Multimodal Large Language Model for Deepfake Video Detection and Description

Haoran Sun[1], Chen Cai[2], Kong Aik LEE[1], Lap-Pui Chau[1], and Yi Wang[1]
[1]Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University, Hong Kong
[2]School of Electrical and Electronic Engineering, Nanyang Technology University, Singapore
E-mail: haoran-eee.sun@connect.polyu.hk, yi-eie.wang@polyu.edu.hk (Correspondence)

***Abstract*—The explosive growth of deepfake video technology is steadily eroding public trust in visual media, necessitating detectors that not only flag forgeries but also provide explanations for them. In this paper, we introduce DVDD-LLaMA, a multimodal large language model that unifies two complementary vision encoders: CLIP for frame-level cross-modal alignment and a SwinT-based Deepfake-Sniffing Encoder for spatio-temporal anomaly capture, followed by a Compact Visual Connector that condenses those features while preserving critical manipulation cues. A lightweight bridging layer then fuses these visual signals with user prompts, enabling the language model to deliver both a real/fake result and a detailed, human-readable rationale. To train and evaluate our model, we construct FF++VQA, a richly annotated deepfake-video question–answer dataset. Fine-tuned on the dataset, DVDD-LLaMA establishes new state-of-the-art performance in both supervised and zero-shot settings and remains robust to previously unseen attack types. Ablation studies confirm the importance of the Deepfake-Sniffing Encoder and Compact Visual Connector. Experiments show that DVDD-LLaMA offers a high-performance and describable solution for deepfake video detection.**

I. Introduction

With the availability of accessible video editing software, artificial intelligence-generated content (AIGC) techniques can be easily used to produce deepfake media, enabling face swapping, expression alteration, and motion manipulation. Such deepfakes threaten personal privacy, enable political misinformation, and erode public confidence in visual evidence [1]. Recently, a surge of research has focused on various deepfake video detection methods [2], [3] to identify manipulated content and mitigate their risks.

Existing conventional deepfake video detection methods [2], [3] primarily rely on deep neural network [4] models, functioning as binary classifiers. As a result, these approaches inherently lack basic interpretability regarding their decision-making processes and are unable to provide explicit, commonsense explanations in textual form for the underlying reasons of authenticity or falsification. Furthermore, the single-perspective binary classification framework of traditional methods, combined with the limitations of current training datasets, restricts the generalization capability of detection models. When confronted with unseen deepfake variants, their performance often degrades significantly. Recent progress in multimodal large language models (MLLMs) has inspired their use in deepfake image detection [5], [6]. Yet, models trained solely at the image level [7], [8] lack access to temporal cues, which limits their effectiveness when the inputs shift from single images to deepfake videos. In the latter, frame-to-frame inconsistency becomes the primary evidence of manipulation for deepfake videos.

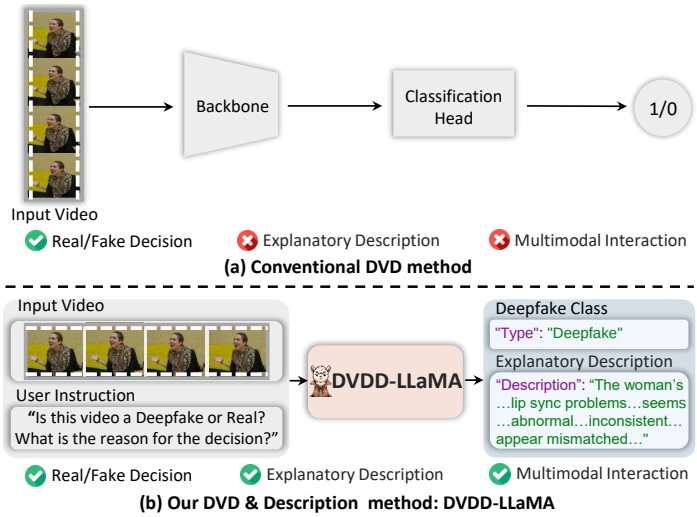

Fig. 1. Illustration of the conventional Deepfake Video Detection (DVD) and our DVD & Description (DVDD) framework.

