

Generating Attribution Reports for Manipulated Facial Images: A Dataset and Baseline

Jingchun Lian¹, Lingyu Liu¹, Yaxiong Wang^{2,†}, Yujiao Wu³,
Lianwei Wu⁴, Li Zhu¹, Zhedong Zheng⁵

¹Xi'an Jiaotong University ²Hefei University of Technology
³CSIRO ⁴Northwestern Polytechnical University ⁵University of Macau

15829901729@stu.xjtu.edu.cn, wangyx@hfut.edu.cn

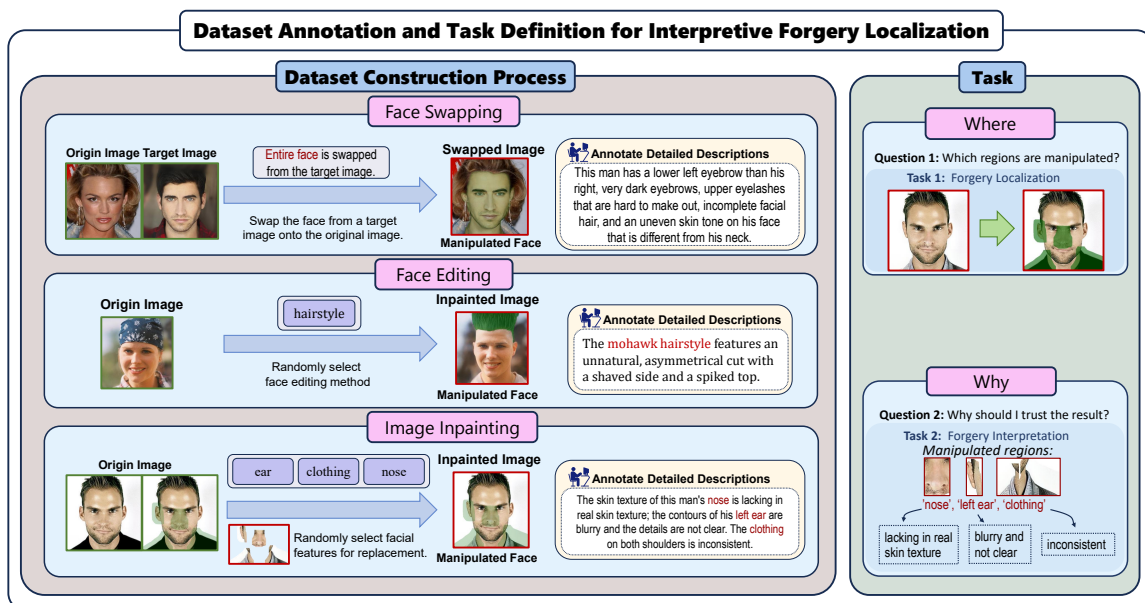


Figure 1: An overview of our proposed benchmark, illustrating the dataset construction process and the joint task definition. The left panel shows the 3 typical manipulation paradigms used for data generation, *i.e.*, Face Swapping, Face Editing, and Image Inpainting. The right panel defines the task of Joint Localization and Explanation, which requires models to answer both “where” a forgery is (Localization) and “why” it is a forgery (Explanation).

Abstract

Existing facial forgery detection methods typically focus on binary classification or pixel-level localization, providing little semantic insight into the nature of the manipulation. To address this, we introduce **Forgery Attribution Report Generation**, a new multimodal task designed to provide post-hoc forensic evidence for manipulated images. This task jointly localizes forged regions (“Where”) and generates natural language explanations grounded in the editing process (“Why”). This dual-focus approach goes beyond traditional binary forensics, providing a comprehensive, interpretable understanding of the manipulation. To enable research in this domain, we present **Multi-**

Modal Tamper Tracing (MMTT), a large-scale dataset of 152,217 samples. Each sample features a process-derived ground-truth mask and a human-authored textual description, ensuring high annotation precision and linguistic richness. We further propose **ForgeryTalker**, a unified end-to-end baseline that integrates vision and language via a shared encoder and dual decoders for mask and text generation. Experiments show that ForgeryTalker achieves competitive performance on both subtasks, *i.e.*, 59.3 CIDEr and 73.67 IoU, establishing a strong baseline for explainable multimedia forensics. Our dataset and code are available at: <https://github.com/NattyLianJc/Generating-Attribution-Reports>.

1 Introduction

The rapid evolution of advanced generative models, notably diffusion models (Ho et al., 2020; Song et al., 2020; Zhang et al., 2025), has significantly enhanced the realism of synthesized images. While promising for creative domains (Dhariwal and Nichol, 2021; Liu et al., 2025a), these technologies raise critical concerns regarding their misuse in misinformation and privacy violations (Rana et al., 2022; Liu et al., 2023b; Zhu et al., 2025b; Liu et al., 2025b; Ma et al., 2024; Zeng et al., 2024), particularly concerning facial manipulation. In response, detection research has rapidly shifted from binary classification to fine-grained forgery localization to address the growing complexity of modern attacks (Verdoliva, 2020; Rossler et al., 2019; Wu et al., 2023; Yu et al., 2021).

Unlike binary classifiers that merely output a simple decision, forgery localization aims to pinpoint specific tampered regions (Verdoliva, 2020). However, binary masks alone provide limited interpretability (Rossler et al., 2019). They treat all manipulated pixels equally, failing to differentiate between subtle and significant alterations or explain the underlying rationale. Furthermore, as modern forgeries become visually indistinguishable from reality, binary masks offer insufficient guidance for human reviewers. Subtle artifacts, such as minute distortions in facial features, are often overlooked, leaving observers without descriptive evidence to verify the detected anomalies and trust the recognition results.

To address these limitations, we introduce the novel task of **Forgery Attribution Report Generation**, aiming to produce a comprehensive report consisting of a pixel-level mask and a textual explanation. To support this, we construct the **Multi-Modal Tamper Tracing (MMTT)** dataset, the first large-scale benchmark with 152,217 samples. A key strength of MMTT lies in its high-quality ground truth: pixel-level masks are programmatically derived from the forgery process to ensure perfect alignment, while textual descriptions are crafted through a rigorous human-in-the-loop pipeline to capture subtle artifacts. Building on this, we propose **ForgeryTalker**, a unified baseline designed to generate these reports end-to-end. At its core, ForgeryTalker utilizes a shared encoder to learn a common, forgery-aware representation, forcing a deep fusion of visual and semantic features. This representation is processed by special-

ized dual decoders to concurrently generate the localization mask and the textual report, ensuring the explanation is semantically grounded in the visual evidence. Notably, our framework operates under the premise that the input image has already been flagged as suspicious by an upstream binary detector (Livernoche et al., 2025; Anan et al., 2025). By focusing exclusively on **post-hoc attribution** (localization and explanation), we shift the forensic objective from merely delivering a binary verdict’ to providing the interpretable evidence’ behind it. This specialized scope ensures that computational resources are dedicated to generating meaningful semantic insights for manipulated images rather than redundant processing of pristine ones. Our primary contributions are summarized as follows:

- **A New Task and Dataset.** We introduce *Forgery Attribution Report Generation*, a multimodal task combining forgery localization and natural language explanation. To support this, we present **Multi-Modal Tamper Tracing (MMTT)**, the first large-scale dataset with 152,217 samples, each annotated with a precise ground-truth mask from the editing process and a human-written attribution report.
- **A Unified and Effective Baseline.** We propose ForgeryTalker, a unified framework that jointly performs forgery localization and report generation. It is designed to facilitate coherent cross-modal reasoning through a shared encoder (image encoder + Q-former) and dual decoders (mask decoder and a Large Language Model).
- **Comprehensive Benchmarking.** We conduct extensive experiments on the proposed dataset for both report generation and forgery localization. The results validate that our baseline achieves competitive performance (59.3 CIDEr and 73.67 IoU) and demonstrates the complementary benefits of jointly addressing both tasks.

2 Multi-Modal Tamper Tracing Dataset

As a comparison with other major forgery datasets in Table 1 highlights, the proposed MMTT is the first to provide detailed textual explanations alongside forgery masks. The dataset contains **152,217** samples distributed across four manipulation paradigms, with diffusion-based inpainting and face swapping being the most prevalent (Figure 3a). Our statistical analysis reveals several key properties: (1) Eyebrows, eyes, and lips are the

Dataset	Tasks	Modality	Source Samples	Unique Forgeries	Manipulation Type	GT Type
FaceForensics++ (Rossler et al., 2019)	Class. / Seg.	Video	1,000	4,000	Multi-Face Mods	Label + Mask
Celeb-DF (Li et al., 2020)	Classification	Video	590	5,639	DeepFake	Image Label
DeeperForensics-1.0 (Jiang et al., 2020)	Classification	Video	50,000	10,000	GAN	Image Label
DFDC (Dolhansky et al., 2020)	Classification	Video	23,654	104,500	DeepFake	Image Label
FaceShifter (Li et al., 2019)	Classification	Video	N/A	5,000	GAN	Image Label
ForgeryNet (He et al., 2021)	Class. / Seg.	Image, Video	116,321	221,247	DeepFake, GAN	Label + Mask
OpenForensics (Le et al., 2021)	Detection / Seg.	Image, Video	45,473	70,325	GAN, Inpainting	BBox, Mask
DF40 (Yan et al., 2024)	Class. / Seg.	Image, Video	N/A	> 1,000,000	Multi-Face Mods	Label + Mask
DiffusionFace (Chen et al., 2024)	Generation	Image	N/A	50,000	Diffusion	Image Label
GenFace (Zhang et al., 2024a)	Generation	Image	10,000	10,000	GAN, Inpainting	Mask
MMTT (Ours)	Seg. / Caption	Text, Image	100,000	152,217	Face Swap, Inpainting, Attribute Edit	Text + Mask

Table 1: Comparison of Face Manipulation Datasets. Our MMTT dataset is highlighted and provides rich text annotations for the “why” problem, a unique feature among existing resources.

most common targets for manipulation across all localized editing methods (Figure 3b). (2) A significant portion of images feature multiple alterations, with 2-5 concurrent modifications being common (Figure 3c). (3) The textual annotations form a rich corpus of over 4 million words, with an average description length of **27.4 words**. The content of these descriptions aligns closely with the visual forgeries, frequently referencing the manipulated facial parts (Figure 3d). As shown in Figure 1, our MMTT dataset provides two complementary types of annotations: *binary forgery masks* in Section 2.1 and *forgery analysis text* in Section 2.2. The forgery analysis text primarily delivers diagnostic summaries of facial images, while the binary masks serve as auxiliary clues, highlighting localized forgery artifacts.

