Journal of Education, Teaching and Social Studies
ISSN 2642-2336 (Print) ISSN 2642-2328 (Online)
Vol. 8, No. 1, 2026
www.scholink.org/ojs/index.php/jetss

*Original Paper*

# DynaLLM-C3N: Dynamic Fusion of Cross-Modal Correlation and Large Language Model Reasoning for Multimodal Fake News Detection

Yuxin Zhao[1] & Jize Li[1]

[1] College of Public Finance and Administration, Harbin University of Commerce, Harbin, Heilongjiang, 150028, China

Received: February 17, 2026 Accepted: March 19, 2026 Online Published: March 24, 2026

doi:10.22158/jetss.v8n1p166 URL: http://dx.doi.org/10.22158/jetss.v8n1p166

***Abstract***

*The growth of false information on social media suggests a need for better ways to spot it. However, current methods are not good at detecting differences between text and images. In fact, they seem to suggest that deep-level problems may be holding back progress in this area. To address these issues, we have developed DynaLLM-C3N. It is a new system that uses language models to detect false information and link it to other data sources. Research shows it uses C3N to detect links between images and text. It can also use language models to verify information and generate new features. Our findings show that it can do all of this better than other methods. It can also use a special process to give different types of data the same importance. We have tested it on the Twitter and Weibo datasets and have found that it is better than current methods. The study shows that dynamic fusion balances the complementary strengths of the two approaches. However, the most significant results suggest that the dynamic fusion process may successfully balance the strengths of data-driven correlation modelling and symbolic reasoning. This suggests that better detection accuracy is possible. However, the evidence suggests that enhanced interpretability could indicate the framework's broader applicability across different detection contexts. This suggests that the source code could be made available to the public when it is published, which would help more people to use these key findings.*

***Keywords***

*Fake news detection, Multimodal learning, Large language models, Cross-modal correlation, Dynamic fusion, Interpretable AI*

## 1. Introduction

The thing is, there's a lot more of this multimodal misinformation on digital platforms these days, and that's a big problem for democratic processes, public health initiatives, and social cohesion [1][2]. It's not like textual misinformation, where you just read something and think, "Hmm, that's not right", you know? With multimodal misleading content, they use text and visuals together to make what they're saying seem more believable and to avoid getting caught. Social media has become so important that it's made it harder to filter information, so now it's up to us to check if what we're seeing is true or not, and that's a big risk for society. All this fake news is a big problem for democratic institutions, social justice, and public trust, so we need to come up with ways to spot it and stop it [5]. Here are some examples: Dissemination of disinformation about novel coronavirus (COVID-19) has resulted in extreme worry among the general public: During the 2016 US Presidential Election, there was a very short period of fake news. Recently, people have tried solving it by using multiple modalities together but those aren't very good either. It depends very much on the global feature similarities of different modalities, it can't find subtler differences about things/words. The second problem is no high order reasoning so claims cant be checked against world knowledge or logic. Thirdly, static fusions can't give different weights to various info sources according to the context.

In this work we introduce DynaLLM-C3N, which is a new framework to solve all the problems by adding some 3 major changes. We first improve upon the Cross-Modality Content Correlation network (C3N) [11] to find out weak links among picture sections and words, dealing with big similarities that do not show much difference. Second is to add on the multimodal LLM reasoning and generate the verified traits via fact-checking. It is more than just 0/1 classification and gives a reason. And finally we suggest an HDM which learns to aggregate these various types of information according to how useful it is for each individual piece of news.

We move away from the traditional method of using multimodal fusion for our detection of fake news. Instead of thinking of text and image as the same kind of input that can be combined in the same way, we see that different news items need different combinations of low-level correlation features and high-level reasoning capabilities. Detection of manipulated celebrity pictures is helped more from fine visual-textual alignment analysis and the discovery of fabricate political claim need fact to be verified more deeply. We will find our dynamic fusion the best way to utilize.

The primary contributions of this work are as follows:

We are suggesting a term named: DynaLLM-C3N which means using new ways to combine cross-modal content correlation modeling and something we call structured reasoning in large language models. And we want to leverage all of this for the multimodal fake news.

We've designed this multimodal LLM reasoning module, right? And what it does is generate all these verification features, and it does it through these fact-checking prompts. So, it's like, it extracts numerical scores and semantic evidence, and it does it all in a structured format.

We have developed a dynamic gating mechanism that adapts to multiple information sources. This mechanism is organised into separate networks for selecting modalities and weighing reasoning dimensions.

We've done loads of experiments on two real-world datasets, and the results speak for themselves. We've also got some detailed studies to show how we designed the system, and they're really impressive.