To address these problems, we propose the DVDD-LLaMA for Deepfake Video Detection and Description task by leveraging a multimodal large language model (MLLM), e.g., LLaMA, to generate deepfake classes and corresponding textual descriptions based on video content and queries. We encourage MLLMs to focus on cognitive perception of video authenticity rather than the traditional framework (i.e., the backbone encoder with a classification head), as shown in Fig. 1. To better encode video data and connect with LLM, we propose a Swin Transformer (SwinT) based Deepfake-Sniffing Encoder (DSEncoder) to compensate for the CLIP [9] encoder and a Compact Visual Connector (CVC) for feature fusion. Moreover, we construct a visual question answering dataset for deepfake video detection and description, termed FF++VQA. The video samples are sourced from the publicly available FaceForensics++ dataset [10], and we design tailored prompts to leverage a powerful multimodal large language

979-8-3315-7206-8/25/$31.00

model [11] with video understanding capabilities for generating diverse VQA pairs. FF++VQA consists of quadruples including videos, queries, deepfake classes, and explanatory descriptions. Using FF++VQA, we fine-tune an MLLM [12] to enable accurate judgment and generation of explanatory descriptions. This process ultimately forms a comprehensive pipeline for detection and description. Our contributions are summarized as follows:

- We propose, for the first time, a MLLM-empowered deepfake video detection and description framework. It is capable of providing both detection results and reasonable explanations, effectively addressing the lack of interpretability in existing DVD methods.
- We develop a DSEncoder to capture spatio-temporal forgery features from consecutive frames and effectively resolve the temporal information loss problem in CLIP. Meanwhile, a CVC is proposed to aggregate features and perform spatio-temporal compression, generating compact task-specific representations.
- Extensive experiments demonstrate that our method can accurately analyze manipulated videos and surpasses previous large video understanding models in both detection and description tasks. In zero-shot experiments across different forgery techniques, our model outperforms the previous state-of-the-art traditional models.

II. Related Work

Video-level deepfake detection. To improve the accuracy of video-level deepfake detection, researchers have explored a variety of approaches, primarily focusing on capturing temporal inconsistencies in forged videos as a general cue. Recent studies have shown that vision transformers (ViT) achieve outstanding performance in deepfake detection tasks [13]. Compared to CNN-based models, these methods deliver superior results, but at the expense of computational efficiency. For example, DFTD [14] considers both global and local information simultaneously, yet fails to address the issue of high computational complexity. On the other hand, the SwinT [15] can generate hierarchical feature representations and exhibits linear computational complexity with respect to the input image size, making it a backbone method for various vision tasks. However, in deepfake detection, SwinT typically focuses on capturing short-term temporal inconsistencies and does not fully exploit long-range dependencies. To overcome these limitations, some studies have explored dual-branch architectures that combine short-term and long-term temporal modeling, such as ICT [16] and DFLL [17]. Nevertheless, these methods still face challenges in generalization, especially when detecting face reenactment and fully synthesized faces, and their ability to detect unseen forgery techniques remains limited. Furthermore, as they usually only output binary classification results, they lack interpretability.

Multimodal large language model. With the rapid advancement of large language models (LLMs), researchers have explored the integration of LLMs with video encoders to leverage their strong generative and understanding capabilities for video tasks [18], [19]. These studies often utilize open-source LLMs, such as Vicuna [20] and LLaMA [12]. The primary distinction among these works lies in how video features are encoded into visual tokens compatible with LLMs. For example, VideoChat [21] employs vision transformers to encode video features and uses a Query Transformer (Q-Former) [22] to compress video tokens. Similarly, VideoLLaMA [23] combines ViT and image Q-Formers to encode individual frames, followed by temporal modeling using a video Q-Former. While existing multimodal models perform well on short video tasks, such as captioning or question-answering, they lack fine-grained temporal modeling, making them ineffective at capturing forgery traces in long videos. To address this limitation, researchers have proposed incorporating temporal information aggregation modules, such as SwinT [15], to improve temporal sensitivity and understanding.

