FAKE-IN-FACEXT: TOWARDS FINE-GRAINED EXPLAINABLE DEEPFAKE ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

The advancement of Vision-Language Models (VLMs) has bridged the gap between vision and language tasks, enabling the implementation of Explainable DeepFake Analysis (XDFA). However, current methods suffer from a lack of fine-grained awareness: the description of artifacts in data annotation is unreliable and coarse-grained, and the models fail to support the output of connections between textual forgery explanations and the visual evidence of artifacts, as well as the input of queries for arbitrary facial regions. As a result, these VLMs can not fully leverage prior knowledge of Face Context (Facext) to enhance performance and broaden application scope. To address this limitation, we propose the **Fake-in-Facext (FiFa)** framework, with contributions focusing on data annotation and model construction. We first define a Facial Image Concept Tree (FICT) to divide facial images into fine-grained regional concepts, thereby obtaining a more reliable data annotation pipeline, FiFa-Annotator, for forgery explanation. We validate that the data generated by FiFa-Annotator can effectively improve detection generalization. Based on this dedicated data annotation, we introduce a novel Artifact-Grounding Explanation (AGE) task, which generates textual forgery explanations interleaved with segmentation masks of manipulated artifacts. We propose a unified multi-task learning architecture, FiFa-VLM, to simultaneously support abundant multimodal inputs and outputs for fine-grained Explainable DeepFake Analysis. With multiple auxiliary supervision tasks, FiFa-VLM can outperform strong baselines on the AGE task and achieve SOTA performance on existing XDFA datasets. The code and data will be made open-source.

1 INTRODUCTION

Artificial Intelligence Generated Content (AIGC) blurs the boundary between fiction and reality. Unauthorized DeepFake images can be exploited to disseminate misinformation, potentially leading to significant social issues and security threats (Tolosana et al., 2020; Turton & Martin, 2020). DeepFake Analysis (DFA) thus becomes crucial for regulating technological applications and mitigating associated societal risks. Over the past two years, the rapid advancement of Vision-Language Models (VLMs) (Liu et al., 2023) has bridged the gap between vision and language tasks. This progress has enabled the research to move beyond simple binary classification, expanding its scope to generative visual question answering. Many works for Explainable DeepFake Analysis (XDFA) (Zhang et al., 2024b; Huang et al., 2024; Qin et al., 2025; Xu et al., 2025; Sun et al., 2025; Guo et al., 2025) have emerged.

However, existing VLMs for XDFA still have a fundamental limitation: they suffer from a lack of fine-grained awareness. Specifically, there are the following issues: (1) in terms of data annotation, existing methods over-rely on GPTs to identify artifact-existing regions while neglecting the application of prior knowledge, leading to unreliability in artifact explanations. Furthermore, the set of concepts used to describe the locations of artifacts is relatively small and only includes a limited number of concepts, which results in imprecise descriptions of artifact locations. For example, an artifact occurring on "the left nasal ala" can only be described as appearing in "the region around the nose". (2) In terms of model construction, these models only output textual explanations and lack the connection between textual forgery explanations and the visual evidence of artifacts. Additionally, they do not support flexible linguistic or visual prompts, thus failing to provide users with targeted discussions regarding facial regions of interest. To address the aforementioned issues, we