## 2. Related Work

Early work mainly concentrated on unimodal approaches, especially those based solely on texts; however, it is still being used at present. The stylometric traits combined along with word vector gave an ensemble which was able to accomplish at 95.49% success in seeking out fake news, verified the value of text attribute [12]. Similar experiments pulled apart numerous textual qualities and applied a mixture technique. Though there is hope even on purely texts basis [13], yet that also leaves something under its surface. Models can get over fitted because of too dependence upon certain cases. Therefore it is about the fusion of multiple modes now. It's an attempt at bringing in more info from things like texts and pictures to improve how well we can find stuff. MPFN gets a 83.3% correct in things like Twitter with it's step by step feature extraction and fusion steps to show how well cross model correlation works [14]. Within and outside of the Modality is emphasized by us through IFIS, and hence we have a better result [15]; Our proposed methods which are modality dependent are created on content alignment; Therefore it uses CLIP for extracting features along with wrap - arround the cross- Modal interactions using convolution. Even though there's remarkable performance shown by Large language Models most times around cross-modal tasks they don't get much use compared to other areas such as Fake new detection [16] So those models become pretty close to some of the smaller models which were fine tuned on a few shots, but there's still room for improvement. A fresh way to detect false information using many different ways like looking at it and reading it, is now made for big computer brains to find mixed-up news. This new idea puts guesses about what's happening in small computer thoughts. By testing with different public dataset we can see how much this structure improves the model for identifying false information [17]. Quantum Multimodal Fake News Identification Model (QMFND) based on Quantum Convolutional Neural Network is proposed. This model, it's like taking the words and pictures and putting them in some quantum state, alright? And then it's using like a quantum conv net for doing the classification. The QMFND has proven it can get up to 87.9% and 84.6% accurate on various datasets. The rate is achieved by this model is more than the rate that has been achieved by all the models which are compared and also showing different advantage regarding the Quantum expressibility and robustness [18].

## 3. Methodology

### *3.1 Problem Formulation*

When we're dealing with a multimodal news post, like N= (T, I) with some text T and an image I, what

we're trying to figure out is how genuine it is. We'll use the label y = {0, 1} L, where y = 1 is for real news and y = 0 is for fake news. The DynaLLM-C3N framework addresses this challenge through three integrated components: First, there's an enhanced Cross-modal Content Correlation Network, then a Multimodal LLM Reasoning Module, and finally, a Hierarchical Dynamic Fusion Mechanism. Figure 1 illustrates the overall architecture of our proposed framework, which is outlined in more detail below.

*3.2 Word and Image-crop Feature Extraction*

Map text and visual information from cross-modal similarity calculations into a unified semantic space. The aim is integration and ensuring the balance and accuracy of multimodal information [19]. The paper's employment of a pre-trained CLIP encoder is for the purpose of leveraging its cross-modal alignment capability to obtain shared semantic features. For the visual modality, extraction of visual content is via object models detected through prior rules; for the textual modality, extraction of nouns and phrases within sentences as semantic content. Subsequent use of the same CLIP encoder is for mapping of textual and visual content into a shared embedding space. Modelling cross-modal associations is achieved by calculating the similarity between their embeddings, which are used to create a representation of the data.

*3.3 Enhanced Cross-modal Content Correlation Network*

In scenarios where fake news is being identified in a variety of forms, although characteristics produced by the CLIP encoder demonstrate considerable transferability, pictures frequently contain intricate backgrounds and semantically unrelated interference. This makes it hard to capture cross-modal shared information by calculating cosine similarity. So, to learn more discriminative association representations, after extracting textual nouns and image region features, we introduce a co-attention transformer module for cross-modal enhancement. As shown in Figure 2, this module adopts a symmetric dual-path attention design. The left path uses text embeddings as queries. It uses image features as keys and values. It performs weighted aggregation of visual information. This is guided by textual semantics. The optimal trajectory employs image features that are symmetrical as inquiries, with text embeddings as keys and values, and thus reconstitutes textual characteristics guided by semantic information. This bidirectional attention mechanism is based on adaptively attaining cross-modal semantic projection and enhancement through content similarity. This process makes it so that shared features between questions are enhanced and semantic gaps between different types of data are narrowed, which then makes it easier to calculate similarity later on. The procedure is as follows:

$$\begin{aligned} MultiHead(Q,K,V)&=(\mathrm{h}ead_1 \oplus \cdots \oplus \mathrm{h}ead_{\mathrm{h}})W^O, \\ Q_i&=QW_i^Q, K_i=KW_i^K, V_i=VW_i^V, \\ \mathrm{h}ead_i&=Att(Q_i,K_i,V_i)=softmax\left(\frac{Q_iK_i^T}{\sqrt{d_{\mathrm{h}}}}\right)V_i. \end{aligned} \tag{1}$$

