This CVPR paper is the Open Access version, provided by the Computer Vision Foundation. Except for this watermark, it is identical to the accepted version; the final published version of the proceedings is available on IEEE Xplore.

Towards General Visual-Linguistic Face Forgery Detection

Ke Sun¹, Shen Chen², Taiping Yao², Ziyin Zhou¹, Jiayi Ji¹, Xiaoshuai Sun¹, Chia-Wen Lin³, Rongrong Ji¹ ¹ Key Laboratory of Multimedia Trusted Perception and Efficient Computing,

Ministry of Education of China, Xiamen University

² Youtu Lab, Tecent, China ³ National Tsing Hua University, Taiwan

Abstract

Face manipulation techniques have achieved significant advances, presenting serious challenges to security and social trust. Recent works demonstrate that leveraging multimodal models can enhance the generalization and interpretability of face forgery detection. However, existing annotation approaches, whether through human labeling or direct Multimodal Large Language Model (MLLM) generation, often suffer from hallucination issues, leading to inaccurate text descriptions, especially for high-quality forgeries. To address this, we propose Face Forgery Text Generator (FFTG), a novel annotation pipeline that generates accurate text descriptions by leveraging forgery masks for initial region and type identification, followed by a comprehensive prompting strategy to guide MLLMs in reducing hallucination. We validate our approach through fine-tuning both CLIP with a three-branch training framework combining unimodal and multimodal objectives, and MLLMs with our structured annotations. Experimental results demonstrate that our method not only achieves more accurate annotations with higher region identification accuracy, but also leads to improvements in model performance across various forgery detection benchmarks. Our Codes are available in https://github.com/skJack/VLFFD.git.

1. Introduction

Face manipulation techniques have achieved remarkable progress in recent years, enabling high-quality modifications of facial attributes [10, 14, 20], expressions [33], and identities [24]. While these advances bring creative possibilities, they also raise serious concerns about potential misuse and social trust [51, 53]. To address these challenges, developing robust face forgery detection methods has become crucial, especially for handling unseen forgeries that exhibit significant domain gaps from training



Figure 1. Differences between annotations generated by human annotation [62], GPT-40 methods and ours for a fake image. The fake image is manipulated only on the mouth region, and the forgery mask is generated by comparing the difference between real and fake images. (Best viewed in color.)

data [30, 35, 44, 46, 54, 61].

Most existing face forgery detection methods rely on unimodal architectures, which often lack interpretability and generalization. Recent advancements in visuallanguage multimodal learning, such as CLIP [40] and multimodal large language models (LLMs), have demonstrated powerful representation learning capabilities for both visual and language tasks. These models create a bridge between vision and language, improving human understanding of visual tasks and enhancing model performance through language-guided learning. For face forgery detection, incorporating language modality could provide interpretable explanations and tap into the rich semantic knowledge embedded in multimodal models [4, 17, 21, 34, 36, 56, 62].

To leverage these powerful multimodal models for forgery detection, high-quality text annotations are essential. However, obtaining accurate text annotations for face forgery data remains challenging. Current approaches for obtaining text annotations primarily fall into two categories:

^{*}Corresponding author

Human Annotation [62], where annotators manually identify forgery regions and provide explanations, and *MLLM Annotation* [21], where prompts are crafted to enable multimodal large language models (e.g., GPT-40) to generate annotations. However, we have observed that both approaches suffer from *hallucination issues*, especially for high-quality forged faces. For instance. as shown in Figure 1, we visualize the annotations of NeuralTexture forgeries in the FFpp [41] dataset produced by DD-VQA [62] and GPT-40. The forgery is limited to the mouth region, while other regions are authentic. Both human and MLLM annotations incorrectly mark the nose area, which remains unaltered in the forged image. Such annotation errors impact the performance and interpretability of downstream tasks.

To address these challenges, we propose a data annotation pipeline called Face Forgery Text Generator (FFTG), which mitigates hallucination by incorporating accurate forgery region localization and type identification as concrete guidance for text generation. Specifically, FFTG first generates forgery maps by comparing real and forged images, assesses the forgery degree of each facial component, and uses handcrafted features to estimate forgery types, combining these elements into a raw annotation. We then design a comprehensive prompting strategy to guide multimodal large language models (e.g., GPT-40 mini) in generating accurate annotations. Our strategy consists of 1) paired real-fake images as visual prompts enabling the model to identify differences through comparison, 2) guide prompts containing the raw annotation and its derivation process to reduce hallucination, 3) task description prompts that guide the model to perform step-by-step analysis through chain-of-thought reasoning, and 4) predefined prompts that standardize output format and provide additional guidelines. As shown in Figure 1, this carefully designed pipeline produces more accurate and diverse annotations compared to existing methods.

We validate the effectiveness of FFTG-generated annotations by fine-tuning both CLIP and multimodal LLMs (e.g., LLaVA). For the CLIP model, we adopt a multimodal joint training approach, aligning and integrating the text and visual modalities to assist classification, allowing the text to better guide the visual encoder. Experimental results demonstrate that FFTG annotations enable better generalization performance compared to traditional methods when fine-tuning CLIP. For multimodal LLMs, our annotations not only provide better interpretability but also achieve higher accuracy compared to human annotations and direct GPT labeling. This indicates that the detailed and structured prompts in FFTG reduce annotation errors, resulting in improved model performance across various metrics.

Our main contributions can be summarized as follows:

• We identify a fundamental challenge in visual-linguistic forgery detection: obtaining accurate text annotations that

align with forgery masks.

- We propose FFTG, a novel annotation pipeline that leverages forgery masks to generate accurate and diverse text annotations for deepfake images.
- We demonstrate the effectiveness of our annotations through extensive experiments with CLIP and MLLM, showing improved generalization and interpretability.