2.1 Forgery Generation

To construct a challenging and diverse dataset, we simulate forgery threats using three distinct manipulation paradigms. For each, we employed state-of-the-art models and developed specific procedures to programmatically generate forged images I_f and their corresponding pixel-perfect ground-truth masks M .

Source Image Collection. We first construct the MMTT dataset from 100,000 high-quality facial images, comprising 30,000 images from CelebAMask-HQ (Zhu et al., 2022) and 70,000 images from Flickr-Faces-HQ (FFHQ) (Karras et al., 2019). All images are resized to 512×512 pixels, which serve as the primary source for subsequent forgery manipulations.

Face Swapping. We used the GAN-based E4S (Abou Akar et al., 2024) model to swap faces between randomly paired images from our source datasets. Crucially, the E4S model automatically generates a precise binary mask M during this process, which directly serves as the ground-truth an-

notation for the manipulated region in the final forged image I_f .

Face Editing. We performed semantic alterations using GAN-inversion models StyleCLIP (Patashnik et al., 2021) and HFGI (Wang et al., 2022). The transformation is applied to an input image I to produce the forged image $I_f = \mathcal{E}_{\text{model}}(I, a)$, where \mathcal{E} is the editing function guided by attribute a , and $\text{model} \in \{\text{StyleCLIP}, \text{HFGI}\}$. The corresponding ground-truth mask, M_{final} , is constructed by taking the union of any pre-existing mask (M_{prev}) and a new semantic mask (M_{semantic}) generated via a face parsing model (Yu et al., 2018), formulated as $M_{\text{final}} = M_{\text{prev}} \cup M_{\text{semantic}}$.

Image Inpainting. We generated localized forgeries using both transformer-based (MAT (Li et al., 2022)) and diffusion-based (SDXL (Podell et al., 2023)) models. The required input masks (M) were created by programmatically selecting and merging facial component segments. The final inpainted image I_f is produced by composing the original image I with the model’s output I_g^{model} using the mask M : $I_f = (1 - M) \cdot I + M \cdot I_g^{\text{model}}$, where $\text{model} \in \{\text{MAT}, \text{SDXL}\}$.

2.2 Diagnosis Text Annotation

Annotation Methodology. To ensure high-quality annotations, we develop a structured pipeline (see Figure 2) where a team of expert annotators receives specific guidelines. Guided by a ground-truth mask for each image pair, annotators are instructed to describe visual inconsistencies or artifacts exclusively within the manipulated regions. They focus only on unnatural or poorly-integrated features, omitting descriptions of authentic areas, and compose self-contained descriptions that do not reference the original image. Each description is limited to a maximum of 120 words.

Annotation Process. Our annotation process involves 30 trained annotators who follow a three-

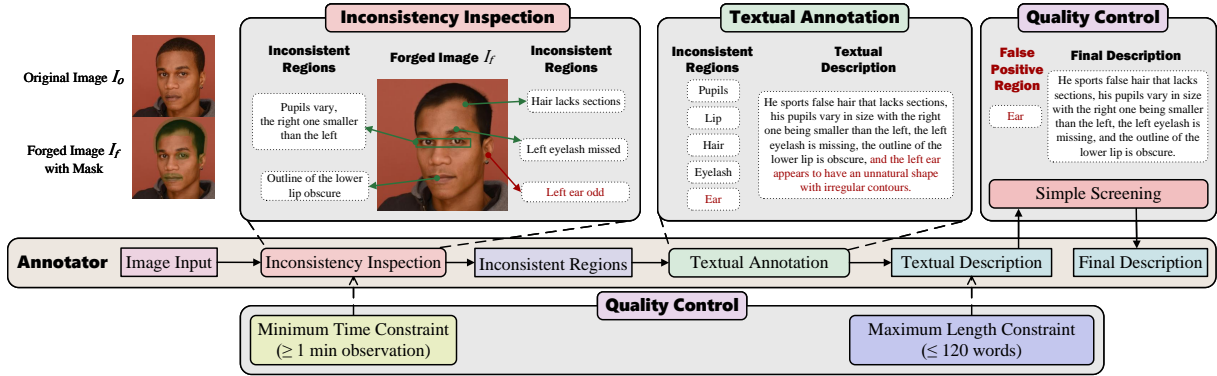


Figure 2: The manual annotation pipeline for the MMTT dataset. Here we show the key stages from **inconsistency inspection** to **textual description** and final **quality control**.

step procedure. First, annotators receive an original-forgery image pair (I_o, I_f) with its corresponding ground-truth mask M . They then inspect the images for inconsistencies within the masked facial regions, such as unnatural textures or asymmetries. Finally, they compose a detailed textual description T , explaining the specific nature of the alteration. This process culminates in a triplet $p = (I_f, M, T)$.

Annotation Quality Assurance. To ensure description reliability and inter-annotator consistency, we implement a rigorous human-in-the-loop quality control pipeline. All 30 annotators were recruited from a professional data annotation agency and provided with standardized guidelines and reference examples to maintain a uniform description style. During the process, we enforce a minimum observation time of one minute per image pair to ensure thorough examination. Furthermore, textual descriptions referencing regions outside the ground-truth mask are automatically flagged for review to prevent false positives. Finally, manual audits and cross-verifications were conducted on sampled batches to ensure that all textual descriptions are highly consistent and strictly grounded in the localized manipulation masks.

3 ForgeryTalker

The architecture of our baseline model, **ForgeryTalker**, extends InstructBlip (Dai et al., 2023) and is structured around a shared encoder and dual decoders. The shared encoder, consisting of a Vision Transformer and a Q-Former, processes the tampered image I to extract multimodal features. Guided by prompts from an integrated Forgery Prompter Network (FPN), these features are then passed to two decoders: a Mask Decoder for forgery localization and a Large Language Model (LLM) that generates the final

attribution report. As shown in Figure 4, training proceeds in two stages. In the **Forgery-aware Pretraining Stage**, we jointly optimize the core modules using a weighted combination of losses to build forgery-sensitive multimodal representations. In the subsequent **Attribution Report Generation Stage**, we first train the FPN to generate accurate region prompts. Then, with the FPN fixed, we fine-tune the mask decoder and Q-Former using segmentation and language modeling losses to improve forgery localization and the final attribution report.

3.1 Forgery-aware Pretraining

To learn robust multimodal representations sensitive to manipulation artifacts, we jointly optimize the proposed model using paired image I and text T via four complementary objectives:

Masked Language Modeling (\mathcal{L}_{mlm}). To enforce local visual-linguistic alignment, we mask a subset of region-related tokens in T (e.g., facial parts). The Q-Former is tasked with reconstructing these masked tokens conditioned on the image embeddings and learnable queries.

Language Modeling (\mathcal{L}_{lm}). We employ a standard generation objective where the T5-based decoder autoregressively generates the explanation text \hat{T} . This is supervised by maximizing the likelihood of the ground-truth tokens given the visual context.

Forgery Localization (\mathcal{L}_{seg}). To capture pixel-level anomalies, the mask decoder predicts a dense binary forgery map from the fused multimodal features. We apply a pixel-wise cross-entropy loss to align the prediction with the corresponding ground-truth manipulation mask.

Cross-model Alignment (\mathcal{L}_{con}). We further align the global image token and the mean-pooled text feature via contrastive learning. This objective pulls paired visual-text representations closer in