So, among these, the vector concatenation operation is denoted by $\oplus$, right? We obtain the query (Q), key (K), and value (V) matrices by stacking text features ET or image segment features EI, each with dimensions RL × d, where L represents the length of text or visual data. $W_i^Q$, $W_i^k$, Wiv Eingang Rd × dh The trainable projection matrix corresponding to the i-th head, which maps the input to the query, key,

and value spaces, where the model dimension is dmodel = 512. Note that the number of heads is h = 8, with each head dimension being $d_h$ = dmodel/h = 64. A linear projection matrix, $W^O$ := R^$hd_h$×d, maps the concatenated outputs of all attention heads back to the model dimension. The subsequent feedforward network consists of two linear layers with intermediate ReLU activations, expressed as follows:

$$\mathrm{FFN}(X)=max(0,XW_1+b_1)W_2+b_2. \quad (2)$$

In the CT module, the input feature X, which belongs to the RL × dmodel space, is processed sequentially through multi-head attention and feedforward network sublayers. Each sublayer uses residual connections and layer normalisation operations. This module is cascaded S times. The final output enhancement results are constituted by the text feature ET and the image feature EI, with both maintaining the same dimensionality as the input feature.

*3.4 Multimodal LLM Reasoning Module*

This paper gives out a framework of cross-modal correlation on different levels. The texts & visuals clip of being encoded, and we got this share one through coattention Transformer to share the same semantics. And after that the framework looks at both macro and micro levels: macro is the world around us in all over the entire scene and complete pictures; micro means it will look into every single unit that matches up thoroughly.

We then took all the obtained features to make an ordered (organized) Cross-Modality similarity matrix per sample. Cosine similarity is calculated systematically to do this. The grid is (N + 1) × (M + 1). N, M represent the amount of diminutive groups of nouns and image parts respectively. In it, (0, 0) indicates how much the text matches with the picture, row 0 shows how much every image part in the text fits with it, column 0 gives the same information but for every noun from the text matched against the image, all remaining places (i, j), where i >= 1 and j >= 1 represent how well the ith noun match with the jth image part. This matrix is the characterization of strengths, patterns of associations among different content modalities. They cover various levels of semantics.

In C3N，it use text-centric mode to get the cross-Modality correlation feature: In this method every piece of text serves as a question asking what part of the image does it match, since text has mostly full info that can be used for detecting fake news [20]. This matrix goes into another multi-channel 1D convolution network with filter going across the sequence of texts. The model uses convolutions of different widths to see both the same and not same things between each text bit and all picture bits close by. Whereas it's a simple scalar that just says "here's how close of a match", this gives a kind of soup, where you have both correlations and contradictions. Paper introduces an approach to detect things inside pictures and also predict our belief regarding those detections. And this is the confidence which is used as prior to focus the convolution on more confident/salient parts. The image patch is sorted according to its confidence score from highest to lowest and then calculates the mean association strength with all other text units. It means we need to line up filter width with visual seq length and then do avg pooling along dim after conv.

We propose a nonlinear conv net after the fine-graded sim calc to learn associated feat from many scalars. Given the augmented text feature sequence, T=[e_1T,…,e_NT], and the image feature sequence, I= [e_1I,...,e_MI], the C3N network constructs a matrix out of each text-image pair features first and then calculates their L2-normalized inner product which is the cosine similarity.

$$\hat{e}_i^x = \frac{\tilde{e}_i^x}{\| \tilde{e}_i^x \|_2}, s_{ij} = \hat{e}_i^w \cdot \hat{e}_j^c. \quad (3)$$

Let S ∈ RN × M be the cosine similarity matrix. N×M Convoltor use a sliding wind of dimension h × M over the Si:i+h-1 to get cross-moldel correlation pattern from local area. And then this is how it goes:

:

$$c_i = max(0, w \cdots_{i:i+h-1} + b). \quad (4)$$

Let S∈NXM to denote the cross-modal cosine similarity matrix where N is the no. of text unit and M is the no. of image regions. The convolution uses a filter of size h × M that goes over the text dimension on matrix S. For every postion i, it does a convolution along the submatrix S(ii+h−1,1:M) and creates some new feature representation which has local cross–modality correlation pattern. Then it is expressed as follows that:

$$C_i = [c_1, c_2, \dots, c_{N-h+1}]. \quad (5)$$

A global average pool is done on every feature map C ∈ R(N-h+1) from the convolutions and then it is compressed to just one scalar value with C3N. The final cross-modal correlation feature vector is formed by taking the concatenation of the pooled outputs from all distinct convolutional kernels. This is how it would look if we expressed it more formally:

$$\tilde{C}_i = \text{average}(C_i), f_c = \tilde{C}_1 \oplus \tilde{C}_2 \oplus \cdots \oplus \tilde{C}_l. \quad (6)$$

Obtain cross-modal correlation features, $f_C$ = RL, where l = k × o, k denotes the number of filters of different sizes, and o represents the output channels corresponding to each size.