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

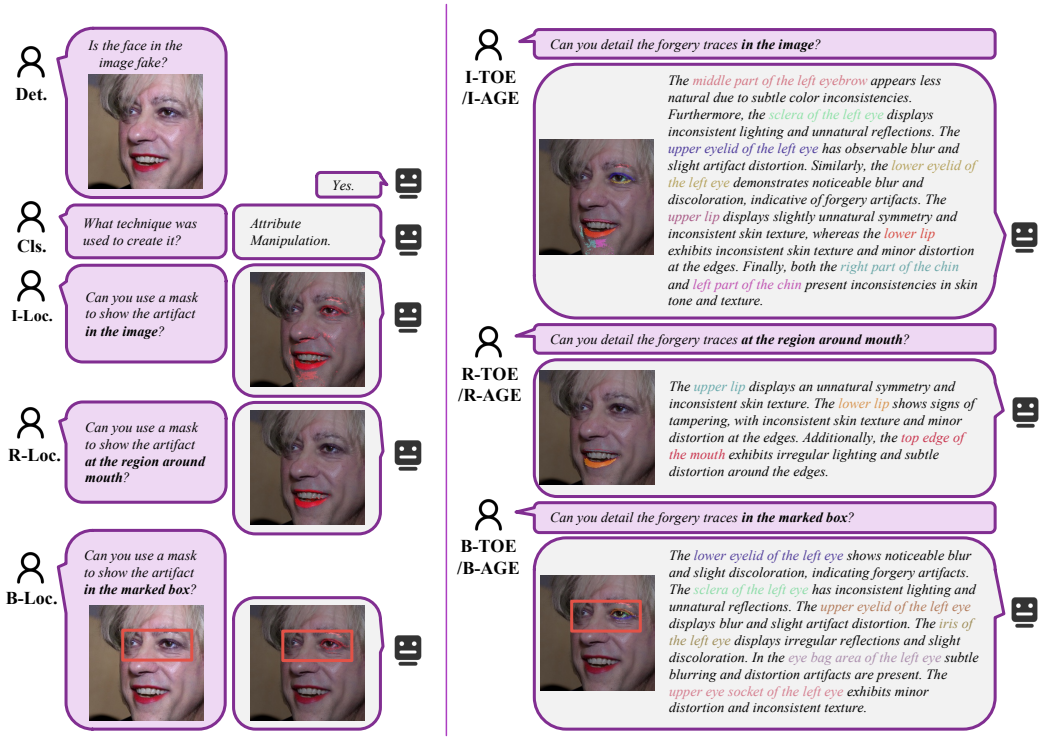

Figure 1: 11 Tasks in **FiFa-11**: Detection (Det.), Classification (Cls.), Image-Level Localization (I-Loc.), Region-Level Localization (R-Loc.), Box-Level Localization (B-Loc.), Image-Level Text-Only Explanation (I-TOE), Region-Level Text-Only Explanation (R-TOE), Box-Level Text-Only Explanation (B-TOE), Image-Level Artifact-Grounding Explanation (I-AGE), Region-Level Artifact-Grounding Explanation (R-AGE), Box-Level Artifact-Grounding Explanation (B-AGE).

propose the **Fake-in-Facext (FiFa)** framework, which aims to empower trained VLMs to fully leverage prior knowledge of fine-grained Face Context (Facext), thereby enhancing detection precision and broadening its application scope.

To bridge the gap in annotation reliability, we develop an automated data annotation pipeline termed **FiFa-Annotator**. We introduce the Facial Image Concept Tree (FICT), a structured prior that models the Face Context (Facext) through 8 hierarchical levels, encompassing 112 Atomic Concepts and 72 Parent Concepts. This allows for an unprecedented fine-grained division of facial images. The pipeline first identifies the list of Atomic Concepts containing artifacts based on the artifact coverage ratio within the corresponding facial region. Subsequently, a powerful VLM (OpenAI, 2024) generates a detailed forgery explanation for each artifact-containing Atomic Concept. Finally, a robust LLM (OpenAI, 2023) aggregates these explanations derived from Atomic Concepts to synthesize comprehensive forgery explanations for Parent Concepts across different hierarchical levels. By grounding the annotation process in the FICT, FiFa-Annotator significantly mitigates hallucinations and achieves a level of linguistic precision that far surpasses existing annotation pipelines. We further demonstrate that the data generated by FiFa-Annotator can significantly improve generalization capabilities in DeepFake Detection, through finetuning CLIP with a multimodal joint strategy.

Empowered by this high-quality data, we further extend the frontiers of XDFA by defining a more versatile set of tasks. While existing VLMs are limited to DeepFake Detection (Det.), Classification (Cls.), Localization (Loc.), and Text-Only Explanation (TOE), our framework introduces the Artifact-Grounding Explanation (AGE) task. AGE requires the model to provide natural language explanations while simultaneously generating segmentation masks that pinpoint the mentioned artifacts. Moreover, we transition from rigid Image-Level queries to flexible, multi-modal queries, enabling the specification of facial regions of interest via textual prompts (Region-Level) and bounding box visual prompts (Box-Level). Collectively, we define a suite of 11 tasks for fine-grained XDFA, called **FiFa-11**. Samples are depicted in Fig. 1. To support these, we have constructed

Table 1: Comparison of capabilities supported by existing work and FiFa framework

Model	Dataset	QA-Pairs	Det.	Cls.	I-Loc.	I-TOE	I-AGE	R-Loc.	R-TOE	R-AGE	B-Loc./B-TOE/B-AGE
DDVQA-BLIP	DD-VQA	15K	✓			✓			✓		
FFAA	FFA-VQA	20K	✓	✓		✓					
DFA-GPT	DFA-Instruct	127K	✓	✓		✓			✓		
FakeShield	MMTD-SET	17K (For Face Forgery)	✓		✓	✓					
FiFa-VLM (Ours)	**FiFa-Instruct-1M (Ours)**	**1.38M**	✓	✓	✓	✓	✓	✓	✓	✓	✓

the **FiFa-Instruct-1M** dataset (comprising 1.38 million samples) and a dedicated evaluation benchmark, **FiFa-Bench**.