2. Related Work

2.1. General Face Forgery Detection

General face forgery detection focuses on improving model generalization to unseen domains, which remains a critical challenge in this field [58, 60]. Existing approaches mainly fall into two categories: forgery simulation and framework engineering. The former simulates various forgery traces through data augmentation, including blending artifacts [7, 15, 28, 43], facial inconsistencies [2, 23, 47, 49, 59, 64], and subtle distortions [25, 38]. The latter enhances network architectures through attention mechanisms [6, 42, 45, 55, 63], frequency-spatial modeling [31, 37, 39], or implicit identity modeling [8, 12, 19]. Recent works also explore local-global relationships [15, 16, 52] and feature disentanglement [5, 13, 18, 48, 57] for better generalization.

2.2. Visual-Language Learning on FFD

The visual-language pretraining paradigm, such as CLIP [40] through multimodal contrastive learning, has recently been extended to face forgery detection. Early attempts like DD-VQA [62] utilized crowdsourcing platforms to collect human annotations for deepfake data and fine-tuned multimodal models like BLIP [27]. With the advancement of multimodal large language models (MLLMs), researchers began exploring their capabilities in forgery detection. Jia et al. [22] first investigated GPT's ability in detecting manipulated faces, while FFAA [21] leveraged GPT-40 for annotation generation and model fine-tuning. X2DFD [4] further proposed a selfenhancement approach for improving MLLM performance in forgery detection. However, the effectiveness of these methods heavily relies on annotation quality. Our work addresses this fundamental challenge by providing a more accurate annotation pipeline that leverages concrete visual evidence to guide text generation.

3. Face Forgery Text Generator

In this section, we introduce our proposed FFTG pipeline, which comprises the **Raw Annotation Generation (RAG)** and **Annotation Refinement**. The goal of RAG is to provide an initial annotation using handcrafted criteria and accurate forged images. Although the generated annotations are limited in diversity and have a relatively fixed structure,



Figure 2. Overall framework of the Face Forgery Text Generator (FFTG). The paired forgery and real image are first fed into the Mask Generation module to generate forgery mask M. Then the Forgery Region Extraction module extracts the selected region R_s . Subsequently, the Forgery Type Decision module decides the forgery type and generates raw annotation. Then the final annotation is generated by GPT with several prompts.



Figure 3. Five typical types of forgery faces. (a) Color Difference. (b) Blur. (c) Structure Abnormal. (d) Texture Abnormal. (e) Blend Boundary. The red circle highlights the region of each forgery type. (Best viewed in color.)

they are highly accurate and reasonable, which helps to reduce the hallucinations that may occur when using large language models for annotation. Annotation Refinement with MLLM leverages advanced multimodal large language models (e.g., GPT-40-mini) to further refine the annotations. To increase diversity and improve accuracy, we employ four types of prompts to guide the large model in this process. The overall framework is shown in Figure 2.

3.1. Raw Annotation Generation

Given a real image $i_r \in \mathbb{R}^{3 \times H \times W}$ and its corresponding forgery image $i_f \in \mathbb{R}^{3 \times H \times W}$, RAG encompasses the following steps:

Mask Generation. To locate the forgery region, similar to [3], we first construct manipulated mask M by computing the absolute pixel-wise difference in the RGB channels, and then normalizing it to the range of [0, 1]:

$$M = |i_r - i_f|/255.$$
(1)

Forgery Region Extraction. This step aims to select a forgery region containing i_f . Facial images are divided into

four areas: mouth, nose, eyes, and face, based on landmarks. We compute the average value of M in each area and set a threshold θ to form the forgery region list L_f . This is defined as:

$$\frac{1}{|R_t|} \sum_{j \in R_t} M(j) > \theta, R_t \to L_f,$$
(2)

where R_t represents one of the four predefined areas, and $|R_t|$ is the sum of pixels in area t. If the value exceeds θ , the corresponding area is included in L_f . After processing all four areas, we randomly select one region R_s from L_f for the next step. $R_s(i_r)$ and $R_s(i_f)$ represent the forgery regions for real and fake pixels, respectively.

Forgery Type Decision. The goal of this step is to determine the type of forgery via a specially designed criterion. According to the previous work and our observation, we categorize the existing forgery types as color difference, blur, structure abnormal, texture abnormal, and blend boundary as shown in Figure 3. We detail each forgery type and corresponding evaluation standard as follows: 1) Color Difference : Occurs in face swaps with notable color variance. We assess this using the distance of average channelwise mean and variance in Lab color space between real and fake regions. 2) Blur: We use the Laplacian operator to quantify local blurring in forgery faces, determining blurriness by the variance after applying the operator to real and fake images in the selected region. 3) Structure Abnormal: Observed deformations in fake face organs are assessed using the SSIM index difference between real and fake images in the selected region R_s . 4) Texture Abnormal: We measure texture clarity using the contrast of the Gray-Level Co-occurrence Matrix (GLCM), defining an area as texture abnormal when the real region's C_d exceeds that of the fake beyond a threshold. 5) Blend Bound**ary**: Existing face manipulation methods conduct blending operation to transfer an altered face into an existing background, which leaves intrinsic cues across the blending boundaries [28], such as the red circle of Figure 3(e). We assess the presence of blending artifacts by analyzing three characteristics in the selected region's boundary: gradient variations, edge transitions, and frequency domain changes. If at least two of these metrics exceed their respective thresholds, we classify the region as having significant blending boundaries.

Supplementary materials provide detailed pseudocodes for each criterion. The identified regions and types are then transformed into natural language expressions using GPT-40 generated descriptive phrases. For instance, "Texture Abnormal" becomes "lacks natural texture" and "Color Difference" translates to "has inconsistent colors". A complete list of these mappings is provided in the supplementary materials. This translation ensures our raw annotations are both technically accurate and linguistically natural, facilitating subsequent refinement by MLLMs.

