REAL: Retrieval-Augmented Prototype Alignment for Improved Fake News Video Detection

Yili Li[1], Jian Lang[1], Rongpei Hong[1], Qing Chen[1], Zhangtao Cheng[1], Jia Chen[1], Ting Zhong[1,2], Fan Zhou[1,3*]
[1]University of Electronic Science and Technology of China, Chengdu, Sichuan, China
[2]Aiwen Tech (Chengdu), Chengdu, Sichuan, China
[3]Key Laboratory of Intelligent Digital Media Technology of Sichuan Province, Chengdu, Sichuan, China
lylxzr@uestc.edu.cn, {jian_lang, rongpei.hong, 202312281002, zhangtao.cheng}@std.uestc.edu.cn
{jchen, zhongting, fan.zhou}@uestc.edu.cn

***Abstract*—Detecting fake news videos has emerged as a critical task due to their profound implications in politics, finance, and public health. However, existing methods often fail to distinguish real videos from their subtly manipulated counterparts, resulting in suboptimal performance. To address this limitation, we propose REAL, a novel model-agnostic REtrieval-Augmented prototype-aLignment framework. REAL first introduces an LLM-driven video retriever to identify contextually relevant samples for a given target video. Subsequently, a dual-prototype aligner is carefully developed to model two distinct prototypes: one representing authentic patterns from retrieved real news videos and the other encapsulating manipulation-specific patterns from fake samples. By aligning the target video's representations with its ground-truth prototype while distancing them from the opposing prototype, the aligner captures manipulation-aware representations capable of detecting even subtle video manipulations. Finally, these enriched representations are seamlessly integrated into existing detection models in a plug-and-play manner. Extensive experiments on three benchmarks demonstrate that REAL largely enhances the detection ability of existing methods. The code and data for reproducing the results are available at https://github.com/Jian-Lang/REAL.**

***Index Terms*—fake news video detection, retrieval augmentation, prototype alignment**

I. Introduction

Online video-sharing platforms like TikTok and YouTube Shorts have become increasingly popular on mobile internet and attracted billions of monthly active users. However, the prevalence of news consumption on these video platforms also boosts the rapid spread of malicious content (e.g., fake news) in videos, posing real-world threats to politics, finance, and public health [1], [2]. Therefore, developing effective methods for Fake News Video Detection (FNVD) is urgent to mitigate their negative impact.

Current methods in FNVD primarily focus on modelling multimodal content and capturing cross-modal correlations to assess video authenticity [1]–[4]. Despite their progress, existing works struggle to effectively identify the nuanced differences between real news videos and their subtly manipulated fake counterparts, incurring limited detection performance.

Previous studies [2], [4] have shown that most fake news videos are commonly created by manipulating real news videos, rather than fabricating entirely new content. To effectively deceive viewers, creators of fake news videos often introduce subtle yet crucial alterations to the original content in real videos, such as distorting the narrative, or editing visual, textual, and audio elements to propagate misinformation [4]. For instance, as shown in Figure 1, the manipulated fake news video on the right introduces only minute modifications to the on-screen text and audio, while retaining nearly identical vision content to the original real version on the left. Due to the high similarity between these two videos, the identification of reliable and distinguishing features becomes particularly challenging, leading existing detection methods to incorrectly classify them as the same category.

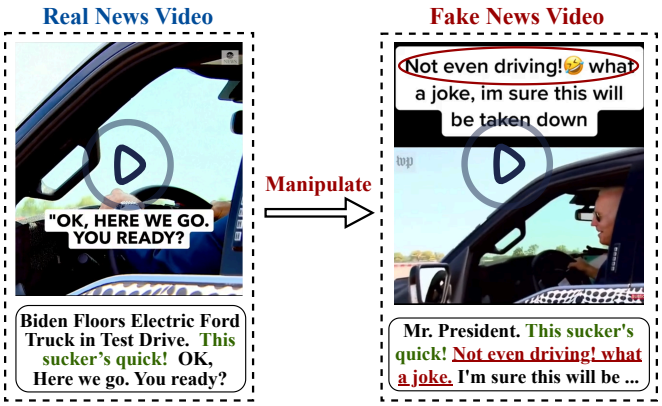


Fig. 1. The left panel shows a real news video of Joe Biden test-driving an electric Ford truck. The right panel illustrates a manipulated version with altered on-screen text and audio, falsely suggesting he was pretending to drive.

