SwinBERT-Fake: Boosting Multimodal Fake News Detection via Hierarchical Visual Modeling and Explainable AI

Yihao Zhao
School of Safety Science and Emergency Management
Wuhan University of Technology
Wuhan, China
ORCID: 0009-0001-5853-5855

Yitian Lu
School of Safety Science and Emergency Management
Wuhan University of Technology
Wuhan, China
ORCID: 0009-0008-6220-2675

Runyang Yu
School of Safety Science and Emergency Management
Wuhan University of Technology
Wuhan, China
ORCID: 0009-0005-5813-0881

Zhan Shi
School of Safety Science and Emergency Management
Wuhan University of Technology
Wuhan, China
ORCID: 0009-0002-0063-1470

Yan Tu ✉
School of Safety Science and Emergency Management
Wuhan University of Technology
Wuhan, China
tuyan_belle@163.com

Abstract*—Multimodal fake news detection faces two critical challenges: capturing fine-grained visual artifacts (e.g., splicing traces) and bridging the semantic gap between textual and visual modalities. Existing state-of-the-art approaches typically rely on standard Vision Transformers (ViT) or augment semantic information via automated image captioning. However, we argue that ViT's global attention mechanism often overlooks local discriminative details, while generated captions introduce semantic hallucinations and significant computational latency. To address these limitations, we propose SwinBERT-Fake, a novel framework that integrates Swin Transformer with BERT to prioritize hierarchical visual modeling over semantic augmentation. By leveraging Swin's shifted window mechanism, our model effectively localizes subtle visual inconsistencies that standard ViT misses. Extensive experiments on the Fakeddit dataset demonstrate that SwinBERT-Fake achieves a state-of-the-art accuracy of 88.09%, significantly outperforming ViT-based baselines. Furthermore, our rigorous ablation studies reveal that semantic augmentation via captioning is redundant, offering no statistical performance gain while increasing inference latency by approximately 4.5 times. Finally, we introduce a comprehensive Explainable AI (XAI) module—incorporating Eigen-CAM, Integrated Gradients, and SHAP—to provide transparent visual, textual, and modal-level evidence for detection. Our findings suggest that optimizing the granularity of visual backbones is a more effective and efficient pathway for misinformation detection than relying on external generative models.

Keywords—Multimodal Fake News Detection, Swin Transformer, Explainable AI (XAI), Image Captioning, Hierarchical Visual Modeling

I. INTRODUCTION

In the era of information explosion, fake news on social media poses a critical threat to economic stability and public safety. For instance, misinformation during the COVID-19 pandemic caused significant economic losses and weakened trust in public health systems [1]-[4]. Consequently, developing robust automated detection systems is urgent. While early deep learning approaches focused on textual analysis [5], modern misinformation is increasingly multimodal, combining text with images or videos to enhance spread and credibility [6]. Simple concatenation of modalities often fails to capture the intricate semantic correlations required for forensics [7]. Thus, advanced multimodal frameworks are essential for preserving trust in complex information environments.

While multimodal approaches have advanced the field, existing state-of-the-art methods still face critical limitations. Most current frameworks rely on standard Vision Transformers (ViT) [8] for visual encoding, which use global attention mechanisms that often overlook fine-grained local artifacts (e.g., splicing traces) crucial for forensics. To compensate, recent studies (e.g., Athira et al. [9]) have introduced automated image captioning as a semantic bridge. However, we argue that this adds computational redundancy and, more dangerously, introduces semantic hallucinations when the captioning model misinterprets manipulated images.

To address these challenges, we propose SwinBERT-Fake, a novel framework that integrates BERT [10] for textual semantics and Swin Transformer [11] for hierarchical visual modeling. Unlike previous methods, we prioritize the extraction of discriminative visual details over semantic augmentation. Our extensive experiments on the Fakeddit [12] dataset demonstrate that SwinBERT-Fake significantly outperforms ViT-based baselines, proving that a robust visual backbone eliminates the need for error-prone image captions while offering transparent multi-level interpretability.

The contributions of this work are threefold. First, we propose SwinBERT-Fake, a hierarchical multimodal framework that integrates BERT [10] and Swin Transformer [11] to effectively capture both semantic inconsistencies and fine-grained visual artifacts. Second, through rigorous ablation studies, we demonstrate that the Swin Transformer significantly outperforms standard ViT, and crucially, that semantic augmentation via automated captioning is redundant, introducing computational latency without meaningful performance gains. Third, we develop a comprehensive Explainable AI (XAI) module—incorporating Eigen-CAM [13], Integrated Gradients [14], and SHAP [15]—that provides granular interpretability at visual, textual, and modal levels, thereby enhancing trust in automated detection systems.