It captures the fine grained semantic relation between the words and pictures. In order to learn richer, the cross modal association is directly from the date by suppressing out all background noises on our observed object with CLIPS capabilities of Multi-modal semantics. So as to bridge the semantic gap, it overcomes the drawback of scalar similarity. At a fine level, learn the saliency first and then use it to capture correlation consistency/inconsistency between text -image. So those kind of connect up with the label that just says whether or not a news article is true when you're trying to train it.

*3.5 Hierarchical Dynamic Fusion Mechanism*

After cross-modal enhancement, the sequence features of $\tilde{E}^T$ and $\tilde{E}^I$ keep their high-dimensional sequence characteristics. This makes sure that the information is kept intact for similarity calculations that are done later. We take vectors from several positions in the sequences of $\tilde{E}^T$ and $\tilde{E}^I$, and we put these into groups. We also put the fused multimodal features and cross-modal correlation features $f_C$ into groups. These vectors mix fine-grained information from another way of sensing through a shared

attention mechanism. At the same time, they keep the global meaning of the original way of sensing. Next, the text global vector, image global vector, and $f_C$ input are put into groups based on their dimensions. These groups are then put into nonlinear mapping through three independent multilayer perceptrons. The resulting low-dimensional discriminative representations are:

$$f_T = \mathrm{MLP}\left(\tilde{E}^T[0]\right), f_I = \mathrm{MLP}\left(\tilde{E}^I[0]\right), f_C = \mathrm{MLP}(f_c). \quad (7)$$

Let$f_T$ , $f_I$ and $f_c$ denote the processed text, image and cross-modal correlation features, respectively. The mapping of $\tilde{E}^T[0]$ and $\tilde{E}^I[0]$ is to $f_T$ and $f_I$ through independent multilayer perceptrons. The use of ReLU activation functions between layers with a 0.5 dropout rate is for the mitigation of overfitting, since each MLP contains three fully connected layers. The three feature vectors are then concatenated and fed into a fully connected layer for fusion, resulting in the following representation:

$$x=\mathrm{MLP}\left(f_T \oplus f_I \oplus f_C\right). \quad (8)$$

*3.6 Classification*

After obtaining the final expression X, the use of a linear projection layer is for the mapping of it to a two-dimensional space and the normalisation of it via a softmax function, with the output being the predicted probabilities that the sample belongs to fake news and real news respectively:

$$\hat{y}=\mathrm{Softmax}(W_x x+b_x). \quad (9)$$

So, when we're doing the classification stage, the model is basically predicting whether a sample is real news or fake news. So, it's like, if it's real news, it's y1, and if it's fake news, it's y2. So, you can work out these probabilities by using something called the 'softmax function' on the linear result $Wx^X+b$. Now, W and b represent the trainable weight matrix and bias term, respectively. The objective of training the model is to minimise the cross-entropy loss, which is expressed as follows:

$$L(\mathbf{\Theta})=-y\log\left(y_1\right)-(1-y)\log\left(y_0\right). \quad (10)$$

While training the model, the actual labels of the examples are shown as y {0,1}, where y=1 indicates false information and y=0 indicates true information. This signifies the entire collection of all the learnable parameters within the model.

## 4. Experiments

*4.1 Experimental Setup*

4.1.1 Datasets

We evaluated the proposed DynaLLM-C3N method on two publicly available, real-world, multimodal datasets: the English Twitter dataset and the Chinese Weibo dataset. The Twitter dataset includes first-person posts with words, pictures, videos, and social background data [21]. In the case of the Weibo dataset, the veracity of news items was ascertained by Xinhua News Agency, China's preeminent news authority, while the authenticity of disinformation was corroborated by Weibo's formal system for the debunking of rumours [19]. In order to concentrate on text and image modalities, we implemented the following preprocessing steps: for Twitter data, we filtered out tweets containing videos or lacking text/images; for Weibo data, we eliminated duplicate images. Final statistics: The Twitter dataset: 8,199

items of fake news, 6,681 items of real news, and 512 shared images. The Weibo dataset: 4,211 items of fake news, 3,639 items of real news, and 7,850 images. To make sure that fair comparisons could be made, the same data partitioning scheme was adopted as that used in the relevant benchmark studies.

4.1.2 Comparison Methods

We compare our proposed model DynaLLM-C3N against a variety of state-of-the-art baseline methods. These include both unimodal and multimodal approaches. The specific selections are as follows:

Unimodal detection methods include:

BERT [22]: It's a deep pre-trained language model based on Transformers, and it uses deep contextual semantics to perform text classification.

dEFEND [23]: It is a way to spot fake news early on. It uses news stories and comments from users to see how they are connected. It represents a detection approach grounded in textual and social context.