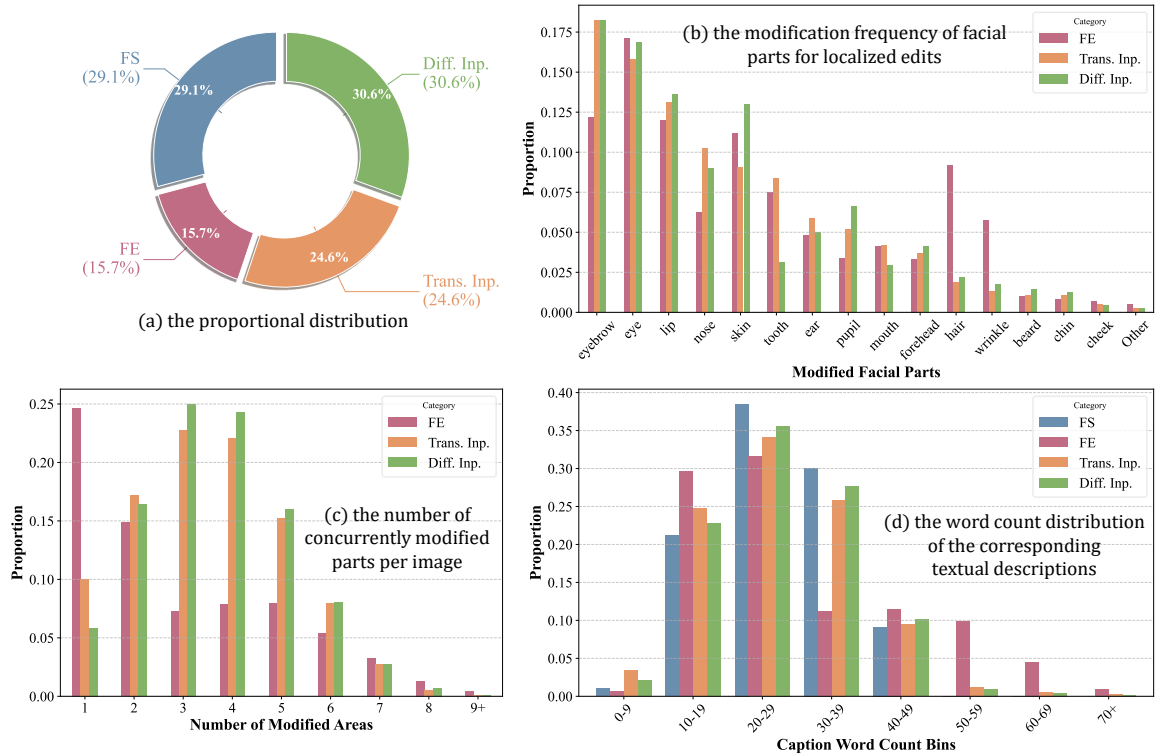


Figure 3: Statistical overview of the MMTT dataset and its four manipulation types: Face Swapping (FS), Face Editing (FE), Transformer-based Inpainting (Trans. Inp.), and Diffusion-based Inpainting (Diff. Inp.). The figure shows: (a) the proportional distribution of these types; (b) the modification frequency of facial parts for localized edits (excluding FS, which alters the entire face); (c) the number of concurrently modified parts per image; and (d) the word count distribution of the corresponding textual descriptions.

the latent space while pushing unpaired ones apart. The final overall objective is a weighted sum of these four complementary losses, equipping the model with both fine-grained localization capabilities and global semantic context.

3.2 Forgery Prompter Network

Motivation. Accurately identifying the most salient manipulated regions in forged images is challenging due to the high visual fidelity of modern manipulation techniques. Even human reviewers must closely inspect the images to detect inconsistencies. Therefore, we propose the Forgery Prompter Network (FPN) to generate an initial set of salient region keywords, which guide downstream reasoning and facilitate the coherent generation of attribution reports.

Region Keywords Extraction. In this step, we extract relevant region labels from the provided textual descriptions. The defined label space comprises 21 distinct facial semantics, where each image is associated with a 21-dimensional ground-truth vector Y ; here, the i -th element is 1 if the corresponding facial part is mentioned in the textual description, and 0 otherwise.

Forgery Prompter Network (FPN) takes the vi-

sion transformers as the main architecture. Considering the crucial role of fine-grained local context in identifying subtle flaws, we introduce a convolution branch at the early m layers to complement the global contexts captured by the vision transformer. As shown in Figure 5, the forgery image I concurrently traverses self-attention blocks and convolution blocks in parallel, producing global-aware features $F_g = \{F_g^0, F_g^2, \dots, F_g^{m-1}\}$ and local-aware features $F_l = \{F_l^0, F_l^2, \dots, F_l^{m-1}\}$. At each encoding level, the corresponding features are element-wise summed and fed into next attention block:

$$F_g^i = \text{MHA}_{i-1}(F_g^{i-1}), F_l^i = \text{Conv}_{i-1}(F_l^{i-1}), \quad (1)$$

$$F_g^i = \text{MHA}_i(F_g^i + F_l^i), \quad i = 1, \dots, m \quad (2)$$

where “MHA” and “Conv” mean the multi-head attention and convolution. Furthermore, we note that the positioning of facial regions in a natural image follows a rigid and predictable structure, with the eyes typically positioned laterally relative to the nose and the eyebrows aligned above the eyes. Leveraging this regularity, we integrate coordinate convolution (Liu et al., 2018) in the initial convolutional layer to detect anomalies in the arrangement of facial features, *i.e.*, $\text{Conv}_0 = \text{CoorConv}$. The

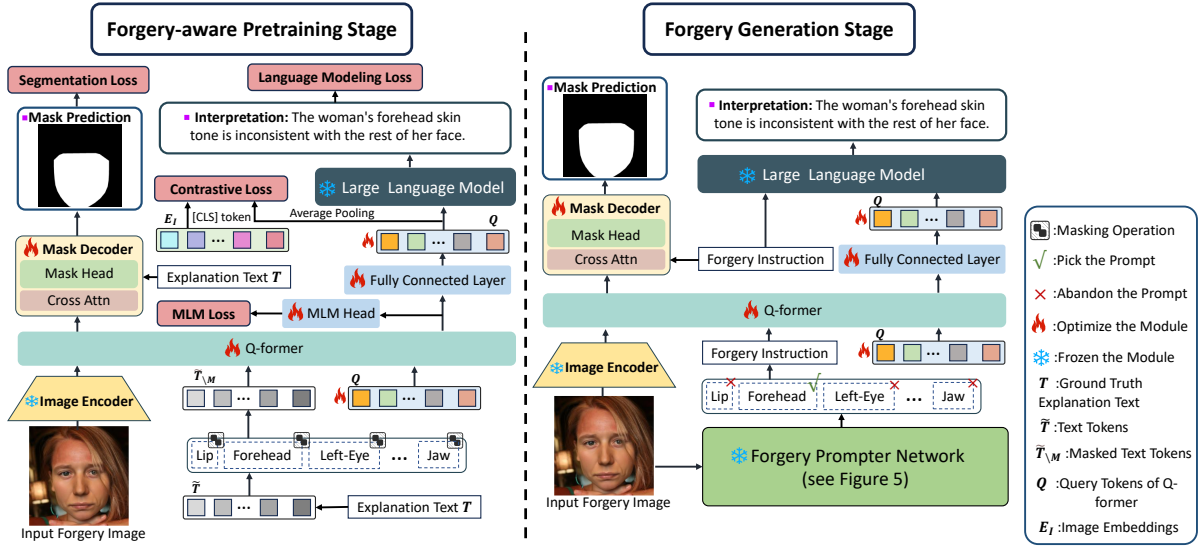


Figure 4: Illustration of our ForgeryTalker framework. The training pipeline has two stages. In the Forgery-aware Pretraining Stage, the Q-former, Mask Decoder, and Language Model are jointly optimized with MLM, language modeling, segmentation, and contrastive losses to build multimodal representations. In the Explanation Generation Stage, the FPN is trained with BCE and Dice losses for region classification and then frozen while the Q-former and Mask Decoder are fine-tuned for improved forgery localization and explanation. Finally, the multimodal features are fed to a Large Language Model to generate explanatory reports.

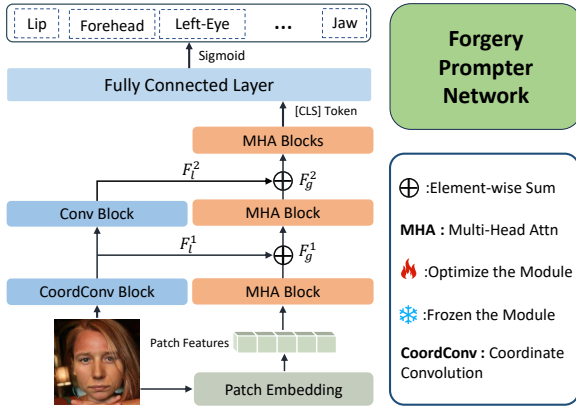


Figure 5: Illustration of the Forgery Prompter Network (FPN). The FPN generates region-aware prompts for forgery localization.

resultant feature F_g^m contains both global and local contexts and is fed into subsequent multi-head attention blocks and a classification head to produce the probability \hat{Y} across regions, while also being used in cross-attention with Q-former features to enhance forgery localization. Finally, the forgery prompter network is trained using a combined loss, incorporating both Binary Cross-Entropy (BCE) and Dice loss to effectively balance region classification and overlap precision:

$$\mathcal{L}_{BCE} = -\frac{1}{21} \sum_{i=1}^{21} Y_i \log \hat{Y}_i + \omega (1 - Y_i) \log (1 - \hat{Y}_i), \quad (3)$$

where ω is a discount factor set to $\omega < 1$ to address the imbalance due to the prevalence of unmodified

regions. The Dice loss is employed to measure the overlap between the predicted labels \hat{Y} and ground truth Y , ensuring that less frequent classes receive more attention:

$$\mathcal{L}_{Dice} = 1 - \frac{2 \sum_{i=1}^{21} Y_i \hat{Y}_i}{\sum_{i=1}^{21} Y_i + \sum_{i=1}^{21} \hat{Y}_i}. \quad (4)$$

Finally, we optimize FPN with the average of the BCE and Dice losses via $\frac{1}{2}(\mathcal{L}_{BCE} + \mathcal{L}_{Dice})$.

3.3 Attribution Report Generation

Subsequently, we fix the trained FPN network and take its region predictions as prior clues to aid both the report generation and the cross-attention process for improved forgery localization. Assume the set of regions from the FPN is $R = \{r_1, r_2, \dots\}$. We design a particular template to include R and form a report-focused instruction T_{instr} :

These facial areas may be manipulated by AI: [R]. Please describe the specific issues in these areas.

This structured prompt serves as the guiding context for the language model, thereby ensuring that the final output accurately reflects the manipulations detected by the FPN. This integration enhances the coherence and quality of the generated reports, offering a comprehensive understanding of the tampered regions. Subsequently, the instruction and the image embeddings are fed into the Q-former, and the resulting features are passed to

the large language model to generate the explanatory text \hat{T} with the length of $L_{\hat{T}}$. This output is then supervised by the language modeling loss as:

$$\mathcal{L}_t = -\mathbb{E}_{(I,T) \sim \mathcal{D}} \left[\sum_{k=1}^{L_{\hat{T}}} \log P(\hat{T}_k | I, \hat{T}_0, \dots, \hat{T}_{k-1}) \right], \quad (5)$$

where $(I, T) \sim \mathcal{D}$ indicates that the expectation is taken over samples from the dataset \mathcal{D} .