II. RELATED WORK

A. Evolution of Multimodal Fake News Detection

Early approaches primarily utilized Convolutional Neural Networks (CNNs) for visual feature extraction. Wang et al.

[16] proposed EANN, which introduced adversarial learning to remove event-specific features. Subsequently, Singhal et al. [17] pioneered SpotFake, fusing VGG-19 [18] visual features with BERT textual embeddings. Building on this, Qian et al. [19] developed HMCAN, integrating ResNet-50 [20] with a hierarchical attention network. Similarly, Zhou et al. [21] presented SAFE, focusing on cross-modal similarity to detect semantic mismatches. Khattar et al. [22] further advanced this by proposing MVAE, utilizing a bimodal variational autoencoder to learn shared representations.

While effective, these CNN-based methods lack the flexibility to capture long-range semantic dependencies. The field has since shifted towards Transformer-based architectures. Dosovitskiy et al. [8] marked a milestone with ViT, applying global self-attention to image patches. Recently, Athira et al. [9] augmented ViT with automated image captioning via BLIP [23] to bridge the semantic gap. However, we argue that ViT's global attention often dilutes fine-grained artifacts, and captioning introduces potential hallucinations. To address this, we adopt the Swin Transformer, developed by Liu et al. [11], which utilizes a hierarchical shifted window mechanism to capture local details akin to CNNs but with the global modeling power of Transformers.

B. Explainable AI (XAI) in Forensics

Explainability is crucial for trust in automated detection, as emphasized by Shu et al. [24] in their Defend framework. Lundberg and Lee [15] introduced SHAP, a game-theoretic approach to quantify feature contributions consistently. For visual explanations, while Grad-CAM [25] is standard for CNNs, it often yields diffuse heatmaps on Transformers. Consequently, we employ Eigen-CAM [13], proposed by Muhammad et al., which visualizes principal components to offer better localization for hierarchical models. For text, Integrated Gradients (IG) by Sundararajan et al. [14] attributes predictions to specific input tokens, satisfying the completeness axiom. Unlike perturbation-based methods like LIME [26], IG offers fine-grained token-level interpretability suitable for analyzing emotional triggers in fake news text.

III. Methodology

In this section, we present the proposed SwinBERT-Fake framework for multimodal fake news detection. Our approach addresses the limitations of global visual attention mechanisms by integrating a hierarchical Swin Transformer with a pre-trained BERT encoder. The overall architecture is illustrated in Fig. 1.

A. Framework Overview

We formulate fake news detection as a binary classification task. Given a dataset $D = (T_i, I_i, y_i)_{(i=1)}^{N}$, where T_i is the text, I_i is the image, and $y_i \in 0, 1$ is the label, we aim to learn a mapping $f: (T, I) \rightarrow y$. The framework consists of three components: a BERT-based textual extractor, a Swin-based visual extractor, and a fusion classifier. An integrated XAI module (Eigen-CAM, IG, SHAP) provides post-hoc interpretability.

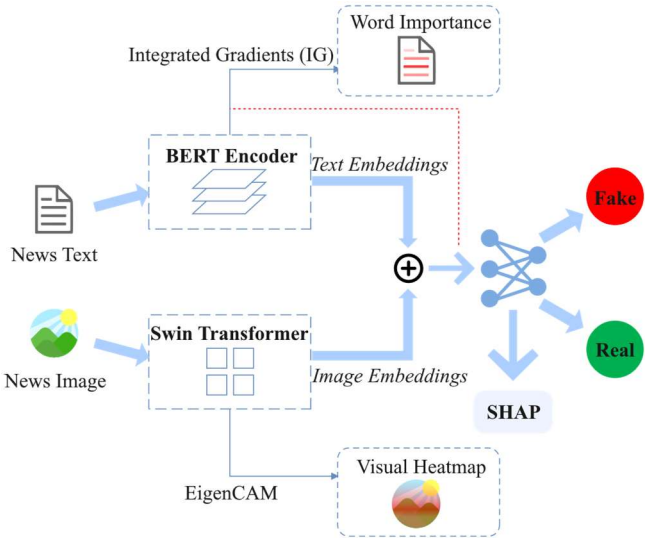

Fig. 1. The Architecture of SwinBERT-Fake Framework.