BiLSTM-Attention [24]: The employment of a bidirectional long short-term memory network is a feature of this method for the extraction of text sequence features, with incorporation of an attention mechanism for the focusing of key words for sequence modeling in text detection.

ResNet and VGG19 [25]: Two classic deep convolutional neural networks are used to extract visual features from news images for the purpose of authenticity classification. These networks serve as visual-content-based detection baselines.

Multimodal detection methods include

attRNN [26]. An attention-based recurrent neural network model that learns text and image features.

MVAE [27]: It is a generative model based on multimodal variational autoencoders that learns shared representations by reconstructing input multimodal data.

MKN [28]: It is a knowledge-aware multimodal network that attempts to enhance semantic associations between modalities by incorporating external knowledge.

SAFE [29]: It is a method. It focuses on fine-grained semantic correlations between text and images. It detects these through object-entity matching.

SpotFaka [30]: Identifies fake news by analysing cross-modal consistency and inconsistency patterns.

LIIMR [31]: Cross-modal relationships are detected by models through hierarchical interaction and memory reconstruction networks.

EM-FEND [32]: An event memory-enhanced multimodal fake news detection method that is designed to improve generalisation capabilities for unknown events.

MMFN [33]: A network that integrates information from different sources and feature extractors using a gating mechanism. This is a multimodal fusion network.

FCINet [34]: The focus of a cross-modal incoherence network is the capture of inconsistencies between text and images.

*4.2 Overall Performance*

**Table 1. Test Results on the Dataset**

| Methods | | Acc | Prec | Recall | F1 |
|---|---|---|---|---|---|
| | BERT | 84.2 | 83.6 | 82.9 | 83.2 |
| | dEFEND | 85.7 | 84.9 | 86.3 | 85.6 |
| Unimodal | BiLSTM-Attention | 83.5 | 82.8 | 84.1 | 83.4 |
| | ResNet | 79.8 | 78.4 | 81.2 | 79.8 |
| | VGG19 | 80.3 | 79.5 | 81 | 80.2 |
| | attRNN- | 85.1 | 84.3 | 85.8 | 85 |
| | MVAE | 86.4 | 85.9 | 86.8 | 86.3 |
| | MKN | 87.2 | 86.7 | 87.6 | 87.1 |
| | SAFE | 88.6 | 87.9 | 89.2 | 88.5 |
| Multimodal | SpotFaka | 89.3 | 88.7 | 89.8 | 89.2 |
| | LIIMR | 90.7 | 90.1 | 91.2 | 90.6 |
| | EM-FEND | 91.8 | 91.3 | 92.2 | 91.7 |
| | MMFN | 93.5 | 93.1 | 93.8 | 93.4 |
| | FCINet | 92.9 | 92.5 | 93.2 | 92.8 |
| | DynaLLM-C3N | 95.4 | 95.1 | 95.6 | 95.3 |

To test the effectiveness of the proposed DynaLLM - C 3N model completely, it is compared with many stateoftheart unimodal and multimodal baseline models under the same condition of experiment. From table 1 we can see the evaluation metrics of every model on the test set:accuracy, precision, recall and F1-score. According to table 1, the proposed DynaLLM-C3N plan performs better than any other plans that we have seen so far. It gets an accuracy of F1 score both at 95.4%, which is much more than every method we compared with it. DynaLLM - C3N improves over the best performing multi - modal baselines, MMFN by 1.9% in terms of Accuracy. We show that we are able to capture this type of complex cross- modality inconsistency in fake news. That's kind of a synergized, if you will it be like more intertwined cross-modal relationship on one hand and combining that with LLM’s reasoning capability but combine them in this flexible way by just flipping the switch. From the experimental data we can see that multimodal detection is usually better than single model detection. This shows the necessity to fuse different kinds of information for detecting fake news. Text-based models do better than purely visual ones at determining whether something is real news, which means text probably matters quite a bit for figuring that out. And even the best single-modal tech is way worse than a multimodal one – so clearly you need visual data. And take dEFEND, which is top single-mode tech: It is 7.8% worse than the current multimodal approach MMFN on the F1 score.

The performance of the multimodal baseline models is gradually getting better and better. Okay well, those early ones that tried to do things like add up the simple bits or use attention seemed quite restricted:

Okay, well now I'm checking out some other things. But in the last few years, those with memory or dynamic fusion have improved even more. Using the novel combination of representation reason and the use of large language model together with a hierarchal dynamical info selection process via our suggested DynaLLM-C3N model it represents an advancement of such. And so it leads to this framework being the best around. It indicates that in order to enhance the efficiency of multimodal fake news, we need besides adopt a data driven Relevance modelling technique and validate with higher level knowledge, also integrate with contextual adaptive decisions fusion.