To address these challenging tasks in **FiFa-11**, we have constructed a unified multi-task learning architecture, **FiFa-VLM**. Current VLMs (Lai et al., 2024; Rasheed et al., 2024) typically introduce additional visual encoders dedicated to segmentation for pixel grounding capability, resulting in inefficient architectures. To address this, our FiFa-VLM employs only one global visual encoder. This encoder concurrently generates suitable visual features for both LLM input and mask prediction. Furthermore, we propose a Multi-Task Decoder. By introducing distinct task-specific query embeddings, this decoder can simultaneously handle Artifact Mask Prediction and multiple auxiliary supervision tasks. Experiments reveal that auxiliary supervision of Region Mask Prediction effectively enhances the accuracy of Artifact Mask Prediction in both Loc. and AGE tasks. We also support bounding box visual prompts by introducing a Box Encoder. Consequently, the resulting VLM demonstrates compelling advantages: despite having 0.94B fewer parameters than the strong baseline GLaMM (Rasheed et al., 2024), it achieves significantly superior performance on nearly all tasks in FiFa-11. On the existing XDFA test benchmarks, DD-VQA and DFA-Bench, FiFa-VLM also achieves SOTA results.

Our contribution of the **Fake-in-Facext (FiFa)** framework for fine-grained Explainable DeepFake Analysis can be summarized as follows:

1. We propose **FiFa-Annotator**, a novel automated pipeline that leverages prior knowledge of Facext to minimize hallucinations. Using this pipeline, we construct the **FiFa-Instruct-1M** training dataset alongside the **FiFa-Bench** evaluation dataset. FiFa-Instruct-1M is the largest training dataset for the XDFA field currently known.
2. We propose **FiFa-VLM**, the first XDFA-aimed model capable of supporting Artifact-Grounding Explanation and responding to bounding box prompts.
3. Through a well-designed architecture, our FiFa-VLM significantly outperforms the strong baseline across almost all tasks in **FiFa-11**.

2 Related Works

Explainable DeepFake Analysis. The advancement of VLMs has expanded the application scope of DFA beyond detection and classification. Models such as DDVQA-BLIP (Zhang et al., 2024b), FFAA (Huang et al., 2024), DFA-GPT (Qin et al., 2025) have demonstrated textual explanation capabilities, while FakeShield (Xu et al., 2025) has further integrated localization. Table 1 contrasts the capabilities supported by existing works with our proposed framework. Training XDFA-aimed VLMs necessitates high-quality image-text annotations. Existing data annotation pipelines can be broadly classified into two categories: (1) Manual annotation (Zhang et al., 2024b). While providing reliable forgery explanations, this method suffers from scalability limitations and often yields monotonous textual descriptions. (2) Automated annotation (Huang et al., 2024; Qin et al., 2025; Xu et al., 2025). This method employs meticulously crafted prompts to leverage powerful VLMs like GPT-4o (OpenAI, 2024) for generating holistic facial forgery explanations. However, this method heavily depends on the forgery understanding abilities of the underlying VLMs, thus reducing the reliability of textual explanations. FFTG (Sun et al., 2025) utilizes landmark localization to localize artifact regions; however, its scope is confined to four specific areas (mouth, nose, eyes, and whole face), which precludes the model from achieving a fine-grained awareness of the overall Face Context.

Pixel Grounding and Box-Level Understanding in VLM. Seminal works, including GPT4RoI (Zhang et al., 2024a), Kosmos-2 (Peng et al., 2023), and Shikra (Chen et al., 2023), pioneer the Box-Level understanding capability in VLMs, enabling conversational focus specification through bounding box visual prompts. Pixel grounding is first introduced into VLMs by LISA (Lai et al.,