3.4 Mask Decoder

We employ SAM’s Two-way Transformer (Kirillov et al., 2023) as the mask decoder. The image encoder of InstructBLIP encodes the forgery image. The resulting features from the Q-former are then enhanced through cross-attention with the FPN’s regional prompts. These enriched features are subsequently fed into the Two-way Transformer to predict the forgery mask \hat{M} . The cross-entropy loss is applied:

$$\mathcal{L}_m = -\frac{1}{HW} \sum_{i=1}^H \sum_{j=1}^W \left[M_{ij} \log \hat{M}_{ij} + (1-M_{ij}) \log (1-\hat{M}_{ij}) \right], \quad (6)$$

where H, W are the height and width of the image. Overall, the full loss in the second stage for report generation and forgery localization is formulated as $\mathcal{L}_{full} = \mathcal{L}_t + \mathcal{L}_m$.

4 Experiment

In this section, we present a series of experiments to evaluate our proposed model, ForgeryTalker, on the MMTT dataset against several baselines.

4.1 Quantitative Results

As shown in Table 2, we benchmark our proposed baseline, ForgeryTalker, against a comprehensive set of existing models adapted for our task: SCA (Huang et al., 2024a), LISA-7B (Lai et al., 2024), Osprey (Yuan et al., 2024), InstructBLIP (Dai et al., 2023), FFAA (Huang et al., 2024b), and FakeShield (Xu et al., 2025).

Report Generation. ForgeryTalker obtains the highest CIDEr score (59.3) on the standard benchmark, significantly surpassing strong competitors like InstructBLIP (51.7) and LISA-7B (44.1), as well as other baselines including SCA (40.6), Osprey (24.5), FFAA (17.4), and FakeShield (10.0). It also achieves the best performance across all BLEU scores and leads in ROUGE-L (28.8). To further verify the generalization capability, we extend the evaluation to the DQ_F++ dataset (Zhang

et al., 2024b) in a zero-shot setting. As shown in Table 4, ForgeryTalker demonstrates superior performance compared to advanced baselines. Specifically, it achieves a dominant CIDEr score of **113.3**, significantly outperforming the runner-up InstructBLIP (98.5). Furthermore, ForgeryTalker ranks first in all BLEU metrics, including BLEU-1 (**48.5**) and BLEU-4 (**32.4**), while maintaining highly competitive performance on ROUGE-L (**47.2**). These results validate that our method generalizes effectively to unseen datasets.

Forgery Localization. ForgeryTalker achieves the highest IoU (73.67) and Precision (91.43). Its competitive Recall (86.22) is second only to SCA, which achieves the highest Recall of 92.11 but with a notably lower IoU (46.69) and Precision (48.49). Other baselines like InstructBLIP, LISA-7B, and FakeShield report lower IoUs of 64.04, 52.45, and 47.44, respectively. Note that Osprey and FFAA do not provide standalone forgery masks, so their localization metrics are not reported. The qualitative results in Figure 6 visually corroborate these findings. While InstructBLIP’s predicted masks (green) often over-segment beyond the ground-truth (red) and its reports are verbose, ForgeryTalker consistently produces more precise masks and concise, relevant reports.

Human Evaluation. To assess the practical utility of our generated reports to non-experts, we conducted a targeted human evaluation. Five independent evaluators performed a blind test on 100 randomly sampled test images, rating the text on a 1-5 Likert scale for *Faithfulness* (accuracy in identifying artifacts without hallucination) and *Helpfulness* (ability to assist human verification). As shown in Table 3, ForgeryTalker significantly outperforms baselines, demonstrating that explicitly grounding text in localized visual evidence substantially calibrates human trust.

4.2 Ablation Study

Effect of the Forgery Prompter Network (FPN). The FPN is shown to be a critical component. Table 5 reveals a significant performance drop in report generation (CIDEr drops to 51.7) when the FPN is removed. Conversely, an oracle FPN using ground-truth prompts (w/ FPN-GT) establishes a high upper bound at 95.1 CIDEr. This large performance gap underscores that the quality of region prompts is a key factor for this task and motivates future work on improving the FPN module.

Effectiveness of Pretraining Stage. Table 6 em-

Setting	Method	Reference	Report Generation					Forgery Localization			
			CIDEr	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	IoU	Precision	Recall
Zero-Shot	Seed1.5VL (Guo et al., 2025)	arXiv25	0.54	17.19	5.35	1.86	0.98	13.74	-	-	-
	Qwen2.5VL (72B) (Bai et al., 2025)	arXiv25	2.72	20.34	6.33	2.55	1.46	16.52	-	-	-
	llava (72B) (Liu et al., 2024)	Blog24	3.06	22.01	8.35	3.30	1.80	16.83	-	-	-
	InternVL3 (78B) (Zhu et al., 2025a)	arXiv25	2.67	22.31	8.08	3.16	1.75	16.82	-	-	-
Fine-Tuning	PSSC-NET (Liu et al., 2022b)	CVPR22	-	-	-	-	-	-	32.33	70.30	37.44
	SCA (Huang et al., 2024a)	CVPR24	40.6	30.6	17.8	11.2	8.2	27.6	46.69	48.49	92.11
	LISA-7B (Lai et al., 2024)	ICCV23	44.1	31.1	17.9	10.8	8.5	28.4	52.45	73.26	71.53
	Osprey (Yuan et al., 2024)	CVPR24	24.5	28.7	16.4	9.4	6.2	25.9	-	-	-
	InstructBLIP (Dai et al., 2023)	NeurIPS23	51.7	31.8	20.3	14.6	11.4	27.7	64.04	87.88	78.53
	FFAA (Huang et al., 2024b)	arXiv24	17.4	12.0	6.6	4.0	12.9	21.7	-	-	-
	FakeShield (Xu et al., 2025)	ICLR25	10.0	9.1	4.3	2.3	12.3	16.7	47.44	58.42	66.37
ForgeryTalker	-	59.3	35.0	22.1	16.0	12.5	28.8	73.67	91.43	86.22	

Table 2: Performance comparison of generated captions and forgery localization across models. The "Zero-Shot" section evaluates large vision-language models without task-specific training, while "Fine-Tuning" includes specialized forgery detection and localization methods.

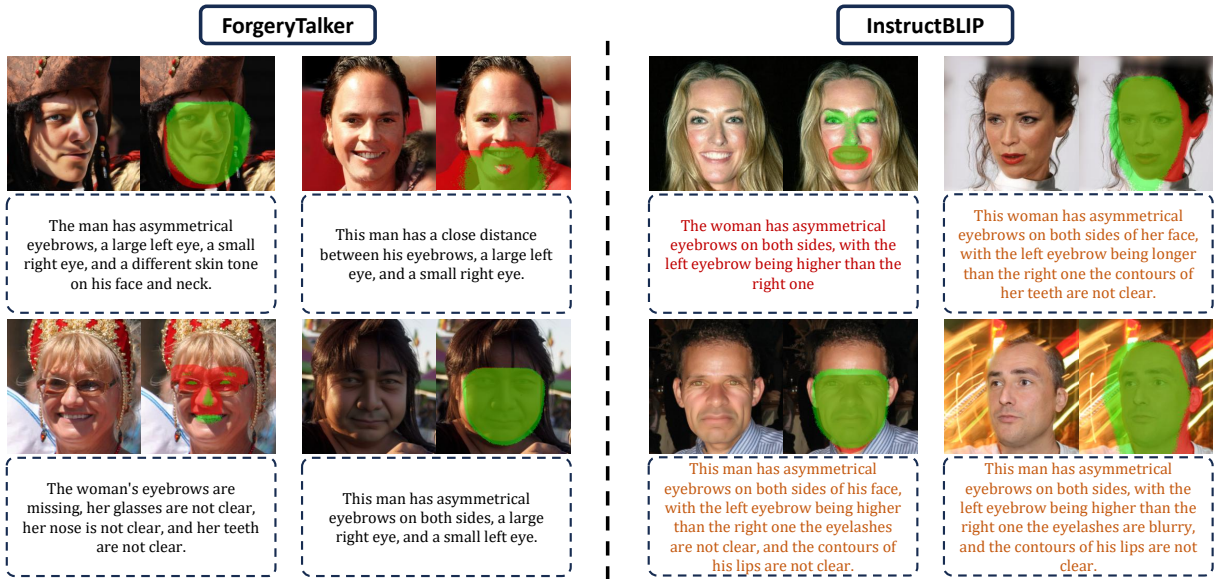


Figure 6: Qualitative comparison of ForgeryTalker and InstructBLIP. For the results, the predicted mask is shown in green and the ground-truth in red to highlight localization errors.

Method	Faithfulness \uparrow	Helpfulness \uparrow
SCA (Huang et al., 2024a)	2.3	3.1
LISA-7B (Lai et al., 2024)	2.7	3.4
InstructBLIP (Dai et al., 2023)	3.6	3.8
ForgeryTalker	4.3	4.4

Table 3: Human Evaluation results (Average Scores out of 5.0) assessing the faithfulness and helpfulness of generated reports.

pirically validates the critical importance of our forgery-aware pretraining strategy. Compared to the baseline model trained purely from scratch, integrating the pretraining stage yields substantial improvements across all evaluated metrics. Specifically, it significantly boosts the localization performance (e.g., IoU +7.8%) and enhances the semantic quality of generated reports (e.g., CIDEr +4.9). This confirms that optimizing the four pre-

training objectives equips the model with robust multimodal representations, which are pivotal for achieving high performance on the downstream joint task.