Parameter-efficient fine-tuning. To transfer the task-solving capabilities of large language models to the visual modality, researchers have proposed vision-language instruction tuning, enabling models to perform various tasks based on image or video content following user instructions [24], [25]. This process typically requires generating high-quality instruction data, which can be divided into two technical branches. The first branch [24], [26] integrates existing multimodal benchmark datasets and converts them into instruction formats, such as MultiInstruct [26] and InstructBLIP [24]. The second branch [27] leverages LLMs to create more diverse conversational-style data. For example, MiniGPT4 [27] and LLaVA [28] provide detailed visual descriptions and generate image- or video-centered conversational data. Despite these advancements, current methods fail to address user requirements for time-sensitive video understanding. Most existing approaches lack fine-grained temporal modeling capabilities, limiting their ability to understand or localize specific segments in long videos. Inspired by the success of instruction tuning in recent LLMs [29], [30], researchers have proposed a deepfake description and a dataset of question answers to improve the adaptability of multimodal models to deepfake-oriented tasks.

III. Method

In this section, we present the DVDD-LLaMA, a multimodal large language model designed for deepfake video detection and description (see Fig. 2). By introducing the Deepfake-Sniffing Encoder (DSEncoder) and Compact Visual Connector (CVC), our method enables fine-grained extraction, fusion, and alignment of visual features with language understanding. We also describe our specialized data curation process for the comprehensive evaluation of deepfake detection and description tasks.

Framework overview. The DVDD-LLaMA framework supports two types of input: user instructions and raw video. A frozen CLIP encoder processes sampled keyframes, providing global semantic cues and stabilizing multimodal dialogue. Meanwhile, the DSEncoder receives thumbnails composed of four consecutive frames and feeds them into resized-window

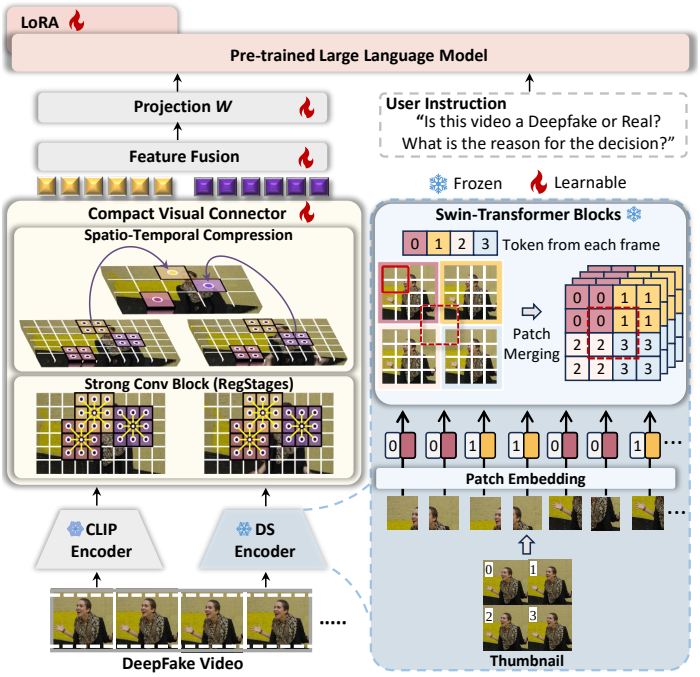

Fig. 2. Overall architecture of DVDD-LLaMA.

SwinT blocks, which efficiently model both spatial forgery features within frames and temporal inconsistencies between frames. The features generated by CLIP and the DSEncoder are fed to the CVC module, where the strong convolution block first processes the features. This block consists of four consecutive RegStage blocks, with the squeeze-and-excitation gating mechanism helping to preserve important forgery cues. Subsequently, Spatio-temporal Compression applies lightweight 3D convolutions to reduce the number of tokens without disrupting the temporal order. Finally, the fused features are mapped to the LLM through a bridging layer, which transforms the compact video tokens. The LLM performs reasoning on the fused representations and outputs the detected deepfake classes along with explanatory descriptions.

A. Deepfake-Sniffing Encoder

The SwinT [15] serves as the backbone of our DSEncoder. It takes four consecutive frames as an input set, where multiple SwinT blocks leverage a shifted window mechanism to extract spatio-temporal features from the sequential frames. These features serve as an essential input for LLM inference.