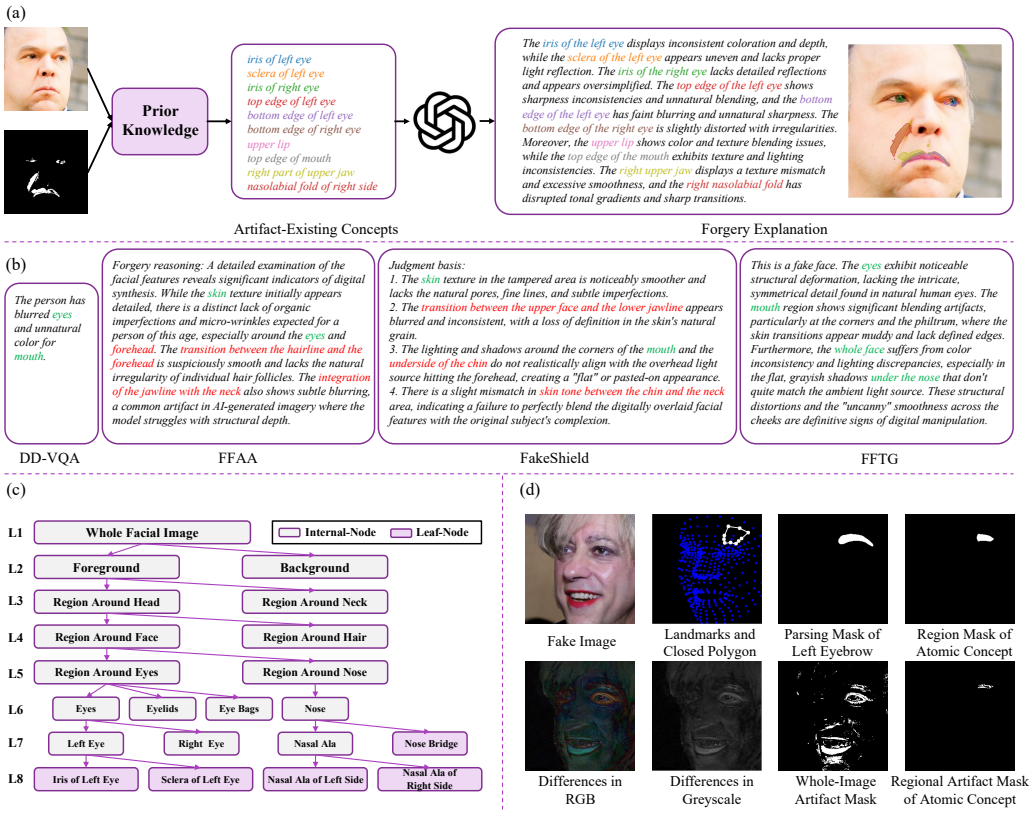

Figure 2: (a) The overall workflow of FiFa-Annotator. Forgery explanations generated by our annotator incorporate a wealth of fine-grained concepts and maintain a one-to-one correspondence with the artifact masks. (b) The forgery explanation annotations generated by existing methods. Actual artifact regions are highlighted in green, while hallucinated ones are indicated in red. (c) Facial Image Concept Tree (FICT). We have selected part of the nodes for illustration. For convenience, some internal-nodes do not have their child-nodes drawn. For the complete list of nodes, refer to Appendix A.1. (d) Image processing procedure in FiFa-Annotator.

2024) and GLaMM (Rasheed et al., 2024). These approaches not only employ the CLIP (Radford et al., 2021) encoder to provide visual features for LLM inputs but also necessitate integrating the SAM (Kirillov et al., 2023) encoder specifically for mask prediction.

Transformer Decoder in Computer Vision. The success of DETR (Carion et al., 2020) has motivated researchers to employ the transformer decoder for solving computer vision problems. MaskFormer (Cheng et al., 2021) proposed a unified approach for semantic and instance-level segmentation tasks, where each segment is represented by a query in the transformer decoder. In Face Perception, Faceptor (Qin et al., 2024), Facexformer (Narayan et al., 2024), and Q-Face (Sun et al., 2024) utilize task-specific queries to handle multiple tasks simultaneously. Our FiFa-VLM implements a Multi-Task Decoder with a group of task-specific query embeddings to concurrently address the primary task of Artifact Mask Prediction alongside a set of auxiliary supervision tasks.

3 Data Annotation Pipeline

We introduce the Facial Image Concept Tree (FICT) and the FiFa-Annotator pipeline built upon it, compare annotation quality with existing pipelines, and present FiFa-Instruct-1M and FiFa-Bench.

3.1 Facial Image Concept Tree

To model context in facial images, we propose the Facial Image Concept Tree (FICT), depicted in Fig. 2(c). This hierarchical tree comprises 8 levels, where each node represents a region concept

within the facial image. We designate concepts on leaf-nodes as Atomic Concepts (totaling 112) and concepts on internal-nodes as Parent Concepts (totaling 72). Leveraging existing facial landmark localization (Lugaresi et al., 2019) and face parsing (Zheng et al., 2022a) tools, we can derive a Region Mask corresponding to each Atomic Concept. The first row of Fig. 2(d) illustrates the Region Mask acquisition process for the "middle part of the left eyebrow" concept: the intersection is computed between a closed polygon formed by specific landmarks and the left eyebrow mask by face parsing. Region Masks for other Atomic Concepts can be similarly produced through analogous predefined procedures. Region Masks for Parent Concepts can be obtained by performing the union operation on the masks of their constituent Atomic Concepts.

3.2 FiFa-Annotator

Building upon the FICT, we construct a novel annotation pipeline termed FiFa-Annotator. The detailed procedure for generating data for the AGE task is outlined below. Data for the Loc. and TOE are concurrently generated.