To address these challenges, we draw inspiration from human cognitive processes [5]: When confronted with ambiguous or difficult-to-distinguish instances, humans instinctively refer to examples from known categories that are closely related to the target. By contrasting the target with these examples, subtle distinguishing patterns—often imperceptible in isolation—become more apparent and actionable. Building upon this intuition, we propose **REAL**, a novel model-agnostic **RE**trieval-**A**ugmented prototype a**L**ignment framework. Unlike existing methods that analyze videos in isolation, REAL leverages semantically target-relevant real and fake reference samples to guide the learning of manipulation-aware represen-

*Corresponding author.

tations for the target videos. These enhanced representations seamlessly integrate with existing FNVD methods, enabling them to distinguish authentic news videos from their altered fake counterparts.

Inspired by the exceptional ability of large language models (LLMs) [6] in comprehending and organizing information, we propose an LLM-driven video retriever to identify semantically relevant examples from two categories (i.e., real and fake) for a given target video. Specifically, the retriever leverages the LLMs to effectively integrate information from audio, text, and visual modalities of the target video into a unified, text-centric representation, which serves as a powerful query for retrieval. The retrieved real and fake samples form contextually relevant reference sets, facilitating the learning of manipulation-aware representations for the target video.

Building on the retrieved samples, we introduce the dual-prototype aligner to refine the target video's features into manipulation-aware representations. Specifically, the aligner leverages a graph attention network [7] to model two distinct prototypes—one for real and one for fake categories—based on the retrieved instances. These prototypes act as "reference points", reflecting authentic patterns in real news videos and manipulation-specific patterns in fake ones. By aligning the target videos with their ground-truth prototypes while distancing them from the opposing prototypes, the aligner produces manipulation-aware representations that highlight subtle discrepancies in real news videos and manipulated counterparts. Our contributions are summarized as follows:

- We propose a novel model-agnostic REtrieval-Augmented prototype aLignment framework (REAL) that generates manipulation-aware representations to enhance existing methods in detecting fake news videos.
- We introduce a fresh LLM-driven video retriever, which provides contextually relevant reference sets to guide representation enhancement for the target video.
- We design a new dual-prototype aligner, which models real and fake prototypes to amplify manipulation-specific signals in altered fake videos representations and enhance authentic characteristics in original real ones.

Extensive experiments on three real-world video datasets demonstrate the excellent capability of REAL, which enhances existing FNVD methods with an average performance improvement of 3.7%.

II. Related Work

The Fake News Video Detection (FNVD) task involves detecting fake news content by analyzing multimodal data within videos, such as textual descriptions, visual content, and audio information. Early detection methods primarily relied on single-modal information to assess video authenticity [8]–[10]. However, due to the inherently multimodal nature of videos, where text, vision, and audio modalities provide complementary aspects of information to describe the content, single-modal models are inadequate for accurate video-based detection [11]. To address this problem, multimodal learning in the FNVD has garnered broad attention in both the research and industry communities [1]–[4], [12]. SV-FEND [1] captures the multimodal correlations within videos and utilizes the social context to assist fake news detection. NEED [4] integrated both explicit and implicit neighborhood relationships to enhance detection performance.

Despite their advancements, existing FNVD methods struggle to identify the subtle differences between real and manipulated fake videos, leading to severe misclassification in real-world scenarios. To address this challenge, our REAL pioneers a retrieval-augmented prototype alignment strategy to generate manipulation-aware representations. These representations exhibit enhanced distinctiveness between original videos and their manipulated variants and can be seamlessly incorporated into existing detection models to boost their performance.

III. Methodology

In this section, we describe the proposed REAL in detail. The overall framework of REAL is illustrated in Figure 2. We first provide the preliminaries for the FNVD. The subsequent subsections present in-depth descriptions of the key components of REAL: the LLM-Driven Video Retriever and the Dual-Prototype Aligner. Finally, we describe how REAL can be seamlessly integrated into existing FNVD models in a plug-and-play manner.

A. Preliminary

Problem Statement Let $\mathcal{S}$ represent a video collected from online video platforms. The video $\mathcal{S}$ is composed of text, vision, and audio modalities, denoted as $\mathcal{S} = \{\mathcal{T}, \mathcal{V}, \mathcal{A}\}$. The objective of FNVD is to detect whether the video $\mathcal{S}$ is **fake** or **real** by considering its modalities $\mathcal{T}$, $\mathcal{V}$, and $\mathcal{A}$. Notably, since REAL is a retrieval-guided framework, we refer to $\mathcal{S}$ as the target video in the following discussion to ensure precise and clear descriptions.

Feature Extraction Prior FNVD methods adopt various methods to extract modality features. To simplify the discussion, we define the modality-specific features of video $\mathcal{S}$ extracted by existing methods as $\mathbf{E}^m$, $m \in \{t, v, a\}$ for text, vision, and audio modalities, respectively.