Impact of Freezing the LLM. Keeping the LLM frozen is a foundational practice in our design to preserve its pre-trained linguistic reasoning and prevent catastrophic forgetting. Empirical results validate this strategy: fine-tuning the LLM with LoRA degrades the text generation performance (dropping from 54.4 to 49.6 CIDEr) compared to the frozen setting. This confirms that preserving the LLM’s pre-trained integrity while tuning alignment modules (e.g., Q-Former and FPN) is optimal for forgery attribution.

Robustness to Real-World Degradations. Images in the wild frequently suffer from social media compression and blurring. To evaluate ro-

Model	Bleu_1	Bleu_2	Bleu_3	Bleu_4	ROUGE_L	CIDEr
SCA (Huang et al., 2024a)	47.6	39.9	35.2	30.1	40.4	71.0
LISA-7B (Lai et al., 2024)	46.5	38.4	33.1	31.2	45.6	74.3
InstructBLIP (Dai et al., 2023)	43.5	38.0	34.2	31.0	47.9	98.5
ForgeryTalker	48.5	41.4	36.5	32.4	47.2	113.3

Table 4: Report generation comparison on the DQ_F++ dataset (Zhang et al., 2024b). The best score for each metric is shown in **bold**.

Method	Report Generation						Forgery Localization		
	CIDEr	Bleu_1	Bleu_2	Bleu_3	Bleu_4	ROUGE_L	IoU	Precision	Recall
ForgeryTalker <i>w/</i> FPN-GT	95.1	41.5	27.6	20.3	16.0	37.0	66.90	88.74	79.83
ForgeryTalker <i>w/o</i> FPN	51.7	31.8	20.3	14.6	11.4	27.7	64.04	87.88	78.53
ForgeryTalker	59.3	35.0	22.1	16.0	12.5	28.8	73.67	91.43	86.22

Table 5: Ablation study on the impact of different variants. *w/* and *w/o* mean equipping or not equipping the following modules.

Method	Report Generation						Forgery Localization		
	CIDEr	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	IoU	Precision	Recall
<i>w/o</i> Pretraining Stage	54.4	33.8	21.6	15.5	12.1	28.3	65.87	89.00	78.87
<i>w/</i> Pretraining Stage	59.3	35.0	22.1	16.0	12.5	28.8	73.67	91.43	86.22

Table 6: Impact of the Forgery-aware Pretraining Stage.

Degradation Condition	CIDEr	Bleu_1	Bleu_4	ROUGE_L	IoU	Precision	Recall
Clean (Original)	59.3	35.0	12.5	28.8	73.67	91.43	86.22
Moderate (0.75× Down-sampling)	59.1	34.8	12.4	28.8	74.31	84.12	84.68
Moderate (Kernel 5 Blur)	44.8	31.9	8.8	26.2	65.89	84.50	74.09
Severe (0.5× Down-sampling)	59.7	35.1	12.5	28.8	74.87	84.22	85.14
Severe (Kernel 11 Blur)	39.1	31.3	7.8	25.4	59.51	80.82	69.44

Table 7: Robustness evaluation across various degradation levels simulating in-the-wild scenarios.

bustness, we tested ForgeryTalker under varying degradation levels (Table 7). Remarkably, under severe 0.5× down-sampling, the CIDEr score remains stable (59.7), and the mask IoU slightly increases to 74.87%, proving the model leverages scale-invariant structural cues rather than fragile pixel-level noise. While severe Gaussian blur (Kernel 11) physically erases high-frequency forensic details and impacts metrics (reducing CIDEr to 39.1), the model still maintains functional semantic reasoning based on global patterns.

5 Conclusion

We address the limitations of traditional forgery localization methods, which typically lack explanatory power. We introduce the novel task of **Forgery Attribution Report Generation** to produce both precise localization masks and rich textual explanations. To catalyze this research, we release the **MMTT** dataset, the first large-scale benchmark featuring high-precision, process-derived masks and meticulously crafted annotations. Furthermore, we propose **ForgeryTalker**, a unified baseline that integrates localization and report generation into an end-to-end framework. Our experiments validate

the effectiveness of ForgeryTalker and establish a solid benchmark on the MMTT dataset. These contributions pave the way for future advancements in explainable and trustworthy facial forgery analysis.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (Grant No. 62302140), the National Key Research and Development Program of China (Grant No. 2023YFC3321600), the Guangdong Basic and Applied Basic Research Foundation (Grant No. 2025A1515012281), the Jiangsu Provincial Science and Technology Program (Grant No. SBZ20250900116), and the Macao Science and Technology Development Fund (Grant No. FDCT/0043/2025/RIA1). Finally, we thank the anonymous reviewers and area chairs for their constructive feedback, which helped improve this paper.

Limitations

Despite the promising performance of ForgeryTalker and the MMTT dataset, we acknowledge several limitations. First, our framework provides post-hoc explanations based

on the premise of an upstream binary decision. As we focus on "where" and "why" for flagged images, the model does not collect interpretable local evidence to *drive* the initial detection. Transitioning from post-hoc attribution to a fully integrated, evidence-based detection pipeline remains an important open research direction. Second, the quality of the generated reports is highly sensitive to the accuracy of initial region proposals. As shown in our ablation studies, while the Forgery Prompter Network (FPN) establishes a robust baseline, there remains a performance gap compared to guidance using ground-truth masks. Bridging this bottleneck is crucial for more reliable forensic reporting. Finally, while ForgeryTalker is efficient during inference, the two-stage training pipeline (forgery-aware pre-training and task-specific fine-tuning) incurs higher computational costs compared to training simpler models from scratch. Additionally, although MMTT covers major manipulation paradigms, potential synthetic bias remains a challenge when encountering entirely unseen, adversarial real-world forgeries.

Ethical Considerations

The proposed MMTT dataset and the associated **ForgeryTalker** framework are developed exclusively to support academic research on deepfake detection and interpretable forensics. We acknowledge the potential dual-use risks inherent in constructing high-fidelity forged facial imagery, which could be misused to analyze or circumvent detection systems. To mitigate such risks, we adopt a strict harm-minimization and controlled-release policy. Specifically, we do **not** disclose the detailed generation pipelines or adversarial editing tools to prevent direct exploitation by malicious actors. Access to the dataset is restricted to vetted academic researchers and institutions under a signed Data Usage Agreement (DUA), which explicitly prohibits malicious content generation and identity re-identification. Moreover, all source images are obtained from publicly available academic datasets and are carefully screened to exclude minors. We reserve the right to revoke access upon any evidence of misuse.

References

Chafic Abou Akar, Rachele Abdel Massih, Anthony Yaghi, Joe Khalil, Marc Kamradt, and Abdallah Makhoul. 2024. Generative adversarial network ap-

plications in industry 4.0: A review. *International Journal of Computer Vision*, 132(6):2195–2254.

Dosovitskiy Alexey. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv: 2010.11929*.

Kafi Anan, Anindya Bhattacharjee, Ashir Intesher, Kaidul Islam, Abrar Assaeem Fuad, Utsab Saha, and Hafiz Imtiaz. 2025. Hybrid deepfake image detection: A comprehensive dataset-driven approach integrating convolutional and attention mechanisms with frequency domain features. *arXiv e-prints*, pages arXiv–2502.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. **Qwen2.5-vl technical report**. *Preprint*, arXiv:2502.13923.

Zhongxi Chen, Ke Sun, Ziyin Zhou, Xianming Lin, Xiaoshuai Sun, Liujuan Cao, and Rongrong Ji. 2024. Diffusionface: Towards a comprehensive dataset for diffusion-based face forgery analysis. *arXiv preprint arXiv:2403.18471*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, and 16 others. 2022. **Scaling instruction-finetuned language models**. *Preprint*, arXiv:2210.11416.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. **Instructblip: Towards general-purpose vision-language models with instruction tuning**. *Preprint*, arXiv:2305.06500.

Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794.

Brian Dolhansky, Joanna Bitton, Ben Pfau, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. 2020. The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397*.

Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, Jingji Chen, Jingjia Huang, Kang Lei, Liping Yuan, Lishu Luo, Pengfei Liu, Qinghao Ye, Rui Qian, Shen Yan, and 178 others. 2025. **Seed1.5-vl technical report**. *Preprint*, arXiv:2505.07062.

Yinan He, Bei Gan, Siyu Chen, Yichun Zhou, Guojun Yin, Luchuan Song, Lu Sheng, Jing Shao, and Ziwei Liu. 2021. Forgerynet: A versatile benchmark for