Step 1: Whole-Image Artifact Mask Extraction. We compute per-pixel differences in RGB channels between real and manipulated facial images, converting the result to grayscale. A binary Whole-Image Artifact Mask is generated by thresholding the top 5% of intensity values (see the second row of Fig. 2(d)).

Step 2: Identification of Artifact-Existing Concepts. Leveraging the Region Mask derived in Sec. 3.1, we can compute the Regional Artifact Mask for each concept via intersection with the Whole-Image Artifact Mask (see the second row of Fig. 2(d)). The artifact coverage ratio is defined as the proportion of artifact pixels within each concept's Region Mask. Atomic/Parent Concepts are flagged as artifact-existing if they simultaneously satisfy: (1) Rank $\leq X$ in artifact coverage ratio descending order. (2) Artifact coverage ratio $\geq Y\%$. (3) Number of artifact pixels $\geq Z$. In our implementation, X, Y, and Z are set to 50, 10, and 50, respectively. The concepts of "whole facial image", "foreground", "region around head", and "region around face" are included in the artifact-existing Parent Concepts by default.

Step 3: Forgery Explanation Generation for Atomic Concepts. For each identified artifact-existing Atomic Concept, GPT-4o (OpenAI, 2024) synthesizes detailed forgery descriptions using well-designed prompts (refer to Appendix A.3). Our prompt refers to forgery analysis perspectives adapted from prior work (Xu et al., 2025).

Step 4: Forgery Explanation Synthesis for Parent Concepts. Following hierarchical relationships in FICT, ChatGPT (OpenAI, 2023) aggregates forgery explanations for Atomic Concepts to generate coherent forgery explanations for artifact-existing Parent Concepts, with well-designed prompts (refer to Appendix A.3). In this way, each Atomic Concept referenced in explanations for Parent Concepts has a precomputed Regional Artifact Mask, enabling I-AGE/R-AGE tasks. We further compute bounding boxes for Region Masks of Parent Concepts to support B-AGE tasks.

Step 5: B-AGE Data Augmentation. To enhance fine-grained forgery analysis, we augment B-AGE data by: (1) Randomly generating 20 bounding boxes per image. (2) Retaining boxes encompassing ≥ 3 artifact-existing Atomic Concepts. (3) Synthesizing Box-Level forgery explanations using a prompt similar to that in Step 4.

3.3 Comparison of Data Annotation Pipelines

Figure 2 compares FiFa-Annotator (a) with existing annotation pipelines (b) on the same sample. Manual annotation (DD-VQA (Zhang et al., 2024b)) accurately identifies artifact regions but lacks detailed analysis. GPT-reliant methods such as FFAA (Huang et al., 2024) and FakeShield (Xu et al., 2025) frequently hallucinate non-existent artifacts. FFTG (Sun et al., 2025) pre-determines artifact locations using landmarks but has coarser spatial granularity than ours. By leveraging a richer set of facial concepts, FiFa-Annotator achieves greater precision and granularity.

3.4 FiFa-Instruct-1M and FiFa-Bench

We used FiFa-Annotator to generate data covering the FiFa-11 task set using the source data from DFFD (Dang et al., 2020). The training data is named FiFa-Instruct-1M, including 1.38M QA-pairs.

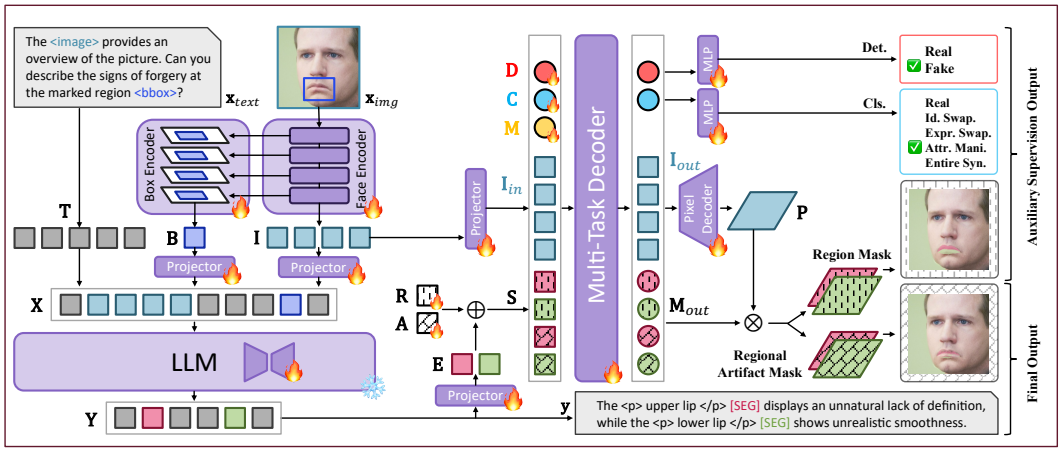

Figure 3: Overall architecture for the proposed FiFa-VLM

This is the largest-scale training dataset for the XDFA field currently known. FiFa-Bench serves as the evaluation benchmark and comprises development and test splits. Detailed per-task data sources and split statistics are provided in Appendix A.4.