B. LLM-Driven Video Retriever

To search for the most semantically relevant video samples for the target video $\mathcal{S}$, we design an LLM-Driven Video Retriever (LDV Retriever), which introduces the exceptional ability of LLMs in information comprehension and organization. The retriever unifies audio, text, and visual modalities from the target video into a shared, text-centric representation that comprehensively represents the video content. This unified representation ensures that information from all three modalities is effectively interpreted and integrated, forming a powerful query for video-to-video retrieval while addressing the inherent heterogeneity between modalities.

Specifically, LDV Retriever first uniformly samples M frames from the target video $\mathcal{S}$ and then employs the pre-trained BLIP [13] to generate captions for each frame, thereby transforming the visual information into textual form, denoted

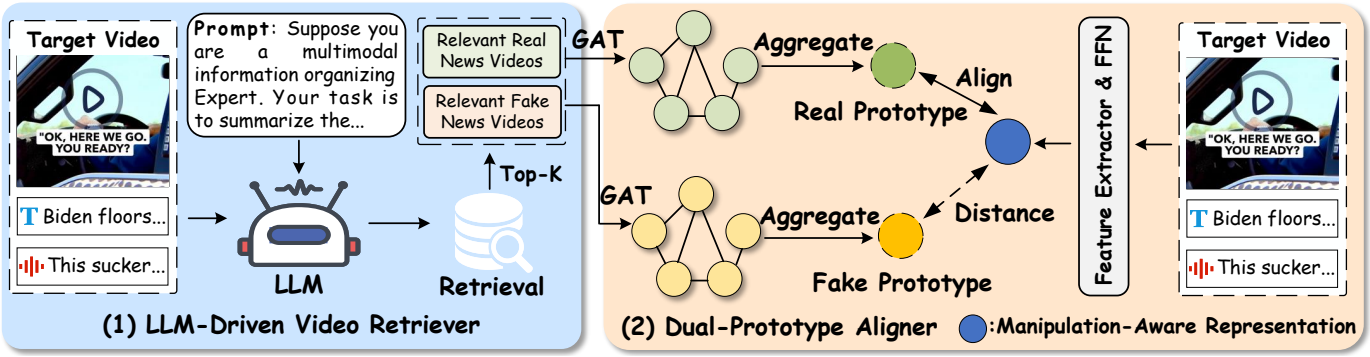


Fig. 2. Overall framework of our proposed REAL. (1) LLM-Driven Video Retriever searches for the most relevant real and fake news videos corresponding to the target real video. (2) Dual-Prototype Aligner leverages the retrieved instances to construct prototypes for real and fake videos using a Graph Attention Network (GAT), aligning the target video representation with its ground-truth prototype while distancing it from the opposing prototype.

as $\mathcal{V}_t = \{\mathcal{V}_t^1, \mathcal{V}_t^2, \cdots, \mathcal{V}_t^M\}$. Second, LDV Retriever leverages Whisper [14] to convert the audio of $\mathcal{S}$ into textual form, denoted as $\mathcal{A}_t$. Subsequently, LDV Retriever utilizes the LLM to integrate information from all three modalities to form the text-centric query $\mathcal{R}$ for retrieval:

$$\mathcal{R} = \mathcal{L}(\mathcal{V}_t, \mathcal{T}, \mathcal{A}_t, \mathcal{P}), \tag{1}$$

where $\mathcal{L}(\cdot)$ represents the LLM, $\mathcal{T}$ is the text modality of $\mathcal{S}$ (e.g., a video's title) and $\mathcal{P}$ is the prompt fed into the LLM, with a concise version shown as:

"Suppose you are a multimodal information organizing expert. Your task is to summarize the information from the visual, textual, and audio content of the given video: [$\mathcal{V}_t$], [$\mathcal{T}$], and [$\mathcal{A}_t$]. Provide a concise and accurate description that effectively represents the video's content."

Based on the query $\mathcal{R}$, LDV Retriever performs the video-to-video retrieval to retrieve the semantically relevant video samples for the target video $\mathcal{S}$:

$$\mathcal{S}^r = \underset{b \in \mathcal{M}}{\text{Top-}K}\left(\frac{\Psi(\mathcal{R})^\top \Psi(\mathcal{R}_b)}{\|\Psi(\mathcal{R})\| \|\Psi(\mathcal{R}_b)\|}\right), \tag{2}$$

where $\Psi(\cdot)$ represents the pre-trained text encoder (i.e., GTE [15]), $\mathcal{M}$ is the memory bank that stores a diverse set of videos, and $\mathcal{S}^r = \{\mathcal{S}_i^r\}_{i=1}^K$ denotes the retrieved top-K videos. We obtain the top-K_f retrieved fake video samples and top-K_r real ones, denoted as $\mathcal{S}^{rf}$ and $\mathcal{S}^{rr}$, respectively. These retrieved videos serve as contextually relevant reference sets, guiding the learning of manipulation-aware representation for the target video $\mathcal{S}$.