- comprehensive forgery analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4360–4369.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.
- Xiaoke Huang, Jianfeng Wang, Yansong Tang, Zheng Zhang, Han Hu, Jiwen Lu, Lijuan Wang, and Zicheng Liu. 2024a. Segment and caption anything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13405–13417.
- Zhengchao Huang, Bin Xia, Zicheng Lin, Zhun Mou, Wenming Yang, and Jiaya Jia. 2024b. Ffaa: Multimodal large language model based explainable open-world face forgery analysis assistant. *arXiv preprint arXiv:2408.10072*.
- Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. 2020. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2889–2898.
- Bachir Kaddar, Sid Ahmed Fezza, Wassim Hamidouche, Zahid Akhtar, and Abdenour Hadid. 2021. Hcitt: Deepfake video detection using a hybrid model of cnn features and vision transformer. In *2021 International Conference on Visual Communications and Image Processing (VCIP)*, pages 1–5. IEEE.
- Hengrui Kang, Siwei Wen, Zichen Wen, Junyan Ye, Weijia Li, Peilin Feng, Baichuan Zhou, Bin Wang, Dahua Lin, Linfeng Zhang, and Conghui He. 2025. **Legion: Learning to ground and explain for synthetic image detection**. *Preprint*, arXiv:2503.15264.
- Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410.
- Atharva Khedkar, Atharva Peshkar, Ashlesha Nagdive, Mahendra Gaikwad, and Sudeep Baudha. 2022. Exploiting spatiotemporal inconsistencies to detect deepfake videos in the wild. In *2022 10th International Conference on Emerging Trends in Engineering and Technology-Signal and Information Processing (ICETET-SIP-22)*, pages 1–6. IEEE.
- Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, and 1 others. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026.
- Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. 2024. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9579–9589.
- S Lalitha and Kavitha Sooda. 2022. Deepfake detection through key video frame extraction using gan. In *2022 International Conference on Automation, Computing and Renewable Systems (ICACRS)*, pages 859–863. IEEE.
- Trung-Nghia Le, Huy H Nguyen, Junichi Yamagishi, and Isao Echizen. 2021. Openforensics: Large-scale challenging dataset for multi-face forgery detection and segmentation in-the-wild. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10117–10127.
- Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. 2019. Faceshifter: Towards high fidelity and occlusion aware face swapping. *arXiv preprint arXiv:1912.13457*.
- Wenbo Li, Zhe Lin, Kun Zhou, Lu Qi, Yi Wang, and Jiaya Jia. 2022. Mat: Mask-aware transformer for large hole image inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10758–10768.
- Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. 2020. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3207–3216.
- Dazhuang Liu, Zhen Yang, Ru Zhang, and Jianyi Liu. 2022a. Maskgan: A facial fusion algorithm for deepfake image detection. In *2022 International Conference on Computers and Artificial Intelligence Technologies (CAIT)*, pages 71–78. IEEE.
- Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. **Llava-next: Improved reasoning, ocr, and world knowledge**.
- Hui Liu, Wenya Wang, and Haoliang Li. 2023a. Interpretable multimodal misinformation detection with logic reasoning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9781–9796.
- Kunlin Liu, Ivan Perov, Daiheng Gao, Nikolay Chervoniy, Wenbo Zhou, and Weiming Zhang. 2023b. Deepfacelab: Integrated, flexible and extensible face-swapping framework. *Pattern Recognition*, 141:109628.
- Lingyu Liu, Yaxiong Wang, Li Zhu, and Zhedong Zheng. 2025a. Every painting awakened: A training-free framework for painting-to-animation generation. *arXiv preprint arXiv:2503.23736*.
- Rosanne Liu, Joel Lehman, Piero Molino, Felipe Petroski Such, Eric Frank, Alex Sergeev, and Jason Yosinski. 2018. An intriguing failing of convolutional neural networks and the coordconv solution. *Advances in neural information processing systems*, 31.

- Xiaohong Liu, Yaojie Liu, Jun Chen, and Xiaoming Liu. 2022b. Psc-net: Progressive spatio-channel correlation network for image manipulation detection and localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11):7505–7517.
- Zhen Liu, Tim Z Xiao, , Weiyang Liu, Yoshua Bengio, and Dinghuai Zhang. 2025b. Efficient diversity-preserving diffusion alignment via gradient-informed gflownets. In *ICLR*.
- Victor Livernoche, Akshatha Arodi, Andreea Musulan, Zachary Yang, Adam Salvail, Gaétan Marceau Caron, Jean-François Godbout, and Reihaneh Rab-bany. 2025. Openfake: An open dataset and platform toward real-world deepfake detection. *arXiv preprint arXiv:2509.09495*.
- Xiaochen Ma, Bo Du, Zhuohang Jiang, Ahmed Y. Al Hammadi, and Jizhe Zhou. 2023. *Iml-vit: Benchmarking image manipulation localization by vision transformer*. Preprint, arXiv:2307.14863.
- Zihan Ma, Minnan Luo, Hao Guo, Zhi Zeng, Yiran Hao, and Xiang Zhao. 2024. Event-radar: Event-driven multi-view learning for multimodal fake news detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5809–5821.
- Joao C Neves, Ruben Tolosana, Ruben Vera-Rodriguez, Vasco Lopes, Hugo Proença, and Julian Fierrez. 2020. Ganprint: Improved fakes and evaluation of the state of the art in face manipulation detection. *IEEE Journal of Selected Topics in Signal Processing*, 14(5):1038–1048.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, and 2 others. 2019. *Pytorch: An imperative style, high-performance deep learning library*. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 8024–8035.
- Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. 2021. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2085–2094.
- Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.
- Sreeraj Ramachandran, Aakash Varma Nadimpalli, and Ajita Rattani. 2021. An experimental evaluation on deepfake detection using deep face recognition. In *2021 International Carnahan Conference on Security Technology (ICCST)*, pages 1–6. IEEE.
- Md Shohel Rana, Mohammad Nur Nobil, Beddhu Murali, and Andrew H Sung. 2022. Deepfake detection: A systematic literature review. *IEEE access*, 10:25494–25513.
- Mj Alben Richards, E Kaaviya Varshini, N Diviya, P Prakash, P Kasthuri, and A Sasithradevi. 2023. Deep fake face detection using convolutional neural networks. In *2023 12th International Conference on Advanced Computing (ICoAC)*, pages 1–5. IEEE.
- T-YLPG Ross and GKHP Dollár. 2017. Focal loss for dense object detection. In *proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2980–2988.
- Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. 2019. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1–11.
- Ekraam Sabir, Jiabin Cheng, Ayush Jaiswal, Wael AbdAlmageed, Iacopo Masi, and Prem Natarajan. 2019. Recurrent convolutional strategies for face manipulation detection in videos. *Interfaces (GUI)*, 3(1):80–87.
- Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.
- YuYang Sun, ZhiYong Zhang, Isao Echizen, Huy H Nguyen, ChangZhen Qiu, and Lu Sun. 2023. Face forgery detection based on facial region displacement trajectory series. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 633–642.
- Luisa Verdoliva. 2020. Media forensics and deepfakes: an overview. *IEEE journal of selected topics in signal processing*, 14(5):910–932.
- Tengfei Wang, Yong Zhang, Yanbo Fan, Jue Wang, and Qifeng Chen. 2022. High-fidelity gan inversion for image attribute editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Yuanlu Wu, Yan Wo, Caiyu Li, and Guoqiang Han. 2023. Learning domain-invariant representation for generalizing face forgery detection. *Computers & Security*, 130:103280.
- Zhipei Xu, Xuanyu Zhang, Runyi Li, Zecheng Tang, Qing Huang, and Jian Zhang. 2025. Fakeshield: Explainable image forgery detection and localization via multi-modal large language models. In *International Conference on Learning Representations*.
- Zhiyuan Yan, Taiping Yao, Shen Chen, Yandan Zhao, Xinghe Fu, Junwei Zhu, Donghao Luo, Chengjie Wang, Shouhong Ding, Yunsheng Wu, and Li Yuan. 2024. *Df40: Toward next-generation deepfake detection*. Preprint, arXiv:2406.13495.

- Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. 2018. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 325–341.
- Peipeng Yu, Zhihua Xia, Jianwei Fei, and Yujiang Lu. 2021. A survey on deepfake video detection. *Iet Biometrics*, 10(6):607–624.
- Yuqian Yuan, Wentong Li, Jian Liu, Dongqi Tang, Xinjie Luo, Chi Qin, Lei Zhang, and Jianke Zhu. 2024. Osprey: Pixel understanding with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28202–28211.
- Fengzhu Zeng, Wenqian Li, Wei Gao, and Yan Pang. 2024. Multimodal misinformation detection by learning from synthetic data with multimodal llms. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10467–10484.
- Guiyu Zhang, Huan-ang Gao, Zijian Jiang, Hao Zhao, and Zhedong Zheng. 2025. Ctrl-u: Robust conditional image generation via uncertainty-aware reward modeling. *ICLR*.
- Yaning Zhang, Zitong Yu, Tianyi Wang, Xiaobin Huang, Linlin Shen, Zan Gao, and Jianfeng Ren. 2024a. Genface: A large-scale fine-grained face forgery benchmark and cross appearance-edge learning. *IEEE Transactions on Information Forensics and Security*.
- Yue Zhang, Ben Colman, Xiao Guo, Ali Shahriyari, and Gaurav Bharaj. 2024b. Common sense reasoning for deepfake detection. In *European Conference on Computer Vision*, pages 399–415. Springer.
- Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. 2022. Celebv-hq: A large-scale video facial attributes dataset. In *European conference on computer vision*, pages 650–667. Springer.
- Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, and 32 others. 2025a. Internv13: Exploring advanced training and test-time recipes for open-source multimodal models. *Preprint*, arXiv:2504.10479.
- Yule Zhu, Ping Liu, Zhedong Zheng, and Wei Liu. 2025b. Seed: A benchmark dataset for sequential facial attribute editing with diffusion models. *arXiv preprint arXiv:2506.00562*.

A Related Work

Facial Manipulation Localization. Detecting manipulated facial regions, especially deepfakes, has garnered attention. CNN-based methods (Sabir et al., 2019) utilize temporal inconsistencies for videos, while GAN-based approaches, such as GANprintR (Neves et al., 2020) and MaskGAN (Liu et al., 2022a), address synthetic artifacts. Hybrid models like HCiT (Kaddar et al., 2021) combine CNNs and ViTs to enhance generalization, and multi-modal methods (Sun et al., 2023; Khedkar et al., 2022) leverage spatial-temporal inconsistencies. However, these models lack interpretability and fine-grained mask generation, which our work addresses by providing both localization masks and textual explanations.

Multi-label Classification for Facial Localization. Multi-label classification captures independent alterations in facial regions but struggles with dependencies across features. CNNs (Lalitha and Sooda, 2022) face limitations in fine-grained tasks, while hybrid models (Kaddar et al., 2021) improve detection by combining local and global features. Weighted loss functions (Ramachandran et al., 2021) and parallel branches (Richards et al., 2023) address class imbalance and refine detection. Yet, few works integrate multi-label classification with localization. Our ViT-based classifier bridges this gap by capturing complex dependencies with parallel branches and weighted loss functions.