Specifically, four consecutive frames are sampled from the video in temporal order, resized to a uniform size, and embedded into a single 224×224 thumbnail using a 2×2 grid layout. This compact arrangement preserves the local information of each frame and explicitly encodes inter-frame information into spatial neighborhoods. As a result, it can optimally leverage the computational principles of self-attention and shifted windows with relative position bias. When a window spans multiple subimages, it captures spatial dependencies across temporal frames. This enables the DSEncoder to combine spatial details with temporal correlations in a unified representation. In this way, both intra-frame features and inter-frame inconsistencies can be perceived simultaneously, transforming high-dimensional visual information into compact embeddings.

Within the four stages of SwinT, we uniformly expand the window size to 14×14 for the first three stages, while the final stage uses a window size equivalent to the feature map dimension (7×7). This configuration allows earlier layers to focus on capturing local forgery traces, while later layers emphasize global temporal consistency. Meanwhile, hierarchical patch merging is applied to progressively downsample the feature maps by factors of 4, 8, and 16. This process constructs a multiscale feature hierarchy. As a result, it achieves a dynamic balance between computational efficiency and representational capacity.

Ultimately, the DSEncoder outputs a set of compact visual embeddings that fuse spatio-temporal information. This representation is robust and scalable, providing an efficient and reliable foundation of deepfake features for subsequent cross-modal inference tasks.

B. Compact Visual Connector

The operational sequence of the Compact Visual Connector is illustrated on the left of Fig. 2. Since autoregressive models (i.e., the LLM backbone) critically depend on token order consistency during both training and inference, preserving the spatio-temporal sequence of visual tokens is essential after frame-wise video encoding. To achieve this, we implement spatio-temporal compression using 3D convolutions. Given that early fusion of frame-level features enhances long-video comprehension, we substitute 2D convolution operators with 3D counterparts when compressing spatio-temporal tokens to reduce token volume. To mitigate information loss from spatio-temporal downsampling, we incorporate four RegStages [31] (strong convolutional blocks) post-downsampling. Each RegStage comprises stacked residual modules employing bottleneck structures and grouped convolution designs. Within these modules, a Squeeze-and-Excitation mechanism [31] recalibrates features to amplify critical spatio-temporal responses. The initial spatial downsampling per RegStage is achieved through convolutional stride adjustment (stride = 2), while subsequent modules preserve resolution while expanding the receptive field. Through hierarchical stacking, RegStages enable multi-level feature abstraction, where shallow layers capture local spatio-temporal patterns and deeper layers model global dependencies. This architecture significantly enhances model capacity and representational power, with residual connections and stochastic depth ensuring gradient flow and training stability, thereby establishing a robust visual foundation for video-language alignment.

C. Data Curation

Since there is currently no dedicated dataset for evaluating the deepfake video detection and description capabilities of MLLMs, we constructed a deepfake video detection and description question-answering dataset (FF++VQA) based on

publicly available deepfake video datasets, like FaceForensics++ [10]. This was achieved using carefully designed prompts and using large video understanding language models [11] to assist in data generation. The dataset consists of quadruples, including video samples, designed questions (user instructions), detected deepfake classes, and explanatory descriptions, as illustrated in Fig. 1(b). For the generated descriptions, we performed manual screening to ensure that the content length is similar and reasonable, thereby guaranteeing high-quality annotations and improving the robustness of the evaluation.

IV. Experiments

A. Implementation Details

We used CLIP [9] and our designed DSEncoder as the two encoders of this large model architecture, and LLaMA2 (7B) [12] as the large language base model. We fine-tuned DVDD-LLaMA on FF++VQA for 20 epochs, using a machine with one 4x NVIDIA RTX 6000 Ada (48GB) and a global batch size of 128 and a local batch size of 1. As shown in Fig. 2, we froze some parameters of the encoder and the large language model, and adjusted the parameters of the Compact Visual Connector, the bridge layer, and LoRA. LoRA is ranked 128, and the number of input frames is 64. In the second stage of multitask fine-tuning, we use video text data to merge high-quality, fine-grained multimodal annotation pairs to finetune DVDD-LLaMA.