4 FIFA-VLM

FiFa-VLM supports queries about arbitrary facial regions and links textual forgery explanations to visual artifact evidence. Below, we describe its architecture and auxiliary supervision through Region Mask Prediction.

4.1 MODEL ARCHITECTURE

The proposed FiFa-VLM (as shown in Fig. 3) comprises five core components: (1) Face Encoder, (2) Box Encoder, (3) Large Language Model (LLM), (4) Multi-Task Decoder, and (5) Pixel Decoder. Several projectors are incorporated to transform features across different representation spaces. Our model is meticulously designed to process both textual and optional visual prompts (bounding boxes), allowing for interaction at multiple levels of granularity and generating Artifact-Grounding text responses.

Facial Image Encoding. To achieve robust facial image encoding, we utilize a 12-layer ViT-B (Dosovitskiy et al., 2021) as the face encoder, pre-trained via the FaRL (Zheng et al., 2022b) framework. Given an input facial image $\mathbf{x}_{img}$, the visual encoder produces:

$$\mathbf{I} = \text{FaceEncoder}(\mathbf{x}_{img}) \in \mathbb{R}^{L_I \times D_I}. \tag{1}$$

Subsequently, $\mathbf{I}$ is projected into the natural language space, yielding $\mathbf{I}' \in \mathbb{R}^{L_I \times D_{llm}}$.

Bounding Box Visual Prompt Encoding. To enhance fine-grained facial context understanding and support bounding box visual prompts, we introduce the Box Encoder. This module constructs a hierarchical feature pyramid from four selected face encoder layers, then employs RoIAlign (He et al., 2017) to generate a feature map. Aggregating these features produces a unified representation:

$$\mathbf{B} = \text{BoxEncoder}(\mathbf{x}_{img}, bbox) \in \mathbb{R}^{1 \times D_B}, \tag{2}$$

which is then projected to the natural language space as $\mathbf{B}' \in \mathbb{R}^{1 \times D_{llm}}$.

Interleaved Vision-Language Sequence Processing. The input text prompt $\mathbf{x}_{text}$ is tokenized into $\mathbf{T} \in \mathbb{R}^{L_T \times D_{llm}}$. The special tokens "<image>" and "<bbox>" in $\mathbf{T}$ are replaced with $\mathbf{I}'$ and $\mathbf{B}'$ respectively, forming an interleaved sequence: $\mathbf{X} \in \mathbb{R}^{L_X \times D_{llm}}$. The LLM generates the output token sequence:

$$\mathbf{Y} = \text{LLM}(\mathbf{X}) \in \mathbb{R}^{L_Y \times D_{llm}}, \tag{3}$$

which is then decoded into human-readable text $\mathbf{y}$.

Multi-Task Processing. The Multi-Task Decoder adopts a two-way transformer architecture (Kirillov et al., 2023). It can process two types of tokens: image tokens and task-specific query tokens:

$$\mathbf{I}_{out}, \mathbf{T}_{out} = \text{MultiTaskDecoder}(\mathbf{I}_{in}, \mathbf{T}_{in}). \tag{4}$$

Here, $\mathbf{I}_{in} \in \mathbb{R}^{L_I \times D_{de}}$ is derived by projecting $\mathbf{I}$. During training, different embeddings are selected as $\mathbf{T}_{in}$ based on the task that the current training sample belongs to. For Det., $\mathbf{T}_{in}$ is set to the detection embedding $\mathbf{D} \in \mathbb{R}^{1 \times D_{de}}$; for Cls., $\mathbf{T}_{in}$ is set to the classification embedding $\mathbf{C} \in \mathbb{R}^{1 \times D_{de}}$; for Loc. or AGE, $\mathbf{T}_{in}$ is set to the mask embedding $\mathbf{M} \in \mathbb{R}^{1 \times D_{de}}$ and the semantic embedding $\mathbf{S} \in \mathbb{R}^{1 \times D_{de}}$. $\mathbf{D}$, $\mathbf{C}$, and $\mathbf{M}$ are all randomly initialized for training.

