy of in-context prompts. In-context prompter distills contextual knowledge from this pertinent multimodal content, providing critical cues related to social feedback that assist the popularity prediction within pre-trained multimodal transformers.

**(O2)**: Our ICPF significantly outperforms both feature-engineering and deep-learning methods. We attribute the performance deficiencies of these baselines to their inherent limitations in adequately exploiting contextual information from relevant multimodal content, as well as their inability to align with the powerful capabilities of pre-trained multimodal transformers in the MVPP task. Furthermore, ICPF demonstrates superior performance compared to prompt-tuning methods, thereby highlighting the limitations of static prompt strategies in the MVPP task. This also indicates that our proposed in-context prompts effectively establish rich correlations between the retrieved and target videos, guiding the popularity predictions of pre-trained models among different micro-videos.

## Ablation Study (RQ2)

We conduct extensive ablation experiments to evaluate the impact of each critical component in ICPF. The results are summarized in Table 3.

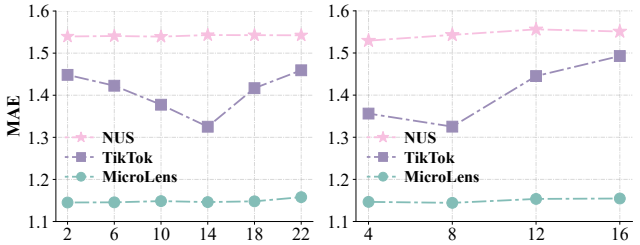
• **Effect of multi-branch retrieval.** To analyze the multi-branch retrieval, we design two variant models: (1) **Cross**: employing a cross-modal retriever to replace the retriever in our ICPF, and (2) **w/o R**: completely removing the retriever and replacing it with random instances. Specifically, we observe a significant decline in performance for both variants. These results validate the existence of model discrepancies within the field of MVPP, indicating that cross-modal retrieval is less effective in identifying suitable matches. Furthermore, they confirm the effectiveness of our design for multi-branch retrieval, which mitigates modality discrepancies and ensures the selection of the most relevant instances.

| Dataset | Metric | SVR | HyFea | MFTM | CLSTM | TMALL | MASSL | CBAN | HMMVE | JAB | MQMC | MMRA | CLIP* | BLIP* | MaPLe | **ICPF** | Improv. | p-val. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MicroLens | nMSE | 0.8132 | 0.8106 | 0.7814 | 0.7966 | 0.9373 | 1.0797 | 0.8106 | 0.8632 | 1.0821 | 1.1130 | 0.7916 | 0.7711 | 0.7853 | 1.0177 | **0.7338** | 4.84% ↑ | 3.98e-05 |
| | MAE | 1.2176 | 1.2321 | 1.2106 | 1.2117 | 1.2990 | 1.4136 | 1.2176 | 1.2524 | 1.4226 | 1.3827 | 1.1981 | 1.1867 | 1.1924 | 1.3685 | **1.1460** | 3.43% ↑ | 3.01e-07 |
| | SRC | 0.4288 | 0.4345 | 0.4477 | 0.4573 | 0.3817 | 0.3875 | 0.4463 | 0.3716 | 0.0141 | 0.2474 | 0.4712 | 0.4795 | 0.4699 | 0.2885 | **0.5200** | 8.45% ↑ | 1.35e-07 |
| NUS | nMSE | 0.6736 | 0.7260 | 0.6443 | 0.6445 | 0.8397 | 1.0052 | 0.6382 | 0.6837 | 0.9372 | 1.0110 | 0.6275 | 0.5978 | 0.6945 | 1.0108 | **0.5730** | 4.15% ↑ | 1.12e-05 |
| | MAE | 1.6688 | 1.7294 | 1.6518 | 1.6530 | 1.8884 | 1.9924 | 1.6068 | 1.6805 | 2.0023 | 2.0583 | 1.6129 | 1.5665 | 1.7141 | 2.0832 | **1.5199** | 2.97% ↑ | 1.39e-06 |
| | SRC | 0.5928 | 0.5318 | 0.5865 | 0.5919 | 0.4418 | 0.5149 | 0.6060 | 0.5587 | 0.3060 | 0.2179 | 0.6067 | 0.6278 | 0.5600 | 0.3956 | **0.6548** | 4.30% ↑ | 2.53e-06 |
| TikTok | nMSE | 0.5073 | 0.5171 | 0.4024 | 0.4146 | 0.5351 | 0.4729 | 0.3923 | 0.4821 | 1.0463 | 0.9881 | 0.4016 | 0.4076 | 0.4859 | 1.0046 | **0.3048** | 22.30% ↑ | 6.71e-06 |
| | MAE | 1.6105 | 1.7577 | 1.5036 | 1.5392 | 1.8086 | 1.6243 | 1.4022 | 1.5531 | 2.3930 | 2.4206 | 1.3861 | 1.5363 | 1.7975 | 2.4822 | **1.2539** | 9.53% ↑ | 3.29e-05 |
| | SRC | 0.6776 | 0.5906 | 0.6708 | 0.6746 | 0.5843 | 0.6340 | 0.6992 | 0.6579 | 0.0080 | 0.2812 | 0.6932 | 0.6683 | 0.6081 | 0.2030 | **0.7554** | 8.04% ↑ | 1.19e-05 |