3.2. Annotation Refinement with MLLM

While our mask-guided analysis provides accurate region localization, the handcrafted features may not fully capture all forgery types, and the generated descriptions lack linguistic diversity. To address these limitations, we leverage GPT-40 mini's strong visual understanding capabilities for refined annotation generation. To ensure both accuracy and diversity while avoiding hallucination, we design a comprehensive prompting strategy with four key components:

Visual Prompt: Instead of presenting single images, we concatenate the real and forged face images as paired inputs to the MLLM. This comparative approach serves two purposes: 1) enables the model to identify forgery artifacts through direct comparison, reducing hallucination by providing explicit visual references, and 2) helps generate more focused annotations for real images by maintaining the forgery detection perspective, avoiding irrelevant descriptions that might emerge from isolated real image.

Guide Prompt: We incorporate the previously generated raw annotations into this component, along with detailed explanations of how each forgery type was determined. For example, we explain how texture abnormalities were identified using GLCM analysis and how structural deformations were determined through SSIM comparisons.

Task Description Prompt: Clear instructions establish an expert forgery detection context, defining specific requirements for analyzing visual evidence and generating comprehensive descriptions of manipulation artifacts.

Pre-defined Prompt: Structured output requirements specify JSON format, mandatory phrases ("This is a real/fake face"), and caption counts for consistent annotation generation, ensuring standardized outputs for downstream tasks.



Figure 4. Overview of our fine-tuning strategies. (a) For multimodal models like CLIP, we employ three training objectives: direct image classification, feature alignment between modalities, and multimodal fusion classification. (b) For MLLM, we utilize our pre-trained image encoder and fine-tune the projector and LLM components.

This strategy enables the model to generate accurate and diverse annotations while maintaining consistency with technical analysis. Due to space limitations, we provide the complete prompt templates in the supplementary material.

4. Model Fine-tuning

To better validate the effectiveness of FFTG for face forgery detection, we provide two baseline approaches for utilizing our annotations, as illustrated in Figure 4. The first baseline focuses on fine-tuning multimodal models like CLIP through a three-branch learning framework that combines both unimodal and multimodal objectives. The second baseline explores enhancing multimodal large language model (MLLM), aiming to improve both their forgery detection accuracy and reasoning capabilities.

4.1. Finetune Multimodal Models

As shown in Figure 4 (a), multimodal models typically consist of two encoders: an image encoder E_i and a text encoder E_t , which extract visual features $v \in \mathbb{R}^{B \times D}$ and text features $l \in \mathbb{R}^{B \times D}$ respectively, where B denotes the batch size and D is the feature dimension. To effectively leverage our FFTG annotations and activate the pretrained knowledge for better forgery localization and type identification, we propose a three-branch training framework that combines unimodal and multimodal learning objectives:

Image Feature Classification. The visual features v extracted by the image encoder E_i are directly used for binary classification through a linear layer. The classification loss L_{img} is defined as:

$$L_{img} = -\frac{1}{B} \sum_{i=1}^{B} y_i \log(\operatorname{softmax}(W_i v_i + b_i)), \quad (3)$$

where $y_i \in \{0, 1\}$ denotes the binary label, and W_i, b_i are learnable parameters.

Multimodal Feature Alignment. To align visual and textual representations, we employ CLIP's contrastive learning between image features v and text features l defined as:

$$L_{align} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp(s(v_i, l_i)/\tau)}{\sum_{j=1}^{B} \exp(s(v_i, l_j)/\tau)}, \quad (4)$$

where $s(\cdot, \cdot)$ denotes normalized cosine similarity and τ is a temperature parameter controlling the sharpness of the distribution.

Multimodal Feature Classification. We fuse visual and textual features through cross-attention and feed the fused features into a classification head. The fusion classification loss L_{fusion} is:

$$L_{fusion} = -\frac{1}{B} \sum_{i=1}^{B} y_i \log(\operatorname{softmax}(W_f(v_i \otimes l_i) + b_f)),$$
⁽⁵⁾

where \otimes denotes cross-attention fusion, and W_f, b_f are learnable parameters.

The overall loss function is:

$$L = L_{img} + L_{align} + L_{fusion}.$$
 (6)

4.2. Finetune Multimodol Large Language Model

Recent advances in multimodal large language models (MLLMs) have demonstrated impressive capabilities in visual understanding and natural language reasoning. In addition to training visual encoders, we explore utilizing FFTG annotations to enhance the forgery detection capabilities of MLLM (e.g., LLaVA). These models typically consist of three components: a vision encoder, an alignment projector, and a large language model (LLM). In our approach, we leverage the pre-trained vision encoder from our previous step and focus on fine-tuning the alignment projector and LLM components, as shown in Figure 4 (b).

To evaluate the model's performance, we design a straightforward yet effective prompt template: "*Do you think this image is of a real face or a fake one?*" followed by "*Please provide your reasons.*". This two-part prompt structure encourages the model to not only make binary decisions but also provide interpretable explanations for its judgments, enabling us to assess both the accuracy and reasoning capabilities of the fine-tuned model.

5. Experiment

5.1. Experimental Setting

Dataset. We conduct experiments on five challenging datasets: FaceForensics++ [41], DFDC-P [11], DFD, Celeb-DF [29], and Wild-Deepfake [66]. FF++ provides

paired real-fake data for generating forgery masks, while others offer diverse forgery types and scenarios. Face detection is performed using DSFD [26]. Detailed dataset descriptions are provided in the supplementary material.