C. Dual-Prototype Aligner

To generate manipulation-aware representation for $\mathcal{S}$, we propose the Dual-Prototype Aligner (DP Aligner), which models both real and fake prototypes and performs prototype alignment learning based on the retrieved video samples.

1) Graph-based Dual-Prototype Generation: To provide expressive prototypes as reference points for the target video $\mathcal{S}$, we introduce a Graph Attention Network (GAT) [7] to model authentic patterns in retrieved real news videos and manipulation-specific patterns in fake ones, respectively.

Specifically, given the retrieved fake video set $\mathcal{S}^{rf} = \{\mathcal{S}_i^{rf}\}_{i=1}^K$, DP Aligner constructs a contextual information graph $\mathcal{G}_f = (\mathcal{V}_f, \mathcal{E}_f)$, where each node $\mathbf{v}_i \in \mathcal{V}_f$ represents the modality-specific feature $\mathbf{E}_i^{f,m}$ of the fake video sample $\mathcal{S}_i^{rf}$, where $m \in \{t, v, a\}$. Each edge $e_{ij} \in \mathcal{E}_f$ reflects pairwise relationships between node i and j based on feature semantic similarity. DP Aligner computes the edge weight e_{ij} between two nodes $\mathbf{v}_i$ and $\mathbf{v}_j$ and normalizes the weight:

$$e_{ij} = \text{LeakyReLU}\left(\mathbf{a}^\top [\mathbf{W}\mathbf{v}_i, \mathbf{W}\mathbf{v}_j]\right), \tag{3}$$

$$\alpha_{ij} = \text{Softmax}(e_{ij}) = \frac{\exp(e_{ij})}{\sum_{k \in \mathcal{N}_i} \exp(e_{ik})}, \tag{4}$$

where $\mathbf{W}$ and $\mathbf{a}$ are trainable parameters, $\mathcal{N}_i$ denotes the set of neighbors of node $\mathbf{v}_i$ in the graph. With the attention weights α_{ij}, the updated representation $\hat{\mathbf{v}}_i$ for each node $\mathbf{v}_i$ is computed as a weighted aggregation of its neighbors:

$$\hat{\mathbf{v}}_i = \sigma\Big(\sum_{j \in \mathcal{N}_i} \alpha_{ij} \cdot \mathbf{W}\mathbf{v}_j\Big), \tag{5}$$

where $\sigma(\cdot)$ is a non-linear activation function (e.g., ReLU). After graph-based aggregation, the embeddings of all nodes in $\mathcal{G}_f$ are pooled to form the fake prototype $\mathbf{P}^{f,m}$:

$$\mathbf{P}^{f,m} = \frac{1}{K_f} \sum_{i=1}^{K_f} \hat{\mathbf{v}}_i, \tag{6}$$

Similarly, the real prototype $\mathbf{P}^{r,m}$ can be obtained via the same process.

2) Prototype Alignment Learning: For the target video $\mathcal{S}$, DP Aligner first feeds its modality-specific feature $\mathbf{E}^m$ into the feed-forward network (FFN) to obtain the original manipulation-aware representation:

$$\mathbf{M}^m = \text{FFN}(\mathbf{E}^m) = \mathbf{W}_2 \cdot \sigma(\mathbf{W}_1 \mathbf{E}^m + \mathbf{b}_1) + \mathbf{b}_2. \tag{7}$$

where $\mathbf{M}^m, m \in \{t, v, a\}$ is the manipulation-aware representation. Next, the dual-prototype alignment loss is defined to align the $\mathbf{M}_i^m$ of a batch of N target videos $\mathcal{B} = \{\mathcal{S}_i\}_{i=1}^N$ to their ground-truth category prototypes and distance the representation from the opposing prototypes:

$$L_p = \sum_{i=1}^{N}\Big(\sum_{m \in \{t,v,a\}}\Big(\Big\|\mathbf{M}_i^m - \mathbf{P}_i^{l,m}\Big\| + \frac{|\mathbf{M}_i^{m\top} \mathbf{P}_i^{n,m}|}{\|\mathbf{M}_i^m\| \|\mathbf{P}_i^{n,m}\|}\Big)\Big), \tag{8}$$

TABLE 2
PERFORMANCE COMPARISON OF BASE MODELS WITH AND WITHOUT OUR REAL. THE BETTER RESULTS IN EACH GROUP USING THE SAME BASE MODEL ARE IN **RED BOLD**, AND THE RELATIVE GAIN (%) IS CALCULATED. THE HIGHER VALUES OF ACC, M-F1, AND M-P INDICATE BETTER PERFORMANCE.