Table 2: Performance comparison on three real-world datasets. The best results are in **bold** font and the second underlined. Lower values of nMSE and MAE, and higher values of SRC, indicate better performance.

| **Datasets** | | **MicroLens** | | | **TikTok** | | |
|---|---|---|---|---|---|---|---|
| Module | Variant | nMSE | MAE | SRC | nMSE | MAE | SRC |
| **ICPF** | **All** | **0.7338** | **1.1460** | **0.5200** | **0.3048** | **1.2539** | **0.7554** |
| Retrieval | Cross | 0.7441 | 1.1484 | 0.5125 | 0.3052 | 1.2708 | 0.7530 |
| | w/o R | 0.7435 | 1.1509 | 0.5090 | 0.3180 | 1.3136 | 0.7430 |
| Prompt | Static | 0.7629 | 1.1670 | 0.5055 | 0.3705 | 1.4144 | 0.7007 |
| | w/o L | 0.7416 | 1.1448 | 0.5198 | 0.3343 | 1.3621 | 0.7187 |
| | w/o P | 0.8788 | 1.2731 | 0.3454 | 0.6487 | 1.9274 | 0.4548 |
| Modal Content | w/o V | 0.8810 | 1.2718 | 0.3533 | 0.5008 | 1.5599 | 0.6232 |
| | w/o T | 0.7408 | 1.1503 | 0.5057 | 0.3989 | 1.4689 | 0.6485 |

Table 3: Ablation study of ICPF on key components.




(a) Effect of $K$ (b) Effect of $L$

Figure 3: Sensitivity analysis of ICPF for the number of retrieved videos $K$ and prompt length $L$ across three datasets.

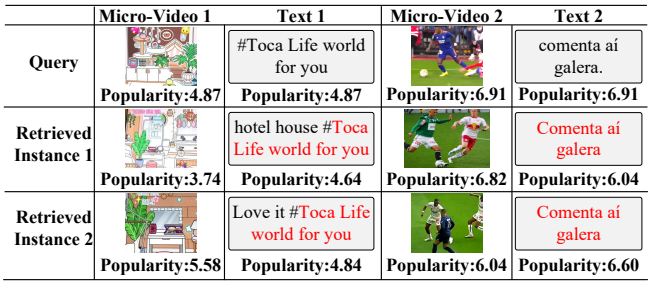



Figure 4: Examples of retrieved Top-2 nearest instances.

• **Effect of in-context prompts.** To assess the effect of in-context prompts, we design three variants: (1) **Static**, which utilizes static prompts to fine-tune pre-trained models; (2) **w/o L**, which removes popularity-level prompts; and (3) **w/o P**, which entirely eliminates the in-context prompts. In Table 3, we observe a significant performance drop when utilizing static prompt-tuning. This finding validates the limitation of static prompts, which are instance-agnostic and provide restricted contextual information for guiding popularity assessment. Additionally, we note a decline in performance when other prompts are removed, indicating that all in-context prompts are beneficial for the MVPP task, as they encompass rich contextual information associated with social feedback from similar videos.

• **Effect of modalities.** We separately remove the visual modality (**w/o V**) and the textual modality (**w/o T**) to evaluate the impact of each modality. Specifically, we find that the removal of either modality results in significant performance degradation. This indicates that both modalities are beneficial for the MVPP task, as multiple modalities provide complementary information.

## Hyper-parameter Analysis (RQ3)

Fig.3 presents an analysis of the key parameters in our ICPF.

• **Number of retrieved instances** $K$. Fig.3(a) illustrates the effect of the number of $K$ across three datasets. We observe that ICPF is not sensitive to the value of $K$ on the MicroLens and NUS datasets. In contrast, on the TikTok dataset, the performance of ICPF initially improves as $K$ increases, but subsequently declines when $K$ becomes larger. We speculate that this decline is due to the introduction of noise (irrelevant videos) into the model. To balance performance across all datasets, we determine that $K = 14$ is optimal.