*4.3 Ablation Study*

We're noticing our slow but getting better multi-model baseline model. Earlier ways that just put things together or used special attention could do little, so we try different better ones now. In fact, in recent years memory based models and dynamic fusion type models are getting better still. Innovative combination of the symbolic reason ability of large language model with hierarchical dynamic information selection is achieved through our proposed DynaLLM-C3N framework which is another major development on the path of evolution of large language models. This shows that, as well as using data to work out how important something is, it is also very important to check that the information is correct and to make decisions based on the situation. We did experiments on the Twitter and Weibo dataset to see how well each part of our DynaLLM-C3N framework worked. The results are presented in Table 4. Removing or replacing any component led to performance degradation, thus validating the necessity of each module in our framework.

**Table 2. The Results of Ablation Experiments on the Datasets**

| **Model** | **Fake** | | | **Real** | | |
|---|---|---|---|---|---|---|
| | Pre. | Rec. | F1 | Pre. | Rec. | F1 |
| DynaLLM-C3N | 0.971 | 0.965 | 0.968 | 0.965 | 0.971 | 0.968 |
| DynaLLM-C3N w/o Dynamic | 0.962 | 0.956 | 0.959 | 0.956 | 0.962 | 0.959 |
| DynaLLM-C3N w/o Enhancement | 0.952 | 0.945 | 0.948 | 0.944 | 0.951 | 0.947 |
| DynaLLM-C3N w/o Correlation | 0.957 | 0.956 | 0.956 | 0.953 | 0.954 | 0.953 |
| DynaLLM-C3N w/o Fusion | 0.948 | 0.953 | 0.950 | 0.950 | 0.945 | 0.947 |
| DynaLLM-C3N w/o Entity | 0.953 | 0.953 | 0.953 | 0.951 | 0.951 | 0.951 |
| DynaLLM-C3N-static | 0.960 | 0.954 | 0.957 | 0.954 | 0.960 | 0.957 |
| DynaLLM-C3N-base (C3N) | 0.964 | 0.953 | 0.958 | 0.951 | 0.962 | 0.956 |

The DynaLLM-C3N model did really well, with 96.8% accuracy, which is 1.1% The main driver behind this improvement is our improved dynamic language adaptation that can now deal more effectively with the features of Chinese social media text. Dropping 0.9 percentage points upon removal of dynamic

adaptation component (DynaLLM- C3N w/o Dynamic) from the system, it's easy to see how much this part matters for dealing with Chinese text full of colloquial expression and internet slang.

From the ablation results, we can see that every component in DynaLLM - C3N is necessary for good performance. The parts which provide the biggest improvement of the capability to spot untrue info on the data set is the part for dynamic language adjustment and intermodal connection.

## 5. Discussion

DynaLLM-C3N produces good results on both the Twitter and Weibo dataset which is a strong point about global education. In order to effectively promote information literacy, framework for identifying misinformation need to take into account the different languages and cultures.

It can also be used as an culturally aware pedagogy. Then we can have teachers use that as information for creating comparisons about analyzing social media in western or eastern places, comparing the misinformation tactics of political memes and health claims. And students could consider how misinformation appears in different cultures.

When introducing our model to IL curricula, we need to prioritize metacognition over giving it a final verdict of "this is correct". Teachers can encourage the asking of questions from their students and promote critical thinking, which are elements of UNESCO's media and information literacy curriculum. Visualising differences, the vague idea of 'critical thinking' becomes tangible and talkable.

Collaborate with educators from different linguistic community to make such tools context aware, and effective for educating the people on misinformation. And it takes nimble, culturally attuned thinkers.

## 6. Conclusion

Proposed: DynaLLM-C3N, a new idea for detecting fake info using different ways of media through special cross-format-content connections, many LLM thoughts, and levels that change over time. We spot the differences and similarities between text and pictures. At the level of both small and big And it does not depend on pre-scalar similarities. Also does not need additional information.

On 2 Real-world datasets, it can be seen that DynaLLM-C3N > latest. In the Weibo overall accuracy is up 3.1%; on Twitter it's 1.9%: An ablation to see what is necessary, and qualitative case studies to show how it would work.

Detecion of fake news in many forms is a problem we focus on, such as missing cross-modality relationships and the use of general features. To identify false information by means of LLMs for understanding and use a hierarchical fusion approach, we offer a more complete and useful way.

The next phase of development will involve expanding our framework to include external knowledge such as factual evidence and user comments to improve the model's capability to identify fake news. We also intend to apply our framework to other multimodal tasks.