B. Textual Feature Extraction

Formally, let H_T be the output sequence of the last transformer layer. We utilize the hidden state corresponding to the [CLS] token as the aggregate semantic representation. The textual vector v_t is obtained as:

$$v_t = H_{T,[\mathrm{CLS}]} \in \mathbb{R}^{d_{text}} \tag{1}$$

where $d_{text} = 768$ is the embedding dimension for the BERT-base model. This vector encapsulates the global semantic context required for consistency checking.

C. Hierarchical Visual Modeling via Swin Transformer

Unlike standard ViT [8], which uses global attention that may smooth out fine-grained artifacts, we adopt the Swin Transformer [11] to capture local visual details through a hierarchical shifted window mechanism. Given an image I, Swin computes self-attention within local windows. For a feature map z^l, the attention is:

$$\mathrm{Attention}(Q, K, V) = \mathrm{Softmax}\left(\frac{QK^T}{\sqrt{d}} + B\right)V \tag{2}$$

where Q, K, V are the query, key, and value matrices derived from the local window features, and B is the relative position bias. To enable cross-window interaction, Swin alternates between Window-based (W-MSA) and Shifted Window-based (SW-MSA) Multi-head Self-Attention. The transformation of features between layers l and $l+1$ is formulated as:

$$\hat{z}^l = \mathrm{W-MSA}\left(\mathrm{LN}\left(z^{l-1}\right)\right) + z^{l-1} \tag{3}$$

$$z = \mathrm{MLP}\left(\mathrm{LN}\left(\hat{z}^l\right)\right) + \hat{z}^l \tag{4}$$

$$\hat{z}^{l+1} = \mathrm{SW-MSA}\left(\mathrm{LN}\left(z^l\right)\right) + z^l \tag{5}$$

$$z^{l+1} = \mathrm{MLP}\left(\mathrm{LN}\left(\hat{z}^{l+1}\right)\right) + \hat{z}^{l+1} \tag{6}$$

where LN denotes Layer Normalization.We apply global average pooling to the final output to obtain the visual vector $v_v \in \mathbb{R}^{1024}$. This representation encodes both high-level semantics and fine-grained structural artifacts essential for forensics.

D. Multimodal Fusion and Classification

We adopt a feature concatenation strategy to synthesize information while preserving individual feature integrity. T The textual vector v_t and visual vector v_v are concatenated to form a joint multimodal representation v_{fusion}:

$$v_{fusion} = v_t \oplus v_v \in \mathbb{R}^{(d_{\text{text}} + d_{\text{visual}})} \tag{7}$$

where $\oplus$ denotes the concatenation operation. This vector is fed into a Multi-Layer Perceptron (MLP) classifier. The model is trained end-to-end by minimizing the Cross-Entropy Loss:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i log(\hat{y}_i) + (1-y_i) log(1-\hat{y}_i)\right] \tag{8}$$

where $\hat{y}_i$ is the predicted probability that the i-th sample is real news.

The overall training pipeline and explainability generation process are summarized in Algorithm 1.

Algorithm 1: Training of SwinBERT-Fake Framework

Input: Dataset $\mathcal{D} = (T_i, I_i, y_i)$, Max Epochs E

Output: Trained Model Θ, Explanation Maps $\mathcal{E}$

for $epoch = 1$ **to** E **do**
  $v_t \leftarrow \text{BERT}(T)_{[\text{CLS}]}$
  $v_v \leftarrow \text{GlobalAvgPool}(\text{SwinTransformer}(I))$
  $v_{joint} \leftarrow \text{Concat}(v_t, v_v)$
  $\hat{y} \leftarrow \text{Softmax}(\text{MLP}(v_{joint}))$
  $\mathcal{L} \leftarrow \text{CrossEntropyLoss}(\hat{y}, y)$
  Update parameters Θ via AdamW
end for
Generate explanations $\mathcal{E}$ via Eigen-CAM, IG, and SHAP.
return Θ, $\mathcal{E}$

E. Integrated Explainability Module

1) Visual Localization via Eigen-CAM

Standard gradient-based methods (e.g., Grad-CAM [25]) often yield diffuse heatmaps when applied to Transformers due to the scattered nature of attention heads. To address this, we employ Eigen-CAM [13], a gradient-free visualization method. Eigen-CAM computes the principal components of the learned representations from the final block of the Swin Transformer. Formally, given the feature map $A \in \mathbb{R}^{(H \times W \times C)}$ from the last layer, we compute the first principal component v_1 using Singular Value Decomposition (SVD). The heatmap M_{visual} is then generated by projecting A onto v_1:

$$M_{visual} = \text{Resize}(A \cdot v_1) \tag{9}$$

For the ViT baseline in our comparative analysis, we similarly apply Eigen-CAM to its final normalization layer (Final Norm) to ensure a fair visual comparison.