B. Experiment Results and Analysis

The experimental results, as shown in Table I, clearly show that DVDD-LLaMA significantly outperforms other models in the FF++VQA set. It achieves an accuracy of 67.9%, an AUC of 67.9%, and an F1 Score of 66.3%, surpassing the previous best-performing model, VideoChat, by 11.5%, 11.3%, and 11.6%, respectively. These results demonstrate DVDD-LLaMA's superior ability to detect deepfake videos with higher reliability and precision. The model excels in capturing fine-grained features and spatio-temporal consistency, which are critical for identifying deepfake manipulations. This significant performance boost is attributed to the integration of innovative components, such as the Deepfake-Sniffing Encoder and the Compact Visual Connector. These modules enable the model to extract and process deepfake-specific features more effectively while maintaining computational efficiency.

Furthermore, as illustrated by the visualization results in Fig. 3, DVDD-LLaMA not only demonstrates a clear advantage in prediction accuracy over the baseline models, but also produces more credible, coherent, and detailed explanatory descriptions. This synergy of high accuracy and high-quality explanations enables users to quickly obtain reliable deepfake classifications while gaining a thorough understanding of the underlying rationale behind the model's decisions. These strengths further underscore the comprehensive advantages of DVDD-LLaMA in both deepfake detection performance and the interpretability of its results.

TABLE I
PERFORMANCE COMPARISON OF VARIOUS MODELS ON FF++VQA.

Model	LLM Size	FF++VQA		
		ACC	AUC	F1 Score
End2end VidLLMs				
Video-LLaVA [32]	7B	51.6	51.1	48.9
VideoChatGPT [33]	7B	52.2	51.5	49.2
VideoChat [21]	7B	56.4	56.2	54.7
DVDD-LLaMA (Ours)	7B	67.9 (+11.5)	67.9 (+11.3)	66.3(+11.6)

C. Zero-shot Performance

The zero-shot performance comparison in Table II highlights the superiority of DVDD-LLaMA over other models in the Deepfake Video Description QA task. With an LLM size of 7B, DVDD-LLaMA achieves an accuracy of 61.3%, an AUC of 60.9%, and an F1 Score of 59.4%, significantly outperforming both transformer-based models (UMMAFormer and TALL4Deepfake) and the video-based model Video-LLaVA. Transformer-based models, such as UMMAFormer and TALL4Deepfake, struggle to handle unfamiliar data in zero-shot settings due to their reliance on pre-learned patterns that may not generalize well to unseen tasks. In contrast, DVDD-LLaMA demonstrates robust adaptability, with improvements of +4.0% in accuracy, +5.1% in AUC, and +4.3% in F1 Score compared to TALL4Deepfake. This performance boost can be attributed to the advanced architecture of DVDD-LLaMA, which effectively captures spatio-temporal features and deepfake-specific patterns.

TABLE II
PERFORMANCE OF LLM-BASED MODELS AND OTHER TRADITIONAL TRANSFORMER-BASED MODELS IN ZERO-SHOT ON FF++VQA.

Model	LLM Size	FF++VQA		
		ACC	AUC	F1 Score
Transformer-based Models				
UMMAFormer [2]	-	55.8	53.1	53.7
TALL4Deepfake [3]	-	57.3	55.8	55.1
VideoLLMs				
Video-LLaVA [32]	7B	51.2	50.5	48.7
DVDD-LLaMA (Ours)	7B	61.3 (+4)	60.9 (+5.1)	59.4(+4.3)

D. Ablation Study

The ablation study in Table III highlights the critical contributions of the DSEncoder and RegStage modules to the performance of DVDD-LLaMA. Without both modules, the model achieves only 54.2% accuracy, indicating their importance. Adding RegStage alone increases the accuracy to 62.7%, showcasing its strength in reasoning and modeling deepfake-specific patterns. Similarly, incorporating DSEncoder alone improves accuracy to 58.1%, demonstrating its effectiveness in extracting spatio-temporal features. When both modules are combined, the model achieves the highest accuracy of 67.9%, confirming their complementary roles. This study underscores the synergy between these modules and validates the effectiveness of the model's architectural design.