Annotation details. We use the open-source DLIB algorithm as the face landmark detector. For the forgery type decision, the threshold of mean and variance is 1.0 and 0.5. For the blur, the threshold is set to 100. If the difference of SSIM is larger than 0.97, we determine the forgery part is structure abnormal. The texture abnormal threshold is set to 0.7. The blending ratio α is set to 0.9. For generating refined annotations, we utilize GPT-40-mini as our multimodal language model annotator. To create a diverse yet manageable dataset from FaceForensics++, we sample 3 frames at regular intervals from each video. During training, we use the temporally closest annotated frame as the ground truth label for intermediate frames.

Training details. For multimodal model fine-tuning, we use CLIP with ViT-L as the image encoder. Input images are resized to 224×224 pixels. The model is optimized using Adam optimizer with a learning rate of 1e - 6 and batch size of 32. For MLLM fine-tuning, we use LLaVA 1.5-7b [32] as our fundation model. we set the learning rate to 2e - 5, batch size to 8, gradient accumulation step to 1, and train for 3 epochs. All experiments are implemented in PyTorch and conducted on $4 \times$ NVIDIA A100 GPUs.

5.2. Experimental Results on FFTG

To evaluate the quality and effectiveness of our FFTG annotations, we compare against three baseline approaches. The first baseline (*w/o text*) trains the model without any textual annotations, serving as a unimodal baseline. The second baseline uses human-annotated text from DD-VQA [62], representing the traditional manual annotation approach. The third baseline employs GPT-4o-mini directly for annotation without our raw description guidance, demonstrating the impact of our structured prompting strategy. The experimental results are shown in Table 1.

We conduct evaluations across three dimensions:

(1) Annotation Evaluation: Using forgery masks as ground truth, we evaluate whether generated annotations correctly identify manipulated regions (mouth, nose, eyes, face) by checking for exact terms or synonyms, measured by precision, recall, and F1-score.

(2) *CLIP Evaluation*: We evaluate the classification performance using AUC and EER metrics from the **Image Feature Classification** branch output. We report the average metrics across five benchmark datasets to evaluate forgery detection performance.

(3) *MLLM Evaluation*: We evaluate MLLM on classification and explanation. For classification, we compute accuracy by matching the occurrence of "real" or "fake" in the model's response with ground truth labels. For explanation

Method	Annotation Evaluation			CLIP Evaluation		MLLM Evaluation			
	Precision	Recall	F1	AVG-AUC	AVG-EER	FFpp-ACC	CDF-ACC	Precision	Recall
w/o Text	-	-	-	84.36	20.64	50.13	65.30	10.41	8.10
DD-VQA (Human)	62.46	51.52	52.06	88.25	18.04	73.54	65.60	62.94	53.62
GPT-4o-mini	61.27	44.00	47.18	87.56	19.21	94.84	73.98	58.26	41.85
FFTG	89.48	57.12	64.96	89.08	17.61	95.84	75.00	88.07	55.30

Table 1. Comparison of different annotation approaches. We report precision, recall and F1-score for annotation quality evaluation, AUC and EER for CLIP-based forgery detection and classification accuracy (ACC) and explanation quality (Precision/Recall) for mLLM evaluation on FFpp and Celeb-DF (CDF) datasets.

Method	FF++		D	FD	DFDC-P		Wild Deepfake		Celeb-DF	
	AUC	EER	AUC	EER	AUC	EER	AUC	EER	AUC	EER
Xception [9]	99.09	3.77	87.86	21.04	69.80	35.41	66.17	40.14	65.27	38.77
EN-b4 [50]	99.22	3.36	87.37	21.99	70.12	34.54	61.04	45.34	68.52	35.61
Face X-ray [50]	87.40	-	85.60	-	70.00	-	-	-	74.20	-
F3-Net [39]	98.10	3.58	86.10	26.17	72.88	33.38	67.71	40.17	71.21	34.03
MAT [63]	99.27	3.35	87.58	21.73	67.34	38.31	70.15	36.53	70.65	35.83
GFF [37]	98.36	3.85	85.51	25.64	71.58	34.77	66.51	41.52	75.31	32.48
LTW [44]	99.17	3.32	88.56	20.57	74.58	33.81	67.12	39.22	77.14	29.34
LRL [3]	99.46	3.01	89.24	20.32	76.53	32.41	68.76	37.50	78.26	29.67
DCL [46]	99.30	3.26	91.66	16.63	76.71	31.97	71.14	36.17	82.30	26.53
PCL+I2G [64]	99.11	-	-	-	-	-	-	-	81.80	-
SBI [43]	88.33	20.47	88.13	17.25	76.53	30.22	68.22	38.11	80.76	26.97
UIA-ViT [65]	-	-	94.68	-	75.80	-	-	-	82.41	-
RECCE [1]	99.32	3.38	89.91	19.95	75.88	32.41	67.93	39.82	70.50	35.34
UCF [57]	97.05	-	80.74	-	75.94	-	-	-	75.27	-
CLIP [40]	99.09	3.16	89.03	17.13	78.83	28.95	77.71	30.38	77.16	29.30
Ours	99.16	3.11	94.81	15.22	83.21	22.43	85.10	23.65	83.15	23.66

Table 2. **Frame-level** cross-database evaluation from FF++(HQ) to DFD, DFDC-P, Wild Deepfake and Celeb-DF in terms of AUC and EER. The FF++ belongs to the intra-domain results while others represent the unseen-domain.

quality, we assess the accuracy of identified forgery regions following the same protocol as Annotation Evaluation.