2) Textual Attribution via Integrated Gradients

For textual explainability, we utilize Integrated Gradients (IG) [14]. IG addresses the saturation problem of gradients by integrating the gradients along a linear path from a baseline input x' (typically a zero embedding) to the actual input x. The attribution score for the i-th token is calculated as:

$$\text{Attr}_i = (x_i - x_i') \times \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha(x - x'))}{dx_i} d\alpha \tag{10}$$

where F is the model's classification function. We further apply stopword filtering to remove non-informative tokens, ensuring that high-attribution scores are assigned only to semantically meaningful keywords.

3) Modality Contribution Analysis via SHAP

To quantify the global contribution of each modality, we adapt the SHAP [15] framework utilized by Athira et al. [9]. Omitting the redundant caption modality, we treat the textual embedding v_t and visual embedding v_v as two distinct coalitions of features. Using the KernelExplainer, we approximate the Shapley value Φ_m for each modality $m \in text, image$, representing its marginal contribution to the prediction logits relative to a background dataset sampled from the training set. This allows us to determine whether the model's decision is text-driven or image-driven for any given sample.

IV. Experiments

In this section, we evaluate our proposed SwinBERT-Fake framework (hereinafter referred to as SwinBERT for brevity). We detail the experimental setup, benchmark against state-of-the-art baselines, and conduct rigorous ablation studies to validate the efficacy of the hierarchical visual backbone and the efficiency trade-offs of semantic augmentation.

A. Dataset and Implementation Details

Our experiments were conducted on the Fakeddit [12] dataset. Specifically, we constructed our experimental subset from the official multimodal_train split. To ensure data quality and computational feasibility, we first filtered out entries with missing labels or image URLs, and then performed stratified sampling based on the binary 2_way_label with a target size of 61,000. After removing instances with inaccessible or corrupted image links during the download phase, we obtained a final, highly balanced subset of 60,547 valid multimodal samples (30,078 real and 30,469 fake). This subset was subsequently stratified—again based on the 2_way_label—into training, validation, and testing sets following a standard 8:1:1 ratio. To ensure full reproducibility, the exact dataset split indices (fixed with random seed 42) and the complete source code are publicly available at: https://github.com/haodashi123/swinbert-fake.

The textual backbone was initialized with bert-base-uncased, while visual backbones used swin-base-patch4-window7-224 and vit-base-patch16-224. To optimize efficiency on limited hardware (RTX 3050, 4GB VRAM) and

mitigate overfitting, we adopted a "Frozen Backbone" strategy. Feature vectors for both modalities were pre-extracted and stored offline using a batch size of 16. For the fusion network, we trained a lightweight Multi-Layer Perceptron (MLP) on the frozen features, avoiding complex cross-attention layers that exacerbate overfitting on small datasets. The training process utilized the AdamW optimizer with an initial learning rate of 1e-4 and a weight decay of 1e-3 to provide strong regularization. A learning rate scheduler, ReduceLROnPlateau, was implemented to reduce the learning rate by a factor of 0.2 if the validation loss did not improve for 3 consecutive epochs. The model was trained for a maximum of 50 epochs with an early stopping patience of 5 epochs. To ensure the reproducibility of our results, the global random seed was fixed at 2026. Furthermore, the explainability modules were implemented using the pytorch-grad-cam library for visual heatmaps and the Captum [27] library for Integrated Gradients.

B. Main Results: Multimodal Performance Comparison

We compared our proposed SwinBERT model against several baselines, including the classic SpotFake (BERT+VGG) and recent Vision Transformer-based approaches. The quantitative results are summarized in TABLE I.