Dataset	**FakeSV**						**FakeTT**						**FVC**					
Model	**ACC**		**M-F1**		**M-P**		**ACC**		**M-F1**		**M-P**		**ACC**		**M-F1**		**M-P**	
BERT	76.88	-	76.40	-	76.78	-	63.54	-	63.01	-	65.33	-	67.17	-	65.08	-	66.07	-
+ **REAL**	**79.81**	3.8↑	**79.52**	4.1↑	**79.52**	3.6↑	**69.23**	8.9↑	**67.16**	6.6↑	**66.91**	2.4↑	**69.75**	3.8↑	**68.04**	4.5↑	**69.11**	4.6↑
ViT	70.84	-	70.84	-	72.09	-	64.21	-	63.89	-	67.14	-	84.00	-	84.09	-	84.29	-
+ **REAL**	**76.19**	7.6↑	**75.95**	7.2↑	**75.87**	5.2↑	**71.90**	12.0↑	**70.51**	10.4↑	**70.46**	5.0↑	**85.31**	1.6↑	**85.22**	1.3↑	**85.20**	1.1↑
AST	68.00	-	67.12	-	67.54	-	60.87	-	60.73	-	64.02	-	76.24	-	75.09	-	75.81	-
+ **REAL**	**70.06**	3.0↑	**69.34**	3.3↑	**69.58**	3.0↑	**62.87**	3.3↑	**62.39**	2.7↑	**65.26**	1.9↑	**77.15**	1.2↑	**76.08**	1.3↑	**77.43**	2.1↑
FANVM	75.70	-	75.18	-	75.64	-	72.24	-	70.53	-	70.23	-	81.96	-	81.62	-	81.68	-
+ **REAL**	**79.70**	5.3↑	**79.13**	5.3↑	**79.74**	5.4↑	**74.24**	2.8↑	**72.66**	3.0↑	**72.27**	2.9↑	**85.17**	3.9↑	**84.67**	3.7↑	**85.39**	4.5↑
SV-FEND	77.22	-	76.46	-	77.46	-	71.57	-	70.30	-	70.31	-	89.16	-	89.02	-	88.88	-
+ **REAL**	**80.07**	3.7↑	**79.67**	4.2↑	**79.85**	3.1↑	**74.58**	4.2↑	**72.56**	3.2↑	**72.05**	2.5↑	**91.10**	2.2↑	**91.01**	2.2↑	**90.87**	2.2↑
NEED	82.84	-	82.60	-	83.56	-	70.23	-	67.12	-	66.84	-	90.77	-	90.53	-	90.83	-
+ **REAL**	**84.42**	1.9↑	**84.00**	1.7↑	**84.76**	1.4↑	**72.57**	3.3↑	**69.19**	3.1↑	**69.10**	3.4↑	**91.98**	1.3↑	**91.82**	1.4↑	**91.84**	1.1↑
FakingRec	84.87	-	84.09	-	86.69	-	78.59	-	77.23	-	76.65	-	91.52	-	91.37	-	91.33	-
+ **REAL**	**85.93**	1.2↑	**85.22**	1.3↑	**87.53**	1.0↑	**79.62**	1.3↑	**78.32**	1.4↑	**77.84**	1.6↑	**92.67**	1.3↑	**92.33**	1.1↑	**92.11**	0.9↑

TABLE 1
STATISTICS OF THREE DATASETS.

Dataset	Language	# Rumor	# Truth	# Total	Duration (s)
FakeSV	Chinese	1,810	1,814	3,624	39.88
FakeTT	English	1,172	819	1,991	47.69
FVC	English (Mainly)	1,633	1,131	2,764	87.83

where $\mathbf{P}_i^{l,m}$ represents the ground-truth category prototype of $\mathcal{S}_i$ and $\mathbf{P}_i^{n,m}$ is the opposite prototype of $\mathcal{S}_i$. By minimizing the distance between the target video $\mathcal{S}_i$ and the ground-truth category prototype, while simultaneously reducing the cosine similarity between $\mathcal{S}_i$ and its opposite prototype, the DP Aligner yields the final manipulation-aware representation.

D. Prediction

We integrate the manipulation-aware representation $\mathbf{M}^m$ into existing FNVD methods in a plug-and-play manner (i.e., Residual Connection [16]) to make more precise prediction:

$$\hat{y} = \text{Predictor}(\mathcal{F}(\mathbf{E}^t + \mathbf{M}^t, \mathbf{E}^v + \mathbf{M}^v, \mathbf{E}^a + \mathbf{M}^a)), \quad (9)$$

where $\hat{y}$ is the predicted category for the target video $\mathcal{S}$, $\mathcal{F}(\cdot)$ and $\text{Predictor}(\cdot)$ denote the multimodal fusion network and the prediction network in various FNVD methods, respectively.