Annotation Evaluation. As shown in Table 1, our FFTG significantly outperforms existing annotation methods in identifying forgery regions. FFTG achieves 89.48% precision and 57.12% recall, surpassing human annotations (DD-VQA) by considerable margins (27.02% and 5.60% respectively). Compared to direct GPT-40 mini annotation without guidance, FFTG improves precision by 28.21% and recall by 17.12%, resulting in a substantially higher F1-score (64.96% vs 47.18%). These results demonstrate that our mask-guided annotation pipeline with structured prompting effectively reduces hallucination and generates more accurate region identifications than both human annotations and direct large model outputs.

CLIP Evaluation. The CLIP evaluation results demonstrate the effectiveness of incorporating textual modality and our training framework. The baseline method (w/o text), which relies solely on image features for binary classification, achieves an average AUC of 84.36% and EER of 20.64%. All methods with textual annotations outperform this unimodal baseline, validating the benefit of leveraging language modality to activate CLIP's pretrained knowledge. Among them, our FFTG achieves the best performance with 89.08% AUC and 17.61% EER, surpassing both human annotations (DD-VQA) and direct GPT-40 mini outputs by significant margins. This improvement demonstrates that high-quality text annotations, combined with our three-branch training framework, can effectively leverage the semantic knowledge embedded in CLIP model and en-

Siginal	Method	Cele	b-DF	DFDC-P		
Sigiliai	Wiethou	AUC	EER	AUC	EER	
Mask	Decoder	77.70	29.56	78.51	29.59	
Digital	Region	79.73	29.24	78.59	29.07	
	Туре	78.45	29.95	78.17	29.92	
	Both	77.18	30.47	77.54	31.03	
Text	Region	81.53	25.11	80.17	26.35	
	Туре	80.40	27.11	78.25	28.19	
	Both (Raw)	82.15	24.58	81.55	24.12	
	Ours	83.15	23.66	83.21	22.43	

Table 3. Ablation study on different supervisory signals.

hance the model's forgery detection capabilities.

MLLM Evaluation. For MLLM evaluation, we assess both the classification accuracy and explanation quality of fine-tuned models. In terms of classification, our FFTGenhanced model achieves the highest accuracy of 95.84% on FFpp (intra-domain) and 75.00% on Celeb-DF (crossdomain), significantly outperforming the baseline without text (50.13% and 65.30%). Notably, while DD-VQA annotations show moderate improvement (73.54% on FFpp), and direct GPT-4o-mini annotations achieve competitive accuracy (94.84% on FFpp), our method consistently performs better across different datasets, demonstrating more robust generalization.

For explanation quality, FFTG generates more accurate forgery region identifications with 88.07% precision and 55.30% recall, substantially surpassing both human annotations and direct GPT-40-mini outputs. These results validate that our structured prompting strategy not only improves the model's classification capability but also enhances its ability to provide accurate and reliable explanations for its decisions, which is crucial for practical applications requiring interpretable outputs.

5.3. Comparison with State-of-the-Art Methods

Cross-dataset evaluation. To evaluate the generalization capability of our fine-tuned CLIP model, we conduct extensive experiments across multiple deepfake datasets. Following standard protocol, we train our model on the high-quality version of FF++ and test on other challenging datasets that exhibit significant domain gaps in terms of forgery types, human identities, video backgrounds, and image quality.

The quantitative results are shown in Table 2. Our method achieves consistent improvements across all unseen datasets. Specifically, on DFDC-P, our method achieves 83.21% AUC, surpassing the recent transformer-based method UIA-ViT (75.80%) by a significant margin of 7.41%. On the challenging Wild Deepfake dataset,

Alignment	Multimodal	Cele	b-DF	Wild Deepfake		
	Withhodai	AUC	EER	AUC	EER	
×	×	77.16	29.30	77.71	30.38	
\checkmark	×	82.19	24.76	82.25	26.77	
×	\checkmark	81.66	24.31	80.35	28.13	
\checkmark	\checkmark	83.15	23.66	85.10	23.65	

Table 4. Ablation study on the impact of different components in terms of AUC and EER. 'Alignment' indicates the Multimodal Feature Alignment (L_{align}). 'Multimodal' signifies the Multimodal Feature Classification (L_{fusion}).

our approach reaches 85.10% AUC, outperforming DCL by nearly 14%. For Celeb-DF, we achieved 83.15% AUC, demonstrating superior performance compared to both traditional methods and recent advances like PCL+I2G (81.80%). These substantial improvements can be attributed to two key factors: 1) the high-quality text annotations from FFTG that help activate CLIP's pretrained knowledge, and 2) our effective three-branch training framework that facilitates both unimodal and multimodal feature learning.

5.4. Ablation Study

Impact of language information. To investigate the effectiveness of different supervisory signals, we compare three approaches: mask-based, digital label-based, and text-based supervision. For mask-based supervision, we employ a decoder to regress forgery masks. For digital supervision, we experiment with region-only, type-only, and both labels as classification targets. As shown in Table 3, mask-based supervision achieves limited performance (77.70% AUC on Celeb-DF), likely due to overfitting to low-level features. Digital supervision performs slightly better, with region-based classification reaching 79.73% AUC.

Our text-based approach significantly outperforms both alternatives, achieving 83.15% AUC on Celeb-DF and 83.21% on DFDC-P. This improvement can be attributed to two factors: 1) the rich semantic information captured by textual descriptions compared to binary or categorical labels, and 2) our FFTG pipeline that generates accurate and diverse annotations. The results also show that using both region and type information (Raw Annotation) performs better than using either alone, demonstrating the benefit of comprehensive text descriptions in guiding the model's learning process.