• **In-context prompt length** $L$. Fig. 3(b) presents the impact of prompt length $L$ on three datasets. Based on the results, we observe that the performance of ICPF initially improves as $L$ increases, but subsequently declines when $L$ becomes larger on the TikTok dataset. The optimal value of $L$ for this dataset is 8. Furthermore, ICPF is not sensitive to varing values of $L$ on the MicroLens and NUS datasets. Therefore, we set $L = 8$ to achieve balance across all datasets.

## Generalizability, Robustness, Scalability & Retrieval Quality (RQ4)

• **Retrieval quality**. We visualize the retrieved instances to evaluate the multi-branch retrieval quality. Fig. 4 presents Top-2 similar instances associated with two randomly selected videos from the TikTok dataset. Specifically, we observe strong correlations between retrieved instances and target videos, including modal content and popularity. This evidence supports the efficacy of our designed multi-branch retriever. Moreover, these results validate our hypothesis that appending social feedback from previously posted similar videos provides valuable contextual information.

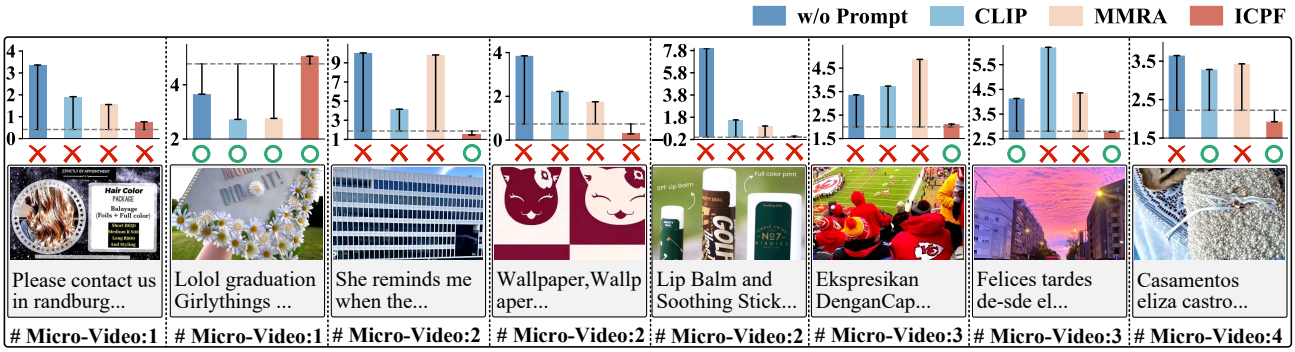
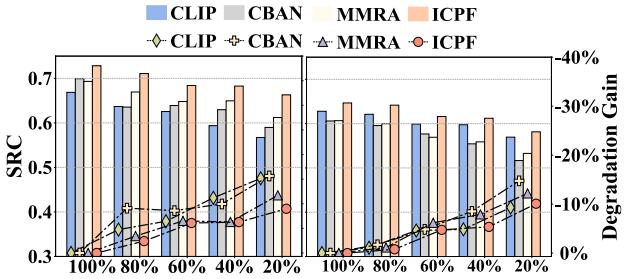



Figure 5: Visualization of model prediction on the TikTok dataset. Top row: error between model prediction scores and ground truth of a variant, CLIP, MMRA and ICPF. Grey dashed line represents the ground truth. The displayed values indicate the prediction errors relative to the ground truth. Red crosses denote error rates exceeding 50%, green circles indicate errors below 50%. Middle row: micro-videos and their texts. Bottom row: the number of published videos for each cold-start user.




(a) TikTok (b) NUS

Figure 6: The effect of training set proportion on TikTok and NUS datasets. The bars represent SRC values and the polylines denote the performance degradation gain.

| Datasets | MicroLens | | MicroLens† | |
|---|---|---|---|---|
| Backbone | MAE | SRC | MAE | SRC |
| **CLIP** | 1.2030 | 0.4648 | $1.1477_{(-0.055)}$ | $0.5174_{(+0.052)}$ |
| **BLIP** | 1.3124 | 0.2731 | $1.1922_{(-0.120)}$ | $0.4949_{(+0.221)}$ |
| **MT1** | 1.2118 | 0.4558 | $1.1460_{(-0.065)}$ | $0.5200_{(+0.064)}$ |
| **MT2** | 1.1923 | 0.4738 | $1.1808_{(-0.011)}$ | $0.4926_{(+0.018)}$ |

Table 4: Performance on different multimdal backbones on the MicroLens dataset. † denotes the method attached with in-context prompts.