Subsequently, the Binary Cross-Entropy loss and the prototype loss are combined to optimize the model's parameters:

$$L_{\text{total}} = \alpha \cdot (\sum_{i=1}^{N} L_{cls}(y_i, \hat{y}_i)) + \beta \cdot L_p, \quad (10)$$

where y_i is the label of $\mathcal{S}_i$, N denotes the batch size, and α and β are parameters to balance the two types of loss. By optimizing the combined loss, REAL improves the ability of existing methods in FNVD.

IV. EXPERIMENTS

A. Experimental Settings

In this section, we provide a summary of the datasets, baselines, evaluation metrics, and implementation details.

Datasets We conducted experiments on three real-world video datasets: FakeSV [1], FakeTT [2], and FVC [17]. Table 1 provides the detailed statistics of three datasets. The splits for each dataset are consistent with the prior works.

Baselines REAL can be extended to any FNVD methods to enhance their prediction. To evaluate its universal efficacy, we select 7 baseline detectors, which are categorized into two groups: (1) *Single Modal Detection Methods*: BERT [8], ViT [9], and AST [10]. (2) *Multimodal Detection Methods*: FANVM [3], SV-FEND [1], NEED [4], and FakingRec [2]. For NEED, we implement it with the base model SV-FEND.

Evaluation Metrics Following prior studies [1], [2], we employ three metrics to evaluate the performance: Accuracy (ACC), Macro F1 score (M-F1), and Macro Precision (M-P).

Implementation Details During the retrieval process, we utilize GPT-4o-mini [6] to organize the multimodal data and employ GTE [15] to generate the retrieval vectors. The memory bank for each dataset is constructed using the corresponding training set. For hyper-parameters, we select K_r and K_l from the set $\{1, 3, 5, 7, 9\}$, while both α and β are fixed at 1.

B. Overall Performance

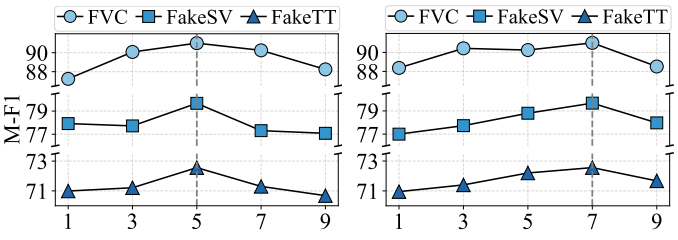
We evaluate the performance of the baseline models without and with REAL in Table 2, and have following observations:

(O1) With the incorporation of REAL, all 7 baseline models exhibit significant performance improvements, achieving gains of 1.2–12% in terms of ACC. These results highlight the effectiveness and versatility of REAL, which enhances the discriminative capability of the baseline models in FNVD by generating more distinctive and expressive representations.

(O2) REAL yields more pronounced improvements on underperforming models. We hypothesize that this phenomenon arises from the susceptibility of these models to misclassify real news videos and their subtly altered counterparts. By leveraging manipulation-aware representations, REAL mitigates this limitation and establishes a more robust lower bound for detection performance in these underperformed methods.

TABLE III
ABLATION STUDY RESULTS OF CORE COMPONENTS WITHIN REAL.

Module	Variant	FakeSV		FakeTT		FVC	
		Acc	M-F1	Acc	M-F1	Acc	M-F1
LDV Retriever	Uni-modal	79.09	78.87	72.57	71.21	89.67	89.54
	w/o LLM	78.96	78.52	73.24	71.26	89.37	89.24
	w/o Retriever	78.59	77.87	70.23	69.23	88.37	88.24
DP Aligner	w/o Real	78.88	78.38	72.57	71.54	89.79	89.70
	w/o Fake	77.49	77.43	72.57	71.61	88.22	88.09
	w/o Graph	78.22	77.71	69.23	68.07	88.07	87.96
REAL	**All**	**80.07**	**79.67**	**74.58**	**72.56**	**91.10**	**91.01**



(a) Effect of K_r. (b) Effect of K_f.

Fig. 3. Hyper-parameter sensitivity analysis of K_r and K_f on three datasets.

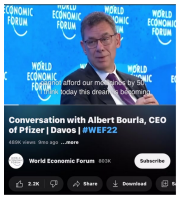
C. Ablation Study

To assess the effectiveness of each core component in REAL, we conduct an ablation study using the base model SV-FEND, with the results presented in Table III.