Impact of different components. To analyze the effectiveness of our three-branch training framework for fintuning multimodal model, we conduct ablation studies on two key components: Multimodal Feature Alignment and Multimodal Feature Classification. As shown in Table 4, using only Image Feature Classification achieves baseline perfor-



Figure 5. Visualization of FFTG annotation pipeline and model inference results. For each example, we show the fake-real image pair, forgery mask, FFTG's annotation, CLIP attention map, and LLaVA's output. FFTG annotations align well with forgery masks and guide both CLIP and LLaVA to focus on genuine manipulation regions.

mance (77.16% AUC on Celeb-DF). Adding Multimodal Feature Alignment improves the AUC by 4.5%, demonstrating the benefit of aligning visual and textual representations. Incorporating Multimodal Feature Classification further boosts performance by leveraging cross-attention fusion. The full model combining all three components achieves the best results (83.15% AUC on Celeb-DF and 85.10% AUC on Wild Deepfake), indicating that the different components complement each other in learning discriminative features for forgery detection.

5.5. Visualizations

Figure 5 showcases our analysis pipeline on FFpp test set examples. FFTG annotations demonstrate high accuracy in identifying manipulated regions and describing forgery types. For instance, in the first example, our annotation captures both eyes and mouth regions distortion along with blending artifacts at boundaries, precisely matching the forgery mask. In the second case, the annotation identifies mouth region's unnatural texture and nose's unrealistic appearance, while also noting blending artifacts around eye regions. The third example shows detailed description of texture abnormalities around eyes and nose, while the last example accurately captures the mouth region's structural distortion and texture inconsistencies. The attention maps from our fine-tuned CLIP model exhibit strong alignment with forgery masks, particularly evident in high-attention areas matching manipulated regions. For example, in the second row, CLIP's attention clearly highlights both the nose and mouth regions identified in the forgery mask. Similarly, LLaVA outputs demonstrate enhanced detection capabilities after fine-tuning, providing precise and consistent explanations. In the third example, LLaVA correctly identifies both the "blurry" appearance of eyes and the distorted nose with unusual texture patterns, showing a correlation with FFTG annotations and forgery masks.

6. Conclusion

In this paper, we analyze the limitations of existing text annotation approaches and present Face Forgery Text Generator (FFTG), a novel annotation pipeline that combines mask-guided analysis with structured prompting strategies to generate accurate and interpretable text descriptions for face forgery detection. Our extensive experiments demonstrate that FFTG effectively addresses the hallucination issues in existing annotation methods, achieving higher accuracy in region identification and leading to substantial improvements when fine-tuning both CLIP and MLLM. These results validate the importance of high-quality text annotations in enhancing both the generalization and interpretability of forgery detection systems, providing a promising direction for future research in multimodal forensics tasks.

7. Acknowledgment

This work was supported by National Key R&D Program of China (No.2023YFB4502804), the National Science Fund for Distinguished Young Scholars (No.62025603), the National Natural Science Foundation of China (No. U22B2051, No. 62302411), the Natural Science Foundation of Fujian Province of China (No.2021J06003), and China Postdoctoral Science Foundation (No. 2023M732948).

References

- [1] Junyi Cao, Chao Ma, Taiping Yao, Shen Chen, Shouhong Ding, and Xiaokang Yang. End-to-end reconstructionclassification learning for face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4113–4122, 2022. 6
- [2] Liang Chen, Yong Zhang, Yibing Song, Lingqiao Liu, and Jue Wang. Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection. In *CVPR*, pages 18710–18719, 2022. 2
- [3] Shen Chen, Taiping Yao, Yang Chen, Shouhong Ding, Jilin Li, and Rongrong Ji. Local relation learning for face forgery detection. AAAI, 2021. 3, 6
- [4] Yize Chen, Zhiyuan Yan, Siwei Lyu, and Baoyuan Wu. A framework for explainable and etendable deepfake detection. *arXiv preprint arXiv:2410.06126*, 2024. 1, 2
- [5] Jikang Cheng, Zhiyuan Yan, Ying Zhang, Li Hao, Jiaxin Ai, Qin Zou, Chen Li, and Zhongyuan Wang. Stacking brick by brick: Aligned feature isolation for incremental face forgery detection. arXiv preprint arXiv:2411.11396, 2024. 2
- [6] Jikang Cheng, Zhiyuan Yan, Ying Zhang, Yuhao Luo, Zhongyuan Wang, and Chen Li. Can we leave deepfake data behind in training deepfake detector? *arXiv preprint arXiv:2408.17052*, 2024. 2
- [7] Jikang Cheng, Ying Zhang, Qin Zou, Zhiyuan Yan, Chao Liang, Zhongyuan Wang, and Chen Li. Ed4: Explicit data-level debiasing for deepfake detection. *arXiv preprint arXiv:2408.06779*, 2024. 2
- [8] Jongwook Choi, Taehoon Kim, Yonghyun Jeong, Seungryul Baek, and Jongwon Choi. Exploiting style latent flows for generalizing deepfake video detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1133–1143, 2024. 2
- [9] François Chollet. Xception: Deep learning with depthwise separable convolutions. In CVPR, pages 1251–1258, 2017.
 6
- [10] Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil K Jain. On the detection of digital face manipulation. In *CVPR*, pages 5781–5790, 2020. 1
- [11] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge dataset. arXiv preprint arXiv:2006.07397, 2020. 5
- [12] Shichao Dong, Jin Wang, Renhe Ji, Jiajun Liang, Haoqiang Fan, and Zheng Ge. Implicit identity leakage: The stumbling block to improving deepfake detection generalization.