• **Model generalizability**. We investigate the generalizability of our ICPF for cold-start users (i.e., those who have published fewer than five videos) in comparison to two strong baselines (i.e., CLIP and MMRA) and a variant (w/o prompt). Fig.5 illustrates the prediction errors of eight randomly selected videos from TikTok's test set. Specifically, MMRA, CLIP, and the variant achieve an error rate below 50% in only one, two, and one instance, respectively. In contrast, ICPF demonstrates superior performance, achieving an error rate below 50% in five instances. Notably, ICPF consistently exhibits lower prediction errors than its counterparts across all instances. These observations underscore the generalizability of our ICPF for newly emergent videos from cold-start users, facilitated by the use of in-context prompts with only 3% of tunable parameters. Furthermore, this indicates that capturing contextual information is beneficial for enhancing model generalizability.

• **Model robustness**. To assess the model's robustness, we compare our ICPF with three competitive baselines (i.e., HMMVED, CLIP, and MMRA) across varying portions of the training set. As shown in Fig.6, we find that our ICPF consistently outperforms all baselines across all training conditions while exhibiting a lower performance drop compared to all baselines. This indicates that our ICPF demonstrates greater robustness, which can be attributed to the incorporation of in-context prompts that provide valuable contextual knowledge for effectively aligning the pre-trained transformers with the MVPP task.

• **Model scalability**. To further validate the model's scalability, we select two multimodal backbones (i.e., CLIP and BLIP) and design two variants: MT1 and MT2. MT1 combines the AngIE and VIT models, while MT2 integrates the AngIE and Dinov2 models. In Table 4, we observe a performance improvement across the four backbones. This finding validates the effectiveness of our design in extracting informative multimodal cues from relevant instances and prompting multimodal transformers.

## Conclusion

In this work, we propose ICPF, a novel in-context prompt-augmented framework designed to enhance MVPP. This model-agnostic architecture incorporates key components, including a multi-branch retriever, an in-context prompt generator, and a knowledge-augmented predictor, to effectively capture valuable contextual information, thereby enhancing the robustness of multimodal transformers in the MVPP task. Extensive experiments conducted on three real-world datasets demonstrate the effectiveness of ICPF. Future work will focus on extending ICPF to other diverse domains, including multimodal recommendation and multimodal graph learning.

## Acknowledgments

This work was supported by National Natural Science Foundation of China (Grant No.62176043, No.62072077, and No.U22A2097).

## References

Bahng, H.; Jahanian, A.; Sankaranarayanan, S.; and Isola, P. 2022. Exploring visual prompts for adapting large-scale models. *arXiv preprint arXiv:2203.17274*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing systems (Neurips)*, 1877–1901.

Chen, J.; Song, X.; Nie, L.; Wang, X.; Zhang, H.; and Chua, T.-S. 2016. Micro tells macro: Predicting the popularity of micro-videos via a transductive model. In *ACM International Conference on Multimedia (MM)*, 898–907.

Cheng, Z.; Ye, W.; Liu, L.; Tai, W.; and Zhou, F. 2023. Enhancing Information Diffusion Prediction with Self-Supervised Disentangled User and Cascade Representations. In *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM)*, 3808–3812.

Cheng, Z.; Zhang, J.; Xu, X.; Trajcevski, G.; Zhong, T.; and Zhou, F. 2024a. Retrieval-augmented hypergraph for multimodal social media popularity prediction. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 445–455.

Cheng, Z.; Zhou, F.; Xu, X.; Zhang, K.; Trajcevski, G.; Zhong, T.; and Philip, S. Y. 2024b. Information Cascade Popularity Prediction via Probabilistic Diffusion. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*.

Cheung, T.-h.; and Lam, K.-m. 2022. Crossmodal bipolar attention for multimodal classification on social media. *Neurocomputing*, 1–12.

Dong, Q.; Li, L.; Dai, D.; Zheng, C.; Wu, Z.; Chang, B.; Sun, X.; Xu, J.; and Sui, Z. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations (ICLR)*.

Du, Y.; Wei, Y.; Ji, W.; Liu, F.; Luo, X.; and Nie, L. 2023. Multi-queue Momentum Contrast for Microvideo-Product Retrieval. In *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM)*, 1003–1011.