V. Conclusions

This paper introduces DVDD-LLaMA, a novel multimodal large language model for Deepfake Video detection and

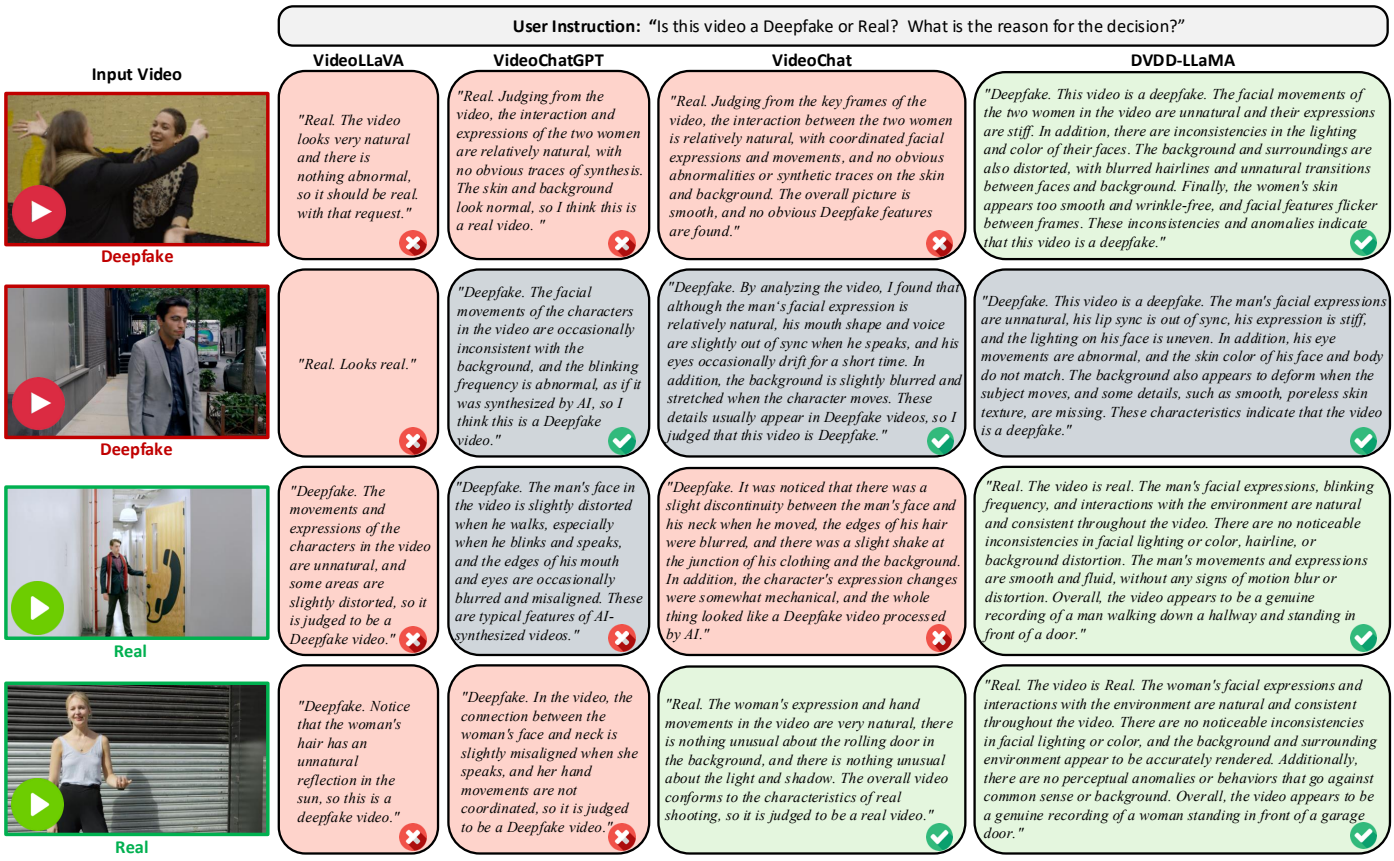

Fig. 3. Comparison of visualization results between DVDD-LLaMA, VideoLLaVA, VideoChatGPT, and VideoChat. Red boxes: incorrect predictions, green boxes: correct predictions, and gray boxes: correct predictions with hallucinated descriptions (i.e., the explanation contains information inconsistent with or fabricated beyond the video content).