As shown in TABLE I, our proposed SwinBERT achieves an accuracy of 88.09% and an F1-score of 0.8825, significantly outperforming the classic SpotFake baseline, which attained an accuracy of 84.10%. Furthermore, SwinBERT surpasses the current state-of-the-art architecture BERT-ViT-BLIP (87.07% accuracy) by a margin of roughly 1.02%. To verify the statistical significance of these improvements, we conducted a comprehensive suite of hypothesis tests.

The comparison between SwinBERT and the BERT-ViT-BLIP baseline yielded a McNemar's [28] p-value of 0.0087 ($p < 0.01$), rejecting the null hypothesis that the two models have the same error rate. Moreover, both the Wilcoxon signed-rank test ($p < 0.0001$) and Paired t-test ($p < 0.0001$) on the per-sample prediction losses confirm that SwinBERT achieves a significantly lower average loss (0.2857 vs. 0.3133). This indicates that our hierarchical visual backbone not only improves classification accuracy but also produces more confident and robust predictions compared to the ViT-based baseline.

C. Ablation Study: Impact of Visual Backbone

To isolate the contribution of the visual encoder, we evaluated the performance of unimodal (image-only) models using VGG-19, ViT, and Swin Transformer. The results are presented in TABLE II.

The results demonstrate a clear hierarchy in visual feature extraction capabilities. The Swin Transformer achieves the highest unimodal accuracy of 82.44%, significantly outperforming the standard ViT (78.56%) and the CNN-based VGG-19 (75.62%). The substantial gap of nearly 4% between Swin and ViT suggests that the hierarchical architecture and shifted window mechanism of Swin are far more effective at capturing the fine-grained visual artifacts and local inconsistencies typical of manipulated news images than the global attention mechanism of standard ViT.

TABLE I. PERFORMANCE COMPARISON OF MULTIMODAL MODELS ON THE FAKEDDIT DATASET.

Model	Acc.	Prec.	Rec.	F1	AUC
SpotFake	0.8410	0.8308	0.8537	0.8421	0.9184
BERT-VIT-BLIP	0.8707	0.8478	0.9016	0.8739	0.9401
BERT-VIT	0.8707	0.8546	0.8913	0.8726	0.9390
SwinBERT (Proposed)	0.8809	**0.8655**	0.9003	0.8825	0.9491
SwinBERT+BLIP	**0.8837**	0.8564	**0.9202**	**0.8872**	**0.9495**

TABLE II. PERFORMANCE COMPARISON OF UNIMODAL IMAGE-ONLY MODELS.

Model	Acc.	Prec.	Rec.	F1	AUC
VGG	0.7562	0.7383	0.7889	0.7628	0.8348
ViT	0.7856	0.7669	0.8168	0.7910	0.8593
Swin Transformer	**0.8244**	**0.7927**	**0.8757**	**0.8321**	**0.8941**

D. Ablation Study: Necessity of Semantic Augmentation

To ensure a fair comparison, the SwinBERT+BLIP variant adopts an identical fusion strategy to the BERT-ViT-BLIP baseline [9], where the caption embedding v_c generated by BLIP is concatenated with the text embedding v_t and visual embedding v_v.

As shown in TABLE I, SwinBERT+BLIP achieves a marginal accuracy gain (88.37%) over standalone SwinBERT (88.09%). However, McNemar's test ($p = 0.3183$) confirms this difference is not statistically significant. While the Wilcoxon signed-rank test ($p < 0.0001$) indicates a slight reduction in prediction loss ranks, the Paired t-test ($p = 0.1729$) reveals no significant change in the mean loss distribution. This suggests that captions may only marginally boost confidence on easy samples without correcting misclassifications, rendering BLIP's semantic augmentation largely redundant to Swin's rich visual features.

Efficiency Analysis. To rigorously quantify the computational cost, we conducted a robust inference latency benchmark using randomized synthetic data to eliminate hardware caching bias. The textual input consisted of a standard news headline dynamically appended with a unique 10-character random suffix per iteration, forcing the tokenizer to re-execute full encoding. Similarly, visual inputs were generated as 224×224 RGB images with uniformly distributed random noise to simulate peak load on the visual backbone. We performed 200 independent inference cycles following a 20-round warm-up phase, enforcing CUDA synchronization and garbage collection to ensure precision. Outliers were removed using Tukey's method (1.5 IQR), and results are reported with 95% confidence intervals.