Output Decoding. For the sample of the Det. and Cls. tasks, $\mathbf{T}_{out} \in \mathbb{R}^{1 \times D_{de}}$ is fed into a 2-layer MLP to obtain predictions. For the sample of the Loc. and AGE tasks, there are some special tokens in the output text $\mathbf{y}$: " " and " " denote the starting and ending points of Atomic Concepts in the forgery explanations, while "[SEG]" indicates where an Artifact Mask should be output. The output token in $\mathbf{Y}$ corresponding to "[SEG]" is projected as Concept Embedding $\mathbf{E} \in \mathbb{R}^{1 \times D_{de}}$ and we use $\mathbf{E}$ as the semantic embedding $\mathbf{S}$ in this section. The output of image tokens $\mathbf{I}_{out}$ is reshaped and upsampled by the Pixel Decoder, obtaining:

$$\mathbf{P} = \text{PixelDecoder}(\mathbf{I}_{out}) \in \mathbb{R}^{D_{de} \times H \times W}. \tag{5}$$

The Regional Artifact Mask for an Atomic Concept is computed via the dot product between $\mathbf{P}$ and the corresponding output of the mask embedding, $\mathbf{M}_{out}$. For simplicity, we have formalized the Mask Prediction process assuming the output text $\mathbf{y}$ contains only one occurrence of "[SEG]". In practice, the input $\mathbf{I}_{in}$ is duplicated based on the number of times "[SEG]" appears to generate the masks for all Atomic Concepts simultaneously.

Notably, during the inference stage, FiFa-VLM's final outputs only include LLM-generated text $\mathbf{y}$ for all tasks and the Regional Artifact Mask for Loc. and AGE tasks. The auxiliary supervision of Det. and Cls. is aimed at enhancing the Face Encoder's forgery understanding capability.

4.2 Auxiliary Supervision of Region Mask Prediction

In Sec. 4.1, we predict the Regional Artifact Mask for each Atomic Concept via Concept Embedding $\mathbf{E}$. As detailed in Sec. 3.2's data annotation pipeline, Region Masks are acquired alongside Regional Artifact Masks for each Atomic Concept. In this section, we incorporate Region Mask Prediction as auxiliary supervision to improve the accuracy of the Artifact Mask Prediction and enhance the fine-grained Facext understanding. To implement this, we introduce randomly initialized Region Embedding $\mathbf{R} \in \mathbb{R}^{1 \times D_{de}}$ and Artifact Embedding $\mathbf{A} \in \mathbb{R}^{1 \times D_{de}}$. The Concept Embedding corresponding to each Atomic Concept is combined with both the Region Embedding and Artifact Embedding, forming two distinct Semantic Embeddings. This enables the simultaneous generation of both a Region Mask and a Regional Artifact Mask for every Atomic Concept. The efficacy of this auxiliary supervision is empirically validated via ablation studies in Sec. 5.4.

5 Experiments

5.1 Implementation Details

We implement FiFa-VLM with PyTorch. The Face Encoder is initialized from a ViT-B pre-trained with the FaRL (Zheng et al., 2022b) framework. The Box Encoder adopts the RoIAlign (He et al., 2017) structure. For the LLM, we utilize the Vicuna (Zheng et al., 2023) LLM with 7B parameters, initialized from the LLM weights of GLaMM-GranD-Pretrained (Rasheed et al., 2024). The Multi-Task Decoder employs the lightweight two-way transformer decoder proposed in SAM (Kirillov et al., 2023). The Pixel Decoder follows Faceptor (Qin et al., 2024) and is configured with two consecutive 2×2 deconvolutional layers. Projectors are implemented as two-layer MLPs with GELU activation functions. For the experimental hyperparameters of the training process, please refer to Appendix B.1.

Table 2: Comparison on FiFa-Bench

Method	Det.			Cls.	I-TOE	I-Loc.	I-AGE	
	ACER ↓	Acc. ↑	F1 ↑	Acc. ↑	METEOR ↑	mIoU ↑	METEOR ↑	mIoU ↑
GLaMM (Multi-Task)	6.97	93.03	92.90	90.48	21.3	20.1	21.4	22.5
FiFa-VLM (Single-Task)	16.59	83.41	83.75	84.29	20.9	25.6	21.2	23.7
FiFa-VLM (Multi-Task)	**4.47**	**95.53**	**95.59**	**93.00**	**23.0**	**31.6**	**23.0**	**29.6**

Method	R-TOE	R-Loc.	R-AGE		B-TOE	B-Loc.	B-AGE	
	METEOR ↑	mIoU ↑	METEOR ↑	mIoU ↑	METEOR ↑	mIoU ↑	METEOR ↑	mIoU ↑
GLaMM (Multi-Task)	18.7	24.9	20.7	25.0	21.2	24.3	20.5	27.2
FiFa-VLM (Single-Task)	18.5	23.0	21.0	23.9	19.1	24.5	20.1	24.7
FiFa-VLM (Multi-Task)	**19.7**	**30.6**	**21.5**	**31.0**	**21.5**	**30.4**	20.3	**30.7**