Segmentation Techniques. Segmentation is crucial for identifying localized manipulations. Models like U-Net and DeepLab (Ross and Dollár, 2017) focus on spatial features, while Transformer models (Alexey, 2020) capture global context. Recent methods like SAM (Kirillov et al., 2023) use a Two-Way Transformer for high-quality masks but lack manipulation-specific context. By integrating SAM with InstructBLIP, we create context-aware forgery masks, unifying segmentation and manipulation detection for enhanced localization.

Explainable Forgery Detection. A recent trend is moving towards explainable forensics. Within the NLP community, multimodal misinformation detection aims to identify inconsistencies between text and images via logic reasoning (Liu et al., 2023a) or synthetic data learning (Zeng et al., 2024), yet these methods often overlook pixel-level manipulation artifacts. In the vision domain, FakeShield (Xu et al., 2025) attempts to explain general image forgeries but relies on synthetic GPT-

4o annotations. Bridging these fields, our work introduces **MMTT**, a large-scale dataset specifically designed for the facial forgery domain. Unlike previous resources, MMTT features meticulous human-in-the-loop annotation, enabling models to ground their textual explanations in precise visual regions.

B Forgery-aware Pretraining

In the main paper, we briefly outlined the Forgery-aware Pretraining stage. To provide a more comprehensive understanding of our optimization strategy, this section details the specific mathematical formulations for the four distinct training objectives: Masked Language Modeling, Language Modeling, Forgery Localization, and Cross-model Alignment Learning.

The goal of our forgery-aware pretraining stage is to learn robust multimodal representations that are sensitive to manipulation artifacts. Given an image I and its corresponding ground-truth explanation text T , we jointly optimize the core modules of our model using four distinct training objectives. The image I is first processed by a frozen visual encoder to yield embeddings E_I , which serve as input alongside the text T for the following loss functions:

Masked Language Modeling (\mathcal{L}_{mlm}): The text T is tokenized into \tilde{T} . Before feeding \tilde{T} into the Q-Former, a subset of region-related tokens (*e.g.*, “ear”, “eye”, etc.) \mathcal{M} is masked, and the masked token results in $\tilde{T}_{\setminus\mathcal{M}}$. Along with the learned query tokens Q and image embeddings E_I , the Q-Former predicts the masked tokens. The loss is computed as:

$$\mathcal{L}_{mlm} = - \sum_{t \in \mathcal{M}} \log P(t | I, \tilde{T}_{\setminus\mathcal{M}}). \quad (7)$$

Language Modeling (\mathcal{L}_{lm}): The Q-Former output is projected and fed to a T5-based decoder (Chung et al., 2022) that generates the explanatory text \hat{T} with the length of $L_{\hat{T}}$. The generated explanation is compared token-by-token with the ground truth via cross-entropy loss:

$$\mathcal{L}_{lm} = - \sum_{k=1}^{L_{\hat{T}}} \log P(\hat{T}_k | I, \hat{T}_0, \dots, \hat{T}_{k-1}), \quad (8)$$

Forgery Localization (\mathcal{L}_{seg}): The non-[CLS] tokens of E_I are seamlessly fused with the text T via cross-attention. The mask decoder predicts a

forgery mask \hat{M} with the height H and width W , which is compared to the ground-truth mask M using pixel-wise cross-entropy loss:

$$\mathcal{L}_{seg} = -\frac{1}{HW} \sum_{i=1}^H \sum_{j=1}^W [M_{ij} \log \hat{M}_{ij} + (1 - M_{ij}) \log(1 - \hat{M}_{ij})], \quad (9)$$

where $M_{ij} = 1$ if the (i, j) pixel is manipulated, 0 otherwise. **Cross-model Alignment Learning (\mathcal{L}_{con}):** To align modalities, we pull the global image feature v (from the [CLS] token) closer to the mean-pooled text feature t with contrastive loss as:

$$\mathcal{L}_{con} = -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(\text{sim}(v_i, t_i)/\tau)}{\sum_{j=1}^N \exp(\text{sim}(v_i, t_j)/\tau)}, \quad (10)$$

where N is the batch size, $\text{sim}(\cdot)$ denotes cosine similarity and τ is a temperature parameter. The overall pretraining loss is defined as:

$$\mathcal{L}_{pretrain} = \lambda_1 \mathcal{L}_{mlm} + \lambda_2 \mathcal{L}_{lm} + \lambda_3 \mathcal{L}_{seg} + \lambda_4 \mathcal{L}_{con}, \quad (11)$$

where $\lambda_1, \lambda_2, \lambda_3,$ and λ_4 being empirically tuned weights. The joint optimization of these losses enables our model to capture both fine-grained local details and global semantic context. This robust initialization is pivotal for the subsequent Attribution Report Generation Stage, where further fine-tuning refines forgery localization and enhances the quality of the generated reports.

C Examples from MMTT Dataset

To enhance the understanding of the MMTT dataset and its unique contributions to facial image forgery localization, we provide a word cloud generated from the textual descriptions (captions) and a series of representative examples. The MMTT dataset is meticulously designed to facilitate fine-grained forgery localization by leveraging multimodal annotations. Each sample consists of three complementary components: a manipulated image, a binary mask delineating the forged regions, and a detailed textual description that explicitly identifies and contextualizes the alterations. These comprehensive annotations provide a robust foundation for research tasks requiring precise localization and explainability of facial manipulations.

The word cloud, presented in Figure 8, visually encapsulates the linguistic distribution within the dataset’s textual annotations. Dominant terms such as "woman," "man," "skin tone," and "facial

skin" highlight the dataset’s focus on describing forgery in specific facial regions. Furthermore, frequent mentions of region-specific features, such as "left eye," "nose bridge," and "right eyebrow," underscore the granularity and specificity of the annotations. This visualization demonstrates the alignment between the textual descriptions and the underlying task of forgery localization, offering an overview of the dataset’s descriptive richness and consistency.

Figure 7 illustrates selected examples from the MMTT dataset, showcasing its multimodal structure and the diversity of forgery types. Each example includes a manipulated image, its corresponding binary mask, and a textual description. For illustrative purposes, we have also included the original (authentic) images alongside the manipulated samples in Figure 7 to provide additional context for understanding the nature and extent of the forgeries. It is important to note that these original images are not part of the MMTT dataset and are shown exclusively to highlight the transformations and to provide clarity on the dataset’s structure. The actual dataset is focused on forged images, binary masks, and detailed captions, without the inclusion of original (authentic) images.

D Experimental Setup

Implementation Details. We implement ForgeryTalker with PyTorch (Paszke et al., 2019) and train on a single NVIDIA H100 (94GB) GPU. Our model is built upon the InstructBLIP framework, utilizing the frozen Flan-T5-XL as the Large Language Model (LLM) backbone. The total number of parameters is approximately 4B. The entire two-stage training process takes approximately 2 days (i.e., 48 GPU hours).

Training Protocol. We use an 8:1:1 train/validation/test split of the MMTT dataset. The two-stage training process is as follows: (1) The FPN is trained for 125k steps (batch size 16, initial lr 7.5e-3 with cosine decay) with the BCE loss weight ω set to 0.2. (2) With the FPN frozen, the main model is trained for 60 epochs (batch size 16, lr 4e-6) using mixed-precision (fp16) training.

Evaluation Metrics. For the report generation task, we employ standard captioning metrics including CIDEr, BLEU, and ROUGE-L, computed using the official pycocoEvalcap toolkit. For forgery localization, we report pixel-level Intersection over Union (IoU), Precision, and Recall to evaluate the