As illustrated in Fig. 2, the integration of BLIP introduces a substantial bottleneck. The inference latency surges from 33.06 ± 0.18 ms for the standalone SwinBERT to 180.25 ± 3.92 ms for SwinBERT+BLIP. This represents an approximate 4.5-fold increase in computational overhead. Given that the semantic augmentation yields no statistically significant gain in accuracy (in McNemar's test) while degrading real-time throughput by over 80%, we conclude that the cost of automated captioning outweighs its benefits in this context. Consequently, we prioritize the streamlined SwinBERT architecture for its superior balance of accuracy and efficiency.

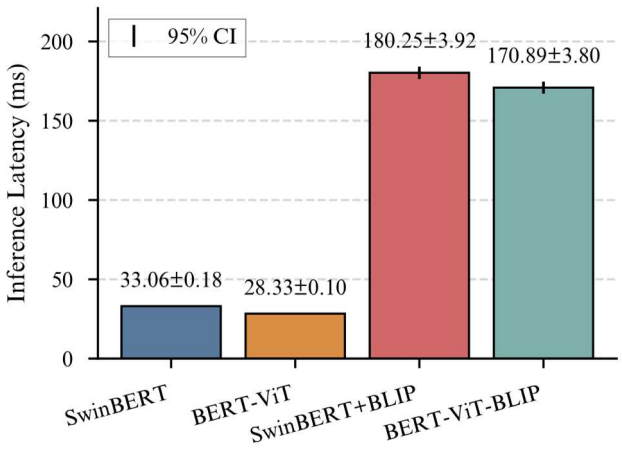

Fig. 2. Inference latency comparison with 95% confidence intervals.

V. Qualitative Analysis

We present three representative case studies to qualitatively validate the effectiveness of our proposed framework and the interpretability of its decisions.

A. Semantic Hallucination in Captioning

Fig. 3 illustrates a "False Context" fake news sample titled "nice puppies," featuring a guinea pig and a puppy. While the image itself is authentic, the misleading caption generated by BLIP—"a man holding two small dogs"—introduces two critical errors: hallucinating a human subject and misidentifying the species. This erroneous semantic context misled the BERT-ViT-BLIP model into predicting "Real" (58.6% confidence). In contrast, both BERT-ViT and our SwinBERT, which rely on intrinsic features without external noise, correctly identified the sample as "Fake," demonstrating the risk of relying on automated captioning for forensic tasks.

B. Localization Precision: Swin vs. ViT

Fig. 4 showcases a deepfake image of a "Saiga Quokkalope" (a spliced mythical creature). This sample poses a significant challenge, deceiving both the BERT-ViT and BERT-ViT-BLIP models into predicting "Real" with high confidence (>89%). The Eigen-CAM visualization reveals that standard ViT (Middle) fails to localize the manipulated subject, focusing instead on background noise. Conversely, SwinBERT correctly flags it as "Fake" (72.1%). Its heatmap (Right) precisely highlights the composite creature, confirming that the hierarchical shifted window mechanism effectively captures the fine-grained splicing artifacts that global attention mechanisms overlook.

C. Adaptive Reasoning via SHAP

Fig. 5 presents a case of pareidolia titled "noticed this guy watching me," where wall cracks resemble a face. SwinBERT confidently classifies this as "Fake" (96.9%). The SHAP modality contribution analysis reveals that the decision was predominantly driven by the Visual Modality (67%). The visual heatmap confirms that the model focused on the specific texture of the "eyes" and "mouth" on the wall. This indicates that SwinBERT successfully distinguished the visual anomaly from a genuine human face, relying on visual evidence when the textual cue was ambiguous.

Fig. 3. Visualization of a fake news sample ("nice puppies") exposed by SwinBERT.

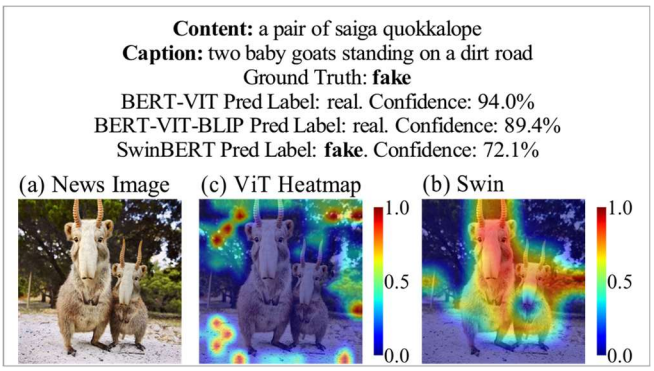

Fig. 4. Visual attention comparison on a deepfake sample ("saiga quokkalope").