In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3994–4004, 2023. 2

- [13] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Ting Zhang, Weiming Zhang, Nenghai Yu, Dong Chen, Fang Wen, and Baining Guo. Protecting celebrities from deepfake with identity consistency transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9468–9478, 2022. 2
- [14] Ester Gonzalez-Sosa, Julian Fierrez, Ruben Vera-Rodriguez, and Fernando Alonso-Fernandez. Facial soft biometrics for recognition in the wild: Recent works, annotation, and cots evaluation. *IEEE Transactions on Information Forensics and Security*, 13(8):2001–2014, 2018. 1
- [15] Fabrizio Guillaro, Davide Cozzolino, Avneesh Sud, Nicholas Dufour, and Luisa Verdoliva. Trufor: Leveraging all-round clues for trustworthy image forgery detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20606–20615, 2023. 2
- [16] Xiao Guo, Vishal Asnani, Sijia Liu, and Xiaoming Liu. Tracing hyperparameter dependencies for model parsing via learnable graph pooling network. *Advances in Neural Information Processing Systems*, 37:116899–116932, 2024. 2
- [17] Xiao Guo, Xiaohong Liu, Iacopo Masi, and Xiaoming Liu. Language-guided hierarchical fine-grained image forgery detection and localization. *International Journal of Computer Vision*, pages 1–22, 2024. 1
- [18] Cheng-Yao Hong, Yen-Chi Hsu, and Tyng-Luh Liu. Contrastive learning for deepfake classification and localization via multi-label ranking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17627–17637, 2024. 2
- [19] Baojin Huang, Zhongyuan Wang, Jifan Yang, Jiaxin Ai, Qin Zou, Qian Wang, and Dengpan Ye. Implicit identity driven deepfake face swapping detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4490–4499, 2023. 2
- [20] Yihao Huang, Felix Juefei-Xu, Qing Guo, Yang Liu, and Geguang Pu. Fakelocator: Robust localization of ganbased face manipulations. *IEEE Transactions on Information Forensics and Security*, 17:2657–2672, 2022. 1
- [21] Zhengchao Huang, Bin Xia, Zicheng Lin, Zhun Mou, and Wenming Yang. Ffaa: Multimodal large language model based explainable open-world face forgery analysis assistant. arXiv preprint arXiv:2408.10072, 2024. 1, 2
- [22] Shan Jia, Reilin Lyu, Kangran Zhao, Yize Chen, Zhiyuan Yan, Yan Ju, Chuanbo Hu, Xin Li, Baoyuan Wu, and Siwei Lyu. Can chatgpt detect deepfakes? a study of using multimodal large language models for media forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4324–4333, 2024. 2
- [23] Chenqi Kong, Anwei Luo, Shiqi Wang, Haoliang Li, Anderson Rocha, and Alex C Kot. Pixel-inconsistency modeling for image manipulation localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 2
- [24] Pavel Korshunov and Sébastien Marcel. Deepfakes: a new threat to face recognition? assessment and detection. *arXiv* preprint arXiv:1812.08685, 2018. 1

- [25] Nicolas Larue, Ngoc-Son Vu, Vitomir Struc, Peter Peer, and Vassilis Christophides. Seeable: Soft discrepancies and bounded contrastive learning for exposing deepfakes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21011–21021, 2023. 2
- [26] Jian Li, Yabiao Wang, Changan Wang, Ying Tai, Jianjun Qian, Jian Yang, Chengjie Wang, Jilin Li, and Feiyue Huang. Dsfd: dual shot face detector. In *CVPR*, pages 5060–5069, 2019. 5
- [27] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 2
- [28] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *CVPR*, pages 5001–5010, 2020. 2, 4
- [29] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A new dataset for deepfake forensics. arXiv preprint arXiv:1909.12962, 2019. 5
- [30] Li Lin, Xinan He, Yan Ju, Xin Wang, Feng Ding, and Shu Hu. Preserving fairness generalization in deepfake detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 16815–16825, 2024.
- [31] Honggu Liu, Xiaodan Li, Wenbo Zhou, Yuefeng Chen, Yuan He, Hui Xue, Weiming Zhang, and Nenghai Yu. Spatialphase shallow learning: rethinking face forgery detection in frequency domain. In *CVPR*, pages 772–781, 2021. 2
- [32] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. Advances in neural information processing systems, 36, 2024. 5
- [33] Ming Liu, Yukang Ding, Min Xia, Xiao Liu, Errui Ding, Wangmeng Zuo, and Shilei Wen. Stgan: A unified selective transfer network for arbitrary image attribute editing. In *CVPR*, pages 3673–3682, 2019. 1
- [34] Ping Liu, Qiqi Tao, and Joey Tianyi Zhou. Evolving from single-modal to multi-modal facial deepfake detection: A survey. arXiv preprint arXiv:2406.06965, 2024. 1
- [35] Anwei Luo, Chenqi Kong, Jiwu Huang, Yongjian Hu, Xiangui Kang, and Alex C Kot. Beyond the prior forgery knowledge: Mining critical clues for general face forgery detection. *IEEE Transactions on Information Forensics and Security*, 19:1168–1182, 2023. 1
- [36] Anwei Luo, Rizhao Cai, Chenqi Kong, Yakun Ju, Xiangui Kang, Jiwu Huang, and Alex C Kot Life. Forgery-aware adaptive learning with vision transformer for generalized face forgery detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 1
- [37] Yuchen Luo, Yong Zhang, Junchi Yan, and Wei Liu. Generalizing face forgery detection with high-frequency features. In *CVPR*, pages 16317–16326, 2021. 2, 6
- [38] Dat Nguyen, Nesryne Mejri, Inder Pal Singh, Polina Kuleshova, Marcella Astrid, Anis Kacem, Enjie Ghorbel, and Djamila Aouada. Laa-net: Localized artifact attention network for quality-agnostic and generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on*