TABLE III
IMPACT OF DIFFERENT MODULES (DSENCODER AND REGSTAGE) TO THE PERFORMANCE IMPROVEMENT OF THE DVDD-LLAMA MODEL ON FF++VQA.

DSEncoder	RegStage	FF++VQA (ACC)
×	×	54.2
✓	×	62.7
×	✓	58.1
✓	✓	67.9

description. By integrating the Deepfake-Sniffing Encoder (DSEncoder), Compact Visual Connector (CVC), and bridging layers with CLIP and LLaMA2, the model effectively captures fine-grained spatio-temporal features and deepfake-specific patterns while maintaining computational efficiency. Extensive experiments on the constructed FF++VQA dataset demonstrate that DVDD-LLaMA significantly outperforms existing methods in both fine-tuned and zero-shot settings, achieving state-of-the-art results in accuracy, AUC, and F1 score. The ablation study highlights the complementary roles of the DSEncoder and RegStage modules, validating the robustness and efficiency of the proposed architecture. With its ability to detect and describe deepfake manipulations with high precision and interpretability, DVDD-LLaMA sets a new benchmark for multimodal deepfake detection and description, paving the way for more reliable and scalable solutions in combating video-based forgery.

REFERENCES

[1] Y. Wang, C. Chen, N. Zhang, and X. Hu, "Watcher: Wavelet-guided texture-content hierarchical relation learning for deepfake detection," *International Journal of Computer Vision*, vol. 132, no. 10, pp. 4746–4767, 2024.

[2] R. Zhang, H. Wang, M. Du, H. Liu, Y. Zhou, and Q. Zeng, "Ummaformer: A universal multimodal-adaptive transformer framework for temporal forgery localization," in *Proceedings of the 31st ACM International Conference on Multimedia*, ACM, 2023, pp. 8749–8759.

[3] Y. Xu, J. Liang, L. Sheng, and X. Y. Zhang, "Learning spatiotemporal inconsistency via thumbnail layout for face deepfake detection," *International Journal of Computer Vision*, vol. 132, no. 12, pp. 5663–5680, 2024.

[4] Y. Li, Y. Wang, W. Wang, D. Lin, B. Li, and K.-H. Yap, "Open world object detection: A survey," *IEEE Trans. Cir. and Sys. for Video Technol.*, vol. 35, no. 2, pp. 988–1008, 2025.

[5] Z. Xu, X. Zhang, R. Li, Z. Tang, Q. Huang, and J. Zhang, "Fakeshield: Explainable image forgery detection and localization via multi-modal large language models," in *International Conference on Learning Representations*, 2025.

[6] Y. Zhang, B. Colman, X. Guo, A. Shahriyari, and G. Bharaj, "Common sense reasoning for deepfake detection," in *European Conference on Computer Vision*, Cham: Springer Nature Switzerland, 2024, pp. 399–415.

[7] C. Cai, S. Wang, K.-H. Yap, and Y. Wang, "Top-down framework for weakly-supervised grounded image captioning," *Knowledge-Based Systems*, vol. 287, p. 111 433, 2024, ISSN: 0950-7051.

[8] C. Cai, Y. Wang, and K.-H. Yap, "Interactive change-aware transformer network for remote sensing image change captioning," *Remote Sensing*, vol. 15, no. 23, 2023, ISSN: 2072-4292.

[9] A. Radford, J. W. Kim, C. Hallacy, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, PmLR, 2021, pp. 8748–8763.

[10] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to detect manipulated facial images," in *International Conference on Computer Vision (ICCV)*, 2019.

[11] Z. Cheng, S. Leng, H. Zhang, *et al.*, *Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms*, 2024. arXiv: 2406.07476 [cs.CV].