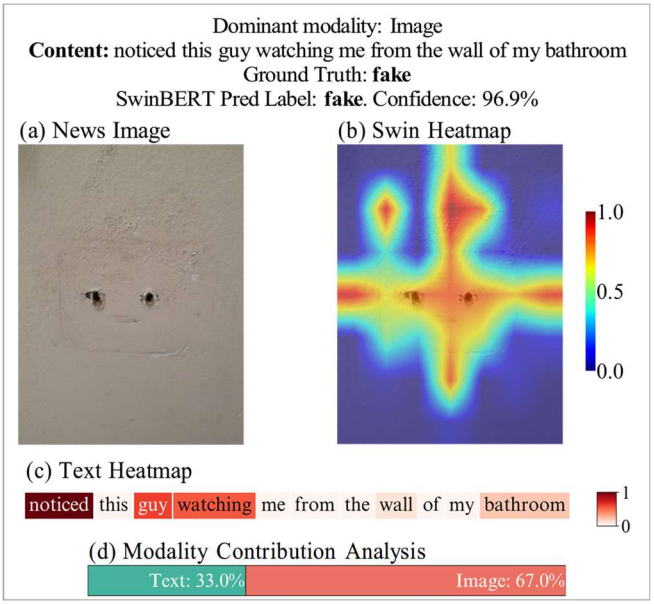

Fig. 5. Multimodal contribution analysis of an image-dominant fake sample.

VI. Conclusion

In this paper, we proposed SwinBERT-Fake, a novel multimodal framework for fake news detection that prioritizes hierarchical visual modeling over semantic augmentation. By replacing the standard Vision Transformer with the Swin

Transformer, we significantly enhanced the model's ability to detect fine-grained visual artifacts and splicing traces. Our rigorous experiments on the Fakeddit dataset demonstrate that SwinBERT-Fake achieves a state-of-the-art accuracy of 88.09%, significantly outperforming baseline models ($p < 0.01$) while maintaining high inference efficiency. Our ablation studies critically re-evaluated the role of image captioning, concluding that it introduces semantic hallucinations and computational overhead without yielding statistically significant performance gains. Furthermore, we developed a comprehensive explainability suite that provides intuitive visual and textual evidence for each prediction.

Despite these promising results, our study has several limitations that outline directions for future research. First, our evaluation was restricted to the single Fakeddit dataset; extending validation to other multimodal datasets (e.g., GossipCop, PolitiFact [29]) is necessary to verify cross-domain generalization. Second, our framework currently processes single-image news posts; handling multi-image or video-based news remains an open challenge. Third, constrained by computational resources, we employed a frozen backbone strategy with a straightforward feature concatenation fusion. Future work will explore end-to-end fine-tuning alongside more sophisticated cross-attention mechanisms to model deeper inter-modal interactions, as well as the integration of Large Language Models (LLMs) to generate human-readable fact-checking reports.

ACKNOWLEDGMENT

This work was supported by the Undergraduate Training Program for Innovation and Entrepreneurship of Hubei Province under Grant No. S202510497323. The authors sincerely thank the anonymous reviewers and the handling editor for their constructive feedback.

REFERENCES

[1] T. Assenza, F. Collard, and P. Feve, "From buzz to bust: How fake news shapes the business cycle," CEPR Discuss. Pap., no. 18912, 2024. [Online]. Available: https://cepr.org/voxeu/columns/buzz-bust-how-fake-news-shapes-business-cycle.

[2] R. Cavazos, "Fake online reviews 2021: The economic cost of bad actors on the internet," CHEQ, Univ. of Baltimore, Baltimore, MD, USA, 2021. [Online]. Available: https://f.hubspotusercontent00.net/hubfs/5228455/Research/Fake%20Online%20Reviews%202021.pdf.

[3] R. Bruns, D. Hosangadi, and M. Trotochaud, "COVID-19 vaccine misinformation and disinformation costs an estimated 50 to 300 million dollars each day," The Johns Hopkins Center for Health Security, 2021. [Online]. Available: https://centerforhealthsecurity.org/sites/default/files/2023-02/20211020-misinformation-disinformation-cost.pdf.