Effect of LDV Retriever To assess the impact of the LDV Retriever, we design three variants: (1) **Uni-modal**: replacing the LDV Retriever with a uni-modal retriever that leverages only textual information for retrieval, (2) **w/o LLM**: removing the LLM within the LDV Retriever, and (3) **w/o Retriever**: entirely discarding the LDV Retriever while using random samples as substitutes for the retrieved instances. As shown in Table III, the uni-modal variant results in a noticeable performance degradation, which highlights the limitations of using uni-modal information alone for accurate video-video retrieval. When the LLM is removed, the performance also drops significantly, demonstrating the powerful ability of the LLM in summarizing and integrating information. Furthermore, replacing the LDV Retriever with random sampling results in a significant performance decline, underscoring the critical importance of high-quality retrieved instances.

Effect of DP Aligner To validate the effectiveness of the DP Aligner, we design three variants: (1) **w/o Real**: removing the real category prototype, (2) **w/o Fake**: removing the fake category prototype, and (3) **w/o Graph**: replacing the GAT mechanism by simply averaging the retrieved instances to form the prototype. As reported in Table III, removing either the real or fake prototype leads to a significant performance drop, which underscores the critical role of each prototype in guiding the learning of manipulation-aware representations. Furthermore, eliminating the GAT mechanism and using a simple averaging strategy also results in a notable decline in performance. This highlights the pivotal role of the GAT in dynamically aggregating information from retrieved instances.

TABLE IV
PRESENTATION OF THE RETRIEVAL QUALITY. SCORE: SIMILARITY SCORE. RED TEXT HIGHLIGHTS SIMILAR CONTENT.

	Target: Fake	Top-1: Real	Top-1: Fake
Vision			
Audio	January in California. And we will reduce the number of people in the world.	We will reduce the number of people in the world that cannot afford medicines.	The first week we will reduce the number of people in the world.
Text	#World #Forum #Economic	#worldeconomic-forums	#world #freedomofspeech
Score	N/A	0.88	0.94

D. Hyper-Parameter Sensitivity Analysis

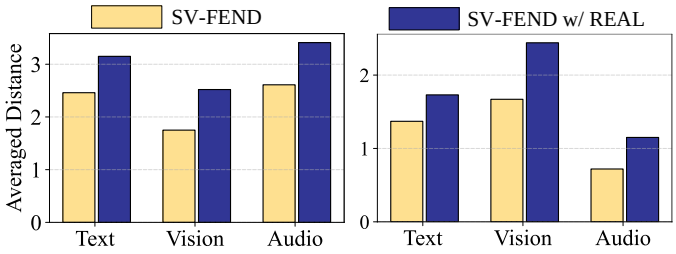
We conduct a sensitivity analysis on two key hyper-parameters of REAL using the base model SV-FEND: the number of retrieved real videos (K_r) and fake videos (K_f). As illustrated in Figure 3, we observe that incorporating retrieved instances consistently improves performance. However, when the number of retrieved instances becomes excessively large, the performance begins to degrade. This decline is primarily attributed to the introduction of noise, such as irrelevant or low-quality samples, which weakens the quality of the prototypes. Consequently, the optimal performance is achieved when $K_r = 5$ and $K_f = 7$ across all three datasets.

E. Retrieval Quality Visualization

To further validate the effectiveness of the proposed LDV Retriever, we randomly select a fake news video from the FakeSV dataset and analyze its retrieved top-1 real and fake news videos. As shown in Table IV, the retrieved real and fake videos demonstrate high semantic relevance to the target video across all modalities. This observation highlights the capability of the LDV Retriever to accurately identify contextually similar instances for the target video.

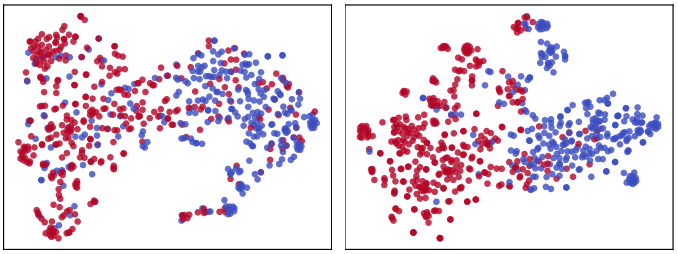
F. Further Analysis on Manipulation-Aware Representation

To further investigate the efficacy of manipulation-aware representations in distinguishing real news videos from their manipulated counterparts, we conduct a quantitative analysis across different modalities. Specifically, we randomly select 50 video pairs from each of the FakeSV and FakeTT datasets, where each pair comprises an authentic news video and its manipulated version. For each pair, we calculate the feature distance between the two videos in three modalities using both the original feature space and the manipulation-aware feature space on the base model SV-FEND. The average distances for each modality are computed and summarized in Figure 4.



(a) FakeSV (b) FakeTT

Fig. 4. Feature distance between real news videos and their manipulated versions on the FakeSV and FakeTT datasets.