**References**

[1] I. Slaughter, A. Peytavin, J. Ugander, & M. Saveski. (2025). Community notes reduce engagement with and diffusion of false information online. *Proc. Natl. Acad. Sci. U.S.A.*, *122*(38), e2503413122. https://doi.org/10.1073/pnas.2503413122

[2] P. Gwiaździński et al. (2023). Psychological interventions countering misinformation in social media: A scoping review. *Front. Psychiatry*, *13*, 974782. https://doi.org/10.3389/fpsyt.2022.974782

[3] O. Razuvayevskaya et al. (2024). Comparison between parameter-efficient techniques and full fine-tuning: A case study on multilingual news article classification. *PLoS ONE*, *19*(5), e0301738. https://doi.org/10.1371/journal.pone.0301738

[4] B. K. Waruwu, E. C. Tandoc, A. Duffy, N. Kim, & R. Ling. (2021). Telling lies together? Sharing news as a form of social authentication. *New Media & Society*, *23*(9), 2516-2533. https://doi.org/10.1177/1461444820931017

[5] X. Zhou, & R. Zafarani. (2021). A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. *ACM Comput. Surv.*, *53*(5), 1-40. https://doi.org/10.1145/3395046

[6] M. Montesi. (2021). Understanding fake news during the Covid-19 health crisis from the perspective of information behaviour: The case of Spain. *Journal of Librarianship and Information Science*, *53*(3), 454-465. https://doi.org/10.1177/0961000620949653

[7] A. Fourney, M. Z. Racz, G. Ranade, M. Mobius, & E. Horvitz. (2017). Geographic and Temporal Trends in Fake News Consumption During the 2016 US Presidential Election. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, Singapore Singapore: ACM, Nov. 2017, pp. 2071-2074. https://doi.org/10.1145/3132847.3133147

[8] J. Wang, J. Zheng, S. Yao, R. Wang, & H. Du. (2023). TLFND: A Multimodal Fusion Model Based on Three-Level Feature Matching Distance for Fake News Detection. *Entropy*, *25*(11), 1533. https://doi.org/10.3390/e25111533

[9] A. Ali et al. (2025). Towards improved fake news detection using a hybrid RoBERTa and metadata enhanced XGBoost model. *Sci Rep*, *16*(1), 1967. https://doi.org/10.1038/s41598-025-29942-y

[10] E. S. Albtoush, K. H. Gan, & S. A. A. Alrababa. (2025). Fake news detection: state-of-the-art review and advances with attention to Arabic language aspects. *PeerJ Computer Science*, *11*, e2693. https://doi.org/10.7717/peerj-cs.2693

[11] Y. Peng, J. Qi, & Y. Zhuo. (2020). MAVA: Multi-Level Adaptive Visual-Textual Alignment by Cross-Media Bi-Attention Mechanism. *IEEE Trans. on Image Process.*, *29*, 2728-2741. https://doi.org/10.1109/TIP.2019.2952085

[12] H. Reddy, N. Raj, M. Gala, & A. Basava. (2020). Text-mining-based Fake News Detection Using Ensemble Methods. *Int. J. Autom. Comput.*, *17*(2), 210-221. https://doi.org/10.1007/s11633-019-1216-5

[13] I. Ahmad, M. Yousaf, S. Yousaf, & M. O. Ahmad. (2020). Fake News Detection Using Machine Learning Ensemble Methods. *Complexity*, *2020*, 1-11. https://doi.org/10.1155/2020/8885861

[14] Y. Jiang, & Y. Wang. (2025). IMFND: In-context multimodal fake news detection with large visual-language models. *Knowledge-Based Systems*, *325*, 113880. https://doi.org/10.1016/j.knosys.2025.113880

[15] P. Zhu, J. Hua, K. Tang, J. Tian, J. Xu, & X. Cui. (2024). Multimodal fake news detection through intra-modality feature aggregation and inter-modality semantic fusion. *Complex Intell. Syst.*, *10*(4), 5851-5863. https://doi.org/10.1007/s40747-024-01473-5

[16] J. Qiao, X. Li, C. Gao, L. Wu, J. Feng, & Z. Wang. (2025). Improving multimodal fake news detection by leveraging cross-modal content correlation. *Information Processing & Management*, *62*(5), 104120. https://doi.org/10.1016/j.ipm.2025.104120

[17] Y. Jiang, & Y. Wang. (2025). IMFND: In-context multimodal fake news detection with large visual-language models. *Knowledge-Based Systems*, *325*, 113880. https://doi.org/10.1016/j.knosys.2025.113880

[18] Z. Qu, Y. Meng, G. Muhammad, & P. Tiwari. (2024). QMFND: A quantum multimodal fusion-based fake news detection model for social media. *Information Fusion*, *104*, 102172. https://doi.org/10.1016/j.inffus.2023.102172

[19] Z. Jin, J. Cao, H. Guo, Y. Zhang, & J. Luo. (2017). Multimodal Fusion with Recurrent Neural Networks for Rumor Detection on Microblogs. In *Proceedings of the 25th ACM international conference on Multimedia*, Mountain View California USA: ACM, Oct. 2017, pp. 795-816. https://doi.org/10.1145/3123266.3123454