[4] S. Bhattacharya and A. Singh, "Unravelling the infodemic: A systematic review of misinformation dynamics during the COVID-19 pandemic," Front. Commun., vol. 10, p. 1560936, Apr. 2025.

[5] E. I. Setiawan, P. Sutanto, C. N. Purwanto, J. Santoso, F. X. Ferdinandus, et al., "An image and text-based fake news detection with transfer learning," PLoS One, vol. 20, no. 6, p. e0324394, 2025.

[6] S. K. Uppada, K. S. Patel, and S. S. S. R. Bhireddy, "An image and text-based multimodal model for detecting fake news in OSN's," J. Intell. Inf. Syst., vol. 61, no. 2, pp. 367–393, Nov. 2022.

[7] Y. Zhu, Y. Wang, Z. Yu, and W. Ding, "Multimodal fake news detection: MFND dataset and shallow-deep multitask learning," in Proc. 34th Int. Joint Conf. Artif. Intell. (IJCAI 25), 2025.

[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in Proc. Int. Conf. Learn. Represent. (ICLR), 2021.

[9] A. B. Athira, S. D. M. Kumar, and A. M. Chacko, "Pretrained transformers for multimodal fake news detection: Explainability using SHapley additive exPlanations for contributions from text, image, and image captions," Eng. Appl. Artif. Intell., vol. 162, p. 112590, 2025.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. 2019 Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol. (NAACL-HLT), vol. 1, 2019, pp. 4171–4186.

[11] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2021, pp. 9992–10002.

[12] K. Nakamura, S. Levy, and W. Y. Wang, "Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection," in Proc. 12th Lang. Resources and Eval. Conf. (LREC), 2020, pp. 6149-6157.

[13] M. B. Muhammad and M. Yeasin, "Eigen-CAM: class activation map using principal components," in Proc. Int. Joint Conf. Neural Netw. (IJCNN), 2020, pp. 1–7.

[14] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in Proc. 34th Int. Conf. Mach. Learn. (ICML), 2017, pp. 3319–3328.

[15] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2017, pp. 4765–4774.

[16] Y. Wang, F. Ma, Z. Jin, Y. Yuan, G. Xun, et al., "EANN: Event adversarial neural networks for multi-modal fake news detection," in Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery and Data Mining (KDD 18), 2018, pp. 849–857.

[17] S. Singhal, R. R. Shah, T. Chakraborty, P. Kumaraguru, and S. Satoh, "SpotFake: A multi-modal framework for fake news detection," in Proc. 2019 IEEE 5th Int. Conf. Multimedia Big Data (BigMM), 2019, pp. 39–47.

[18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in Proc. Int. Conf. Learn. Represent. (ICLR), 2015. [Online]. Available: https://arxiv.org/abs/1409.1556.

[19] S. Qian, J. Wang, J. Hu, Q. Fang, and C. Xu, "Hierarchical multi-modal contextual attention network for fake news detection," in Proc. 44th Int. ACM SIGIR Conf. Research and Develop. Inf. Retrieval (SIGIR 21), 2021, pp. 153–162.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 770–778.

[21] X. Zhou, J. Wu, and R. Zafarani, "SAFE: Similarity-aware multi-modal fake news detection," Adv. Knowl. Discov. Data Min. (PAKDD), vol. 12085, pp. 27–40, 2020.

[22] D. Khattar, J. S. Goud, M. Gupta, and V. Varma, "MVAE: Multimodal variational autoencoder for fake news detection," in Proc. World Wide Web Conf. (WWW), 2019, pp. 2915–2921.

[23] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in Proc. Int. Conf. Mach. Learn. (ICML), 2022, pp. 12888–12900.

[24] K. Shu, L. Cui, S. Wang, D. Lee, and H. Liu, "dEFEND: Explainable fake news detection," in Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery and Data Mining (KDD), 2019, pp. 1916–1926.

[25] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2017, pp. 618–626.

[26] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery and Data Mining (KDD), 2016, pp. 1135–1144.

[27] N. Kokhlikyan, V. Mastropietro, B. Low, R. Carvajal-Reyes, J. Schmidt, et al., "Captum: A unified and generic model interpretability library for PyTorch," arXiv preprint arXiv:2009.07896, 2020.

[28] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," Psychometrika, vol. 12, no. 2, pp. 153–157, 1947.

[29] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, "FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media," Big Data, vol. 8, no. 3, pp. 171–188, 2020.