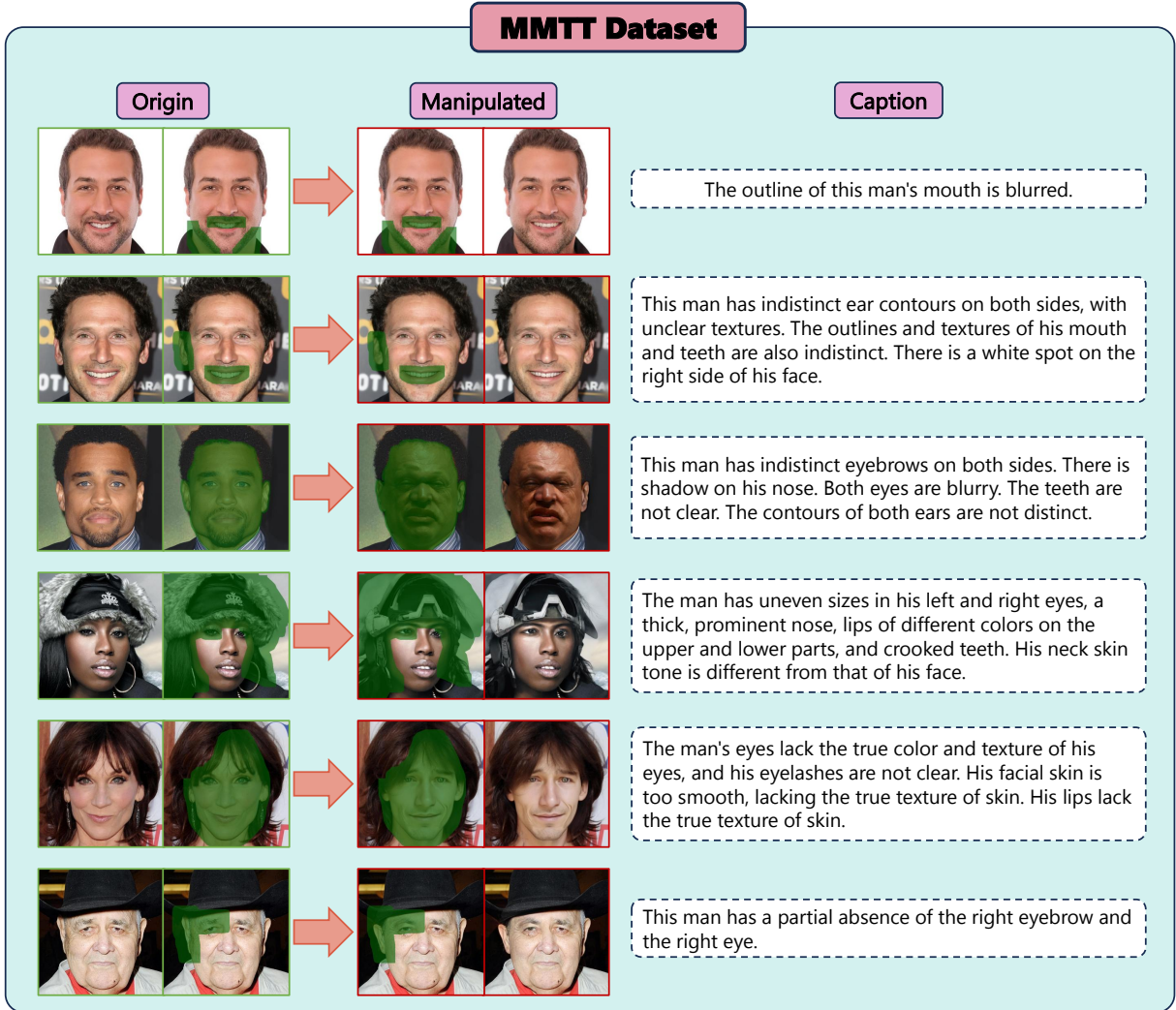


Figure 7: Examples from the MMTT dataset. Each row illustrates a case from the dataset, comprising a manipulated image, its corresponding binary mask (overlaid in green), and a textual description detailing the altered facial regions. For illustrative purposes, the original (authentic) images are also included in this figure to highlight the extent and nature of the manipulations. The green regions indicate the localized areas of forgery as identified by the binary masks. It is important to note that the original images are not part of the MMTT dataset; the dataset itself consists only of manipulated images, binary masks, and their associated textual descriptions.

mask alignment quality.

E Comparison with State-of-the-art Models

We compare the performance of our proposed ForgeryTalker framework against a range of recent models, including specialist forgery localization models and general-purpose Large Vision-Language Models (LVLMs). It is important to note that while the specialist models and our ForgeryTalker were trained on our dataset, the general LVLMs were evaluated in a zero-shot setting to assess their out-of-the-box capabilities for this novel task. The evaluation covers both forgery localization (IoU, Precision, Recall) and interpreta-

tion generation (CIDEr, ROUGE-L, BLEU scores), with results summarized in Table 8.

For forgery localization, specialist models like IML-ViT expectedly achieve the highest scores in IoU (77.89) and Recall (90.04), as they are solely optimized for this task. However, our ForgeryTalker demonstrates highly competitive localization capabilities, achieving a strong IoU of **73.67** and securing the best Precision score (**91.43**) among all compared models. This indicates our model’s superior ability to avoid over-predicting forged regions.

For the primary task of interpretation generation, ForgeryTalker significantly outperforms all other LVLMs across every text-based metric. It

Model	Interpretation Generation						Forgery Localization		
	CIDEr	ROUGE_L	Bleu_1	Bleu_2	Bleu_3	Bleu_4	IoU	Precision	Recall
Seed1.5VL (Guo et al., 2025)	0.54	13.74	17.19	5.35	1.86	0.98	-	-	-
Qwen2.5VL (72B) (Bai et al., 2025)	2.72	16.52	20.34	6.33	2.55	1.46	-	-	-
Qwen2.5VL (32B) (Bai et al., 2025)	2.53	16.89	22.44	8.16	3.3	1.78	-	-	-
Qwen2.5VL (7B) (Bai et al., 2025)	2.48	17.0	22.33	8.2	3.24	1.78	-	-	-
llava (72B) (Liu et al., 2024)	3.06	16.83	22.01	8.35	3.3	1.8	-	-	-
llava (8B) (Liu et al., 2024)	2.1	18.05	20.75	7.66	3.15	1.75	-	-	-
InternVL3 (78B) (Zhu et al., 2025a)	2.67	16.82	22.31	8.08	3.16	1.75	-	-	-
InternVL3 (38B) (Zhu et al., 2025a)	2.29	17.18	21.21	7.83	3.2	1.76	-	-	-
InternVL3 (14B) (Zhu et al., 2025a)	2.29	16.85	20.81	7.54	3.03	1.71	-	-	-
IML-ViT (Ma et al., 2023)	-	-	-	-	-	-	77.89	83.76	90.04
PSCC-NET (Liu et al., 2022b)	-	-	-	-	-	-	32.33	70.3	37.44
ForgeryTalker	59.3	28.8	35.0	22.1	16.0	12.5	73.67	91.43	86.22

Table 8: Performance comparison against state-of-the-art models in the supplementary material. Specialist models (IML-ViT, PSCC-NET) and our ForgeryTalker are trained on our dataset. **General Large Vision-Language Models (LVLMs) are evaluated in a zero-shot setting.** Best performance for each metric is in **bold**.

Model	ω	Loss	PLM
ViT	1	BCE	34.23
ViT	0.2	BCE	38.92
FPN	0.2	BCE	39.16
FPN	0.2	BCE + Dice	41.05

Table 9: Ablation Study on the Impact of the FPN

of 6.5, significantly surpassing the second-best results. It is worth noting that the absolute scores for n-gram based metrics, particularly CIDEr (2.2), are relatively modest for all models. This is largely attributed to the inherent stylistic discrepancy between the ground-truth captions in SynthScars and our MMTT training data. Since metrics like CIDEr are highly sensitive to specific linguistic patterns, the domain shift in annotation style leads to lower numerical scores. Despite this, our method demonstrates superior transferability and relative performance compared to the state-of-the-art baselines.

H Ablation Study on Manipulation Type Contributions

To investigate the individual contributions of different forgery types within our MMTT dataset—namely Face Swapping, Face Editing, and Image Inpainting—we conducted an ablation study based on data composition. Specifically, we trained ForgeryTalker on subsets containing only two out of the three manipulation types and evaluated the performance on the full test set. This “leave-one-type-out” setting allows us to quantify the impact

of the missing type on the model’s generalization capability.

The results are presented in Table 11. The comparison reveals distinct roles for each manipulation type:

- **Impact of Face Editing on Report Quality:** When Face Editing data is excluded from training (see the column “Swapping + Inpainting”), the report generation performance suffers the most dramatic decline, with the CIDEr score dropping from 56.9 to **16.2**. This suggests that the Face Editing samples in MMTT contain the most diverse and semantically complex linguistic descriptions. Without them, the model struggles to generate high-quality, descriptive attribution reports.
- **Impact of Inpainting on Localization and Fluency:** Excluding Image Inpainting data (see the column “Swapping + Editing”) leads to the lowest localization accuracy (IoU drops to 61.28) and a collapse in ROUGE-L score (2.7). This indicates that Inpainting samples are critical for the model to learn precise boundary localization and to maintain the structural fluency of the generated text.
- **Impact of Face Swapping:** Removing Face Swapping data (see the column “Editing + Inpainting”) results in a moderate performance drop across both localization and generation metrics (e.g., IoU drops to 63.12, CIDEr to 49.2). This implies that Face Swapping contributes to the overall robustness of the model but is less specialized than the other two types

Model	Bleu_1	Bleu_2	Bleu_3	Bleu_4	ROUGE_L	CIDEr
SCA (Huang et al., 2024a)	9.23	6.45	3.89	0.93	15.74	0.54
Osprey (Yuan et al., 2024)	7.54	6.09	4.36	2.29	13.28	1.40
ForgeryTalker	10.80	8.80	7.40	6.50	32.20	2.20

Table 10: Zero-shot report generation performance on the face-modification subset of the SynthScars dataset (Kang et al., 2025). All models were fine-tuned on the MMTT dataset and evaluated on SynthScars without further tuning.

Table 11: Ablation study on the contribution of different forgery types within the MMTT dataset. Each row represents a model trained on a specific subset of data (leaving one type out) and evaluated on the full test set. ‘‘Full Set’’ indicates training with all forgery types.

Training Data Composition	Forgery Localization			Report Generation					
	IoU	Precision	Recall	Bleu_1	Bleu_2	Bleu_3	Bleu_4	ROUGE_L	CIDEr
Full Set (All Types)	71.00	90.27	84.04	31.90	19.60	14.20	11.20	28.30	56.90
w/o Face Swapping	63.12	84.81	73.80	28.10	17.80	12.70	10.00	26.30	49.20
w/o Image Inpainting	61.28	72.95	80.55	25.00	15.90	11.80	9.50	2.70	48.50
w/o Face Editing	68.20	89.77	78.72	23.50	12.30	7.00	4.20	21.50	16.20

in driving specific extreme metrics.

In summary, the diversity of manipulation types in MMTT is essential, with each type contributing uniquely to either the visual localization precision or the semantic richness of the generated reports.

I Statements

I.1 Reproducibility Statement

To ensure transparency and facilitate future research, we commit to the full release of our code, data, and model weights. First, we will make the complete source code for ForgeryTalker publicly available on GitHub upon publication, covering both the pre-training and report generation stages. This repository will include comprehensive training scripts, configuration files for ablation studies, and detailed environment setup instructions. Second, the MMTT dataset, which consists of 152,217 image-text-mask triplets, will be released to the research community under our Data Usage Agreement. Finally, to establish a standardized benchmark and lower the barrier for subsequent work, we will provide pre-trained checkpoints for our best-performing models and baselines.

I.2 LLM Usage Statement

During the preparation of this work, the authors used a large language model to assist with improving grammar, rephrasing sentences, and ensuring terminological consistency. The authors reviewed and edited all model-generated text and take full responsibility for the final content of this paper.