[20] P. Qi et al. (2021). Improving Fake News Detection by Using an Entity-enhanced Framework to Fuse Diverse Multimodal Clues. In *Proceedings of the 29th ACM International Conference on Multimedia*, Virtual Event China: ACM, Oct. 2021, pp. 1212-1220. https://doi.org/10.1145/3474085.3481548

[21] C. Boididou, S. Papadopoulos, M. Zampoglou, L. Apostolidis, O. Papadopoulou, & Y. Kompatsiaris. (2018). Detection and visualization of misleading content on Twitter. *Int J Multimed Info Retr*, *7*(1), 71-86. https://doi.org/10.1007/s13735-017-0143-x

[22] M. Luqman, M. Faheem, W. Y. Ramay, M. K. Saeed, & M. B. Ahmad. (2024). Utilizing Ensemble Learning for Detecting Multi-Modal Fake News. *IEEE Access*, *12*, 15037-15049. https://doi.org/10.1109/ACCESS.2024.3357661

[23] B. Cui et al. (2023). Intra-graph and Inter-graph joint information propagation network with third-order text graph tensor for fake news detection. *Appl Intell*, *53*(16), 18971-18988. https://doi.org/10.1007/s10489-023-04455-1

[24] K. Bai, S. Jin, Z. Zhang, & S. Dai. (2024). Drilling Rate of Penetration Prediction Based on CBT-LSTM Neural Network. *Sensors*, *24*(21), 6966. https://doi.org/10.3390/s24216966

[25] Y. Fan, K. Zhang, B. Zheng, Y. Zhou, J. Zhou, & W. Pan. (2025). GCSA-ResNet: a deep neural network architecture for Malware detection. *Sci Rep*, *15*(1), 24098. https://doi.org/10.1038/s41598-025-10561-6

[26] Y. Liang, T. Tohti, & A. Hamdulla. (2022). Multimodal false information detection method based on Text-CNN and SE module. *PLoS ONE*, *17*(11), e0277463. https://doi.org/10.1371/journal.pone.0277463

[27] D. Khattar, J. S. Goud, M. Gupta, & V. Varma. (2019). MVAE: Multimodal Variational Autoencoder for Fake News Detection. in *The World Wide Web Conference*, San Francisco CA USA: ACM, May 2019, pp. 2915-2921. https://doi.org/10.1145/3308558.3313552

[28] J. Wang, Z. Zhu, C. Liu, R. Li, & X. Wu. (2024). LLM-Enhanced multimodal detection of fake news. *PLoS ONE*, *19*(10), p. e0312240. https://doi.org/10.1371/journal.pone.0312240

[29] X. Zhou, J. Wu, & R. Zafarani. (2020). SAFE: Similarity-Aware Multi-Modal Fake News Detection. in *Advances in Knowledge Discovery and Data Mining*, *12085*, H. W. Lauw, R. C.-W. Wong, A. Ntoulas, E.-P. Lim, S.-K. Ng, & S. J. Pan, Eds., in Lecture Notes in Computer Science, vol. 12085. Cham: Springer International Publishing, 2020, pp. 354-367. https://doi.org/10.1007/978-3-030-47436-2_27

[30] S. Singhal, A. Kabra, M. Sharma, R. R. Shah, T. Chakraborty, & P. Kumaraguru. (2020). SpotFake+: A Multimodal Framework for Fake News Detection via Transfer Learning (Student Abstract). *AAAI*, *34*(10), 13915-13916. https://doi.org/10.1609/aaai.v34i10.7230

[31] S. Singhal, T. Pandey, S. Mrig, R. R. Shah, & P. Kumaraguru. (2022). Leveraging Intra and Inter Modality Relationship for Multimodal Fake News Detection. in *Companion Proceedings of the Web Conference 2022*, Virtual Event, Lyon France: ACM, Apr. 2022, pp. 726-734. https://doi.org/10.1145/3487553.3524650

[32] P. Qi et al. (2021). Improving Fake News Detection by Using an Entity-enhanced Framework to Fuse Diverse Multimodal Clues. In *Proceedings of the 29th ACM International Conference on Multimedia*, Virtual Event China: ACM, Oct. 2021, pp. 1212-1220. https://doi.org/10.1016/j.eswa.2023.120469

[33] Y.-K. Li, Q.-H. Meng, Y.-X. Wang, & H.-R. Hou. (2023). MMFN: Emotion recognition by fusing touch gesture and facial expression information. *Expert Systems with Applications*, *228*, 120469. https://doi.org/10.1016/j.eswa.2023.120469

[34] H. Liu, W. Xie, & S. Wang. (2024). Feature fusion and context interaction for RGB-D indoor semantic segmentation. *Applied Soft Computing*, *167*, 112379. https://doi.org/10.1016/j.asoc.2024.112379