[12] H. Touvron, L. Martin, K. Stone, *et al.*, *Llama 2: Open foundation and fine-tuned chat models*, 2023. arXiv: 2307.09288 [cs.CL].

[13] H. Zhao, W. Zhou, D. Chen, W. Zhang, and N. Yu, *Self-supervised transformer for deepfake detection*, 2022. arXiv: 2203.01265 [cs.CV].

[14] A. Khormali and J.-S. Yuan, "Dfdt: An end-to-end deepfake detection framework using vision transformer," *Applied Sciences*, vol. 12, no. 6, 2022.

[15] Z. Liu, Y. Lin, Y. Cao, *et al.*, *Swin transformer: Hierarchical vision transformer using shifted windows*, 2021. arXiv: 2103.14030 [cs.CV].

[16] X. Dong, J. Bao, D. Chen, *et al.*, *Protecting celebrities from deepfake with identity consistency transformer*, 2022. arXiv: 2203.01318 [cs.CV].

[17] S. A. Khan and H. Dai, *Video transformer for deepfake detection with incremental learning*, 2021. arXiv: 2108.05307 [cs.CV].

[18] P. Jin, R. Takanobu, W. Zhang, X. Cao, and L. Yuan, *Chat-univi: Unified visual representation empowers large language models with image and video understanding*, 2024. arXiv: 2311.08046 [cs.CV].

[19] Y. Su, Y. Wang, Q. Hu, C. Yang, and L.-P. Chau, "Annexe: Unified analyzing, answering, and pixel grounding for egocentric interaction," in *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2025, pp. 9027–9038.

[20] L. Zheng, W.-L. Chiang, Y. Sheng, *et al.*, "Judging llm-as-a-judge with mt-bench and chatbot arena," *Advances in Neural Information Processing Systems*, vol. 36, pp. 46 595–46 623, 2023.

[21] K. Li, Y. He, Y. Wang, *et al.*, *Videochat: Chat-centric video understanding*, 2024. arXiv: 2305.06355 [cs.CV].

[22] J. Li, D. Li, S. Savarese, and S. Hoi, *Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models*, 2023. arXiv: 2301.12597 [cs.CV].

[23] H. Zhang, X. Li, and L. Bing, *Video-llama: An instruction-tuned audio-visual language model for video understanding*, 2023. arXiv: 2306.02858 [cs.CL].

[24] W. Dai, J. Li, D. Li, *et al.*, *Instructblip: Towards general-purpose vision-language models with instruction tuning*, 2023. arXiv: 2305.06500 [cs.CV].

[25] R. Luo, Z. Zhao, M. Yang, *et al.*, *Valley: Video assistant with large language model enhanced ability*, 2023. arXiv: 2306.07207 [cs.CV].

[26] Z. Xu, Y. Shen, and L. Huang, *Multiinstruct: Improving multi-modal zero-shot learning via instruction tuning*, 2023. arXiv: 2212.10773 [cs.CL].

[27] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, *Minigpt-4: Enhancing vision-language understanding with advanced large language models*, 2023. arXiv: 2304.10592 [cs.CV].

[28] H. Liu, C. Li, Q. Wu, and Y. J. Lee, *Visual instruction tuning*, 2023. arXiv: 2304.08485 [cs.CV].

[29] L. Ouyang, J. Wu, X. Jiang, *et al.*, *Training language models to follow instructions with human feedback*, 2022. arXiv: 2203.02155 [cs.CL].

[30] Y. Wang, S. Mishra, P. Alipoormolabashi, *et al.*, *Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks*, 2022. arXiv: 2204.07705 [cs.CL].

[31] I. Radosavovic, R. P. Kosaraju, R. Girshick, K. He, and P. Dollár, *Designing network design spaces*, 2020. arXiv: 2003.13678 [cs.CV].

[32] B. Lin, Y. Ye, B. Zhu, *et al.*, *Video-llava: Learning united visual representation by alignment before projection*, 2024. arXiv: 2311.10122 [cs.CV].

[33] M. Maaz, H. Rasheed, S. Khan, and F. S. Khan, *Video-chatgpt: Towards detailed video understanding via large vision and language models*, 2024. arXiv: 2306.05424 [cs.CV].