Computer Vision and Pattern Recognition, pages 17395–17405, 2024. 2

- [39] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *ECCV*, pages 86–103. Springer, 2020. 2, 6
- [40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 1, 2, 6
- [41] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *ICCV*, pages 1–11, 2019. 2, 5
- [42] Rui Shao, Tianxing Wu, and Ziwei Liu. Detecting and grounding multi-modal media manipulation. In *Proceedings* of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6904–6913, 2023. 2
- [43] Kaede Shiohara and Toshihiko Yamasaki. Detecting deepfakes with self-blended images. In CVPR, pages 18720– 18729, 2022. 2, 6
- [44] Ke Sun, Hong Liu, Qixiang Ye, Jianzhuang Liu, Yue Gao, Ling Shao, and Rongrong Ji. Domain general face forgery detection by learning to weight. In AAAI, pages 2638–2646, 2021. 1, 6
- [45] Ke Sun, Hong Liu, Taiping Yao, Xiaoshuai Sun, Shen Chen, Shouhong Ding, and Rongrong Ji. An information theoretic approach for attention-driven face forgery detection. In *European Conference on Computer Vision*, pages 111–127. Springer, 2022. 2
- [46] Ke Sun, Taiping Yao, Shen Chen, Shouhong Ding, Rongrong Ji, et al. Dual contrastive learning for general face forgery detection. In AAAI, 2022. 1, 6
- [47] Ke Sun, Shen Chen, Taiping Yao, Hong Liu, Xiaoshuai Sun, Shouhong Ding, and Rongrong Ji. Diffusionfake: Enhancing generalization in deepfake detection via guided stable diffusion. arXiv preprint arXiv:2410.04372, 2024. 2
- [48] Ke Sun, Shen Chen, Taiping Yao, Xiaoshuai Sun, Shouhong Ding, and Rongrong Ji. Continual face forgery detection via historical distribution preserving. *International Journal of Computer Vision*, pages 1–18, 2024. 2
- [49] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28130–28139, 2024. 2
- [50] Mingxing Tan and Quoc V Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *ICML*, 2019. 6
- [51] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. Deepfakes and beyond: A survey of face manipulation and fake detection. arXiv preprint arXiv:2001.00179, 2020. 1
- [52] Yuan Wang, Kun Yu, Chen Chen, Xiyuan Hu, and Silong Peng. Dynamic graph learning with content-guided spatial-

frequency relation reasoning for deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7278–7287, 2023. 2

- [53] Ruiyang Xia, Dawei Zhou, Decheng Liu, Jie Li, Lin Yuan, Nannan Wang, and Xinbo Gao. Inspector for face forgery detection: Defending against adversarial attacks from coarse to fine. *IEEE Transactions on Image Processing*, 2024. 1
- [54] Ruiyang Xia, Dawei Zhou, Decheng Liu, Lin Yuan, Shuodi Wang, Jie Li, Nannan Wang, and Xinbo Gao. Advancing generalized deepfake detector with forgery perception guidance. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6676–6685, 2024. 1
- [55] Yuting Xu, Jian Liang, Gengyun Jia, Ziming Yang, Yanhao Zhang, and Ran He. Tall: Thumbnail layout for deepfake video detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22658–22668, 2023. 2
- [56] Zhiyuan Yan, Peng Sun, Yubo Lang, Shuo Du, Shanzhuo Zhang, Wei Wang, and Lei Liu. Multimodal graph learning for deepfake detection. arXiv preprint arXiv:2209.05419, 2022. 1
- [57] Zhiyuan Yan, Yong Zhang, Yanbo Fan, and Baoyuan Wu. Ucf: Uncovering common features for generalizable deepfake detection. arXiv preprint arXiv:2304.13949, 2023. 2, 6
- [58] Zhiyuan Yan, Yong Zhang, Xinhang Yuan, Siwei Lyu, and Baoyuan Wu. Deepfakebench: A comprehensive benchmark of deepfake detection. arXiv preprint arXiv:2307.01426, 2023. 2
- [59] Zhiyuan Yan, Yuhao Luo, Siwei Lyu, Qingshan Liu, and Baoyuan Wu. Transcending forgery specificity with latent space augmentation for generalizable deepfake detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8984–8994, 2024. 2
- [60] Zhiyuan Yan, Taiping Yao, Shen Chen, Yandan Zhao, Xinghe Fu, Junwei Zhu, Donghao Luo, Chengjie Wang, Shouhong Ding, Yunsheng Wu, et al. Df40: Toward next-generation deepfake detection. arXiv preprint arXiv:2406.13495, 2024. 2
- [61] Zhiyuan Yan, Yandan Zhao, Shen Chen, Mingyi Guo, Xinghe Fu, Taiping Yao, Shouhong Ding, and Li Yuan. Generalizing deepfake video detection with plug-and-play: Video-level blending and spatiotemporal adapter tuning. arXiv preprint arXiv:2408.17065, 2024. 1
- [62] Yue Zhang, Ben Colman, Ali Shahriyari, and Gaurav Bharaj. Common sense reasoning for deep fake detection. *ECCV*, 2024. 1, 2, 5
- [63] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-attentional deepfake detection. *CVPR*, 2021. 2, 6
- [64] Tianchen Zhao, Xiang Xu, Mingze Xu, Hui Ding, Yuanjun Xiong, and Wei Xia. Learning self-consistency for deepfake detection. In *CVPR*, pages 15023–15033, 2021. 2, 6
- [65] Wanyi Zhuang, Qi Chu, Zhentao Tan, Qiankun Liu, Haojie Yuan, Changtao Miao, Zixiang Luo, and Nenghai Yu. Uiavit: Unsupervised inconsistency-aware method based on vision transformer for face forgery detection. *ECCV*, 2022. 6

[66] Bojia Zi, Minghao Chang, Jingjing Chen, Xingjun Ma, and Yu-Gang Jiang. Wilddeepfake: A challenging real-world dataset for deepfake detection. In ACM MM, pages 2382– 2390, 2020. 5