Ghosh, S.; Vinyals, O.; Strope, B.; Roy, S.; Dean, T.; and Heck, L. 2016. Contextual lstm (clstm) models for large scale nlp tasks. *arXiv preprint arXiv:1602.06291*.

Hsu, C.-C.; Lee, C.-M.; Hou, X.-Y.; and Tsai, C.-H. 2023. Gradient Boost Tree Network based on Extensive Feature Analysis for Popularity Prediction of Social Posts. In *ACM International Conference on Multimedia (MM)*, 9451–9455.

Hu, Z.; Nakagawa, S.; Zhuang, Y.; Deng, J.; Cai, S.; Zhou, T.; and Ren, F. 2024. Hierarchical Denoising for Robust Social Recommendation. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning (ICML)*, 4904–4916.

Jia, M.; Tang, L.; Chen, B.-C.; Cardie, C.; Belongie, S.; Hariharan, B.; and Lim, S.-N. 2022. Visual prompt tuning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 709–727.

Jiang, X.; Qiu, R.; Xu, Y.; Zhang, W.; Zhu, Y.; Zhang, R.; Fang, Y.; Chu, X.; Zhao, J.; and Wang, Y. 2024. RAGraph: A General Retrieval-Augmented Graph Learning Framework. *arXiv preprint arXiv:2410.23855*.

Khattak, M. U.; Rasheed, H.; Maaz, M.; Khan, S.; and Khan, F. S. 2023. Maple: Multi-modal prompt learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 19113–19122.

Khosla, A.; Das Sarma, A.; and Hamid, R. 2014. What makes an image popular? In *Proceedings of the ACM Web Conference (WWW)*, 867–876.

Lai, X.; Zhang, Y.; and Zhang, W. 2020. HyFea: Winning solution to social media popularity prediction for multimedia grand challenge 2020. In *ACM International Conference on Multimedia (MM)*, 4565–4569.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. C. H. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *International Conference on Machine Learning (ICML)*, 12888–12900.

Li, X.; and Li, J. 2023. Angle-optimized text embeddings. *arXiv preprint arXiv:2309.12871*.

Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; and Neubig, G. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9): 1–35.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Mosbach, M.; Andriushchenko, M.; and Klakow, D. 2021. On the Stability of Fine-tuning BERT: Misconceptions, Explanations, and Strong Baselines. In *International Conference on Learning Representations (ICLR)*.

Ni, Y.; Cheng, Y.; Liu, X.; Fu, J.; Li, Y.; He, X.; Zhang, Y.; and Yuan, F. 2023. A Content-Driven Micro-Video Recommendation Dataset at Scale. *arXiv preprint arXiv:2309.15379*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 8748–8763.

Tatar, A.; De Amorim, M. D.; Fdida, S.; and Antoniadis, P. 2014. A survey on predicting the popularity of web content. *Journal of Internet Services and Applications*, (1): 1–20.

Weissburg, E.; Kumar, A.; and Dhillon, P. S. 2022. Judging a book by its cover: Predicting the marginal impact of title on Reddit post popularity. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 1098–1108.

Xie, J.; Zhu, Y.; and Chen, Z. 2023. Micro-Video Popularity Prediction Via Multimodal Variational Information Bottleneck. *IEEE Transactions on Multimedia (TMM)*, 24–37.

Xin, S.; Li, Z.; Zou, P.; Long, C.; Zhang, J.; Bu, J.; and Zhou, J. 2021. ATNN: adversarial two-tower neural network for new item's popularity prediction in E-commerce. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, 2499–2510.

Yu, X.; Zhou, C.; Fang, Y.; and Zhang, X. 2024. MultiGPrompt for multi-task pre-training and prompting on graphs. In *Proceedings of the ACM on Web Conference 2024*, 515–526.

Zhang, Z.; Xu, S.; Guo, L.; and Lian, W. 2022. Multi-modal Variational Auto-Encoder Model for Micro-video Popularity Prediction. In *Proceedings of the International Conference on Communication and Information Processing (ICCIP)*, 9–16.

Zhong, T.; Lang, J.; Zhang, Y.; Cheng, Z.; Zhang, K.; and Zhou, F. 2024. Predicting Micro-video Popularity via Multimodal Retrieval Augmentation. In *Proceedings of the ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, 2579–2583.

Zhou, F.; Xu, X.; Trajcevski, G.; and Zhang, K. 2021. A survey of information cascade analysis: Models, predictions, and recent advances. *ACM Computing Surveys (CSUR)*, 54(2): 1–36.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision (IJCV)*, (9): 2337–2348.