(a) SV-FEND (b) SV-FEND w/ REAL

Fig. 5. T-SNE visualization of model SV-FEND and with the help of REAL. Red points indicate fake news videos while blue points represent real ones.

From the results, we can draw the conclusion: in the original feature space, the feature distances between real and manipulated videos are relatively small, indicating limited discriminative power. In contrast, the manipulation-aware representations exhibit significantly larger feature distances, effectively amplifying the discrepancy between the real and manipulated videos, and facilitating more precise detection.

G. Visualization

Figure 5 presents a t-SNE [18] visualization of the embedding distributions for the two categories on the test set of the FVC dataset. We visualize the output embeddings from the last layer of the classifier in both the original base model SV-FEND and SV-FEND enhanced with REAL. The visualization reveals that the integration of REAL enables SV-FEND to generate more discriminative representations, resulting in clearer class boundaries compared to the original one.

V. Conclusion

In this work, we present REAL, a novel, model-agnostic retrieval-augmented prototype alignment framework for enhancing the performance of existing methods for FNVD. REAL introduces two key components: (1) an LLM-driven video retriever that identifies the most semantically relevant video instances for a given target video, and (2) a dual-prototype aligner that constructs two distinct prototypes: one modeling authentic patterns in retrieved real news videos and another summarizing manipulation-specific characteristics from fake samples. By aligning the target video with its ground-truth prototype while distancing it from the opposing prototype, REAL effectively captures manipulation-aware representations, thereby enhancing the ability of existing FNVD methods to differentiate real news videos from their manipulated counterparts. Extensive experiments conducted on three real-world video datasets validate the effectiveness of REAL.

Acknowledgments

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

References

[1] Peng Qi, Yuyan Bu, Juan Cao, Wei Ji, Ruihao Shui, Junbin Xiao, Danding Wang, and Tat-Seng Chua, "Fakesv: A multimodal benchmark with rich social context for fake news detection on short video platforms," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2023, vol. 37, pp. 14444–14452.

[2] Yuyan Bu, Qiang Sheng, Juan Cao, Peng Qi, Danding Wang, and Jintao Li, "Fakingrecipe: Detecting fake news on short video platforms from the perspective of creative process," in *Proceedings of the ACM International Conference on Multimedia (MM)*, 2024, pp. 1351–1360.

[3] Hyewon Choi and Youngjoong Ko, "Using topic modeling and adversarial neural networks for fake news video detection," in *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM)*, 2021, pp. 2950–2954.

[4] Peng Qi, Yuyang Zhao, Yufeng Shen, Wei Ji, Juan Cao, and Tat-Seng Chua, "Two heads are better than one: Improving fake news video detection by correlating with neighbors," in *Findings of the Association for Computational Linguistics*, 2023, pp. 11947–11959.

[5] Douglas R Hofstadter, *Fluid concepts and creative analogies: Computer models of the fundamental mechanisms of thought.*, Basic books, 1995.

[6] OpenAI, "Gpt-4 Technical Report," *arXiv*, 2023.

[7] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio, "Graph attention networks," in *International Conference on Learning Representations (ICLR)*, 2018.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the Conference of the North American chapter of the Association for Computational Linguistics (NAACL)*, 2019, pp. 4171–4186.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations (ICLR)*, 2020.

[10] Yuan Gong, Yu-An Chung, and James Glass, "Ast: Audio spectrogram transformer," *arXiv preprint arXiv:2104.01778*, 2021.

[11] Jian Lang, Rongpei Hong, Jin Xu, Yili Li, Xovee Xu, and Fan Zhou, "Biting off more than you can detect: Retrieval-augmented multimodal experts for short video hate detection," in *The Web Conference (WWW)*. ACM, 2025.

[12] Rongpei Hong, Jian Lang, Jin Xu, Zhangtao Cheng, Ting Zhong, and Fan Zhou, "Following clues, approaching the truth: Explainable micro-video rumor detection via chain-of-thought reasoning," in *The Web Conference (WWW)*. ACM, 2025.

[13] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International Conference on Machine Learning (ICML)*. PMLR, 2023, pp. 19730–19742.

[14] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning (ICML)*. PMLR, 2023, pp. 28492–28518.

[15] Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang, "Towards general text embeddings with multi-stage contrastive learning," *arXiv preprint arXiv:2308.03281*, 2023.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[17] Olga Papadopoulou, Markos Zampoglou, Symeon Papadopoulos, and Ioannis Kompatsiaris, "A corpus of debunked and verified user-generated videos," *Online information review*, vol. 43, no. 1, pp. 72–88, 2019.

[18] Laurens Van der Maaten and Geoffrey Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.