Table 3: Comparison on DD-VQA

Method	Det.				I-TOE			
	Acc. ↑	Recall ↑	Precision ↑	F1 ↑	BLEU-4 ↑	CIDEr ↑	ROUGE_L ↑	METEOR ↑
DDVQA-BLIP-TI	87.49	93.41	86.97	90.07	40.8	2.057	60.9	34.6
FiFa-VLM (Ours)	**88.76**	**95.24**	**87.72**	**91.32**	**48.1**	**2.869**	**65.4**	**40.4**

Table 4: Comparison on DFA-Bench

Method	Det.	Cls.	I-TOE
	ACER ↓	Acc. ↑	ROUGE-L ↑
DFA-GPT	5.04	92.74	42.54
FiFa-VLM (Ours)	**4.54**	**92.95**	**42.78**

Table 5: Superiority of FiFa-Annotator

I-TOE Data Source	ACER	Precision	Recall
None	7.71	-	-
DD-VQA (Human)	7.23	-	-
GPT-4o	6.91	52.75	54.19
FiFa-Annotator (Ours)	**6.31**	74.18	76.68

5.2 Performance Evaluation

We report the performance of the multi-task models on the test set of FiFa-Bench in Tab. 2. For a strong baseline, we fine-tuned GLaMM with frozen vision encoders, using the same training hyperparameters. Our proposed FiFa-VLM achieves significantly superior performance over GLaMM on nearly all evaluation metrics. We further compare FiFa-VLM's performance under multi-task and single-task settings. To ensure fair comparison, all single-task models are trained with sample amounts identical to their corresponding subsets in the multi-task setting. Results demonstrate that multi-task learning effectively enhances data efficiency, yielding better-performing models while utilizing equivalent training data. We also evaluate the FiFa-VLM on existing XDFA datasets: DD-VQA (Zhang et al., 2024b) and DFA-Bench (Qin et al., 2025), as shown in Tabs. 3 and 4. Our model achieves competitive performance across both benchmarks, with particularly notable gains on the I-TOE task of DD-VQA. These quantitative results substantiate the effectiveness of our architecture design.

5.3 Superiority of FiFa-Annotator

Following the evaluation protocols in FFTG (Sun et al., 2025), we first utilize our FiFa-Annotator to generate I-TOE data for the high-quality version of the FF++ (Rössler et al., 2019) dataset. We then follow the three-branch training framework from FFTG (Sun et al., 2025) to fine-tune the CLIP (Radford et al., 2021) model using multimodal data. During testing, results on FF++ represent in-domain performance, while results on DFD, DFDC-P (Dolhansky et al., 2020), and Celeb-DF (Li et al., 2020b) represent cross-domain performance. The experimental results in Tab. 6 demonstrate that by training with I-TOE data constructed via the FiFa-Annotator, the CLIP model achieves results across most metrics that surpass traditional methods as well as the CLIP model fine-tuned with I-TOE data constructed by FFTG. The complete comparison is provided in Appendix B.2.

Table 5 further employs FiFa-VLM as the training model to investigate performance when not combining I-TOE data, combining I-TOE data from DD-VQA (Zhang et al., 2024b), combining I-TOE data directly generated by GPT-4o (OpenAI, 2024) (prompts refer to Appendix A.3), and combining I-TOE data generated by our FiFa-Annotator. The experimental results indicate that the I-TOE data generated by our FiFa-Annotator yields the most significant performance improvement for DeepFake Detection. In Table 5, ACER evaluates detection, whereas precision and recall evaluate the correctness and coverage of Artifact-Existing Concepts mentioned in FiFa-VLM's generated explanations. The ground-truth concepts are obtained via Step 2 in Section 3.2. Training with FiFa-Annotator's I-TOE data yields higher concept precision and recall than training with GPT-4o's annotations.

Table 6: Frame-level cross-dataset evaluation with selected methods. FF++ (HQ) is in-domain; DFD, DFDC-P, and Celeb-DF are cross-domain.

Methods	FF++		DFD		DFDC-P		Celeb-DF	
	AUC ↑	EER ↓	AUC ↑	EER ↓	AUC ↑	EER ↓	AUC ↑	EER ↓
LRL (Chen et al., 2021)	**99.46**	**3.01**	89.24	20.32	76.53	32.41	78.26	29.67
DCL (Sun et al., 2022)	99.30	3.26	91.66	16.63	76.71	31.97	82.30	26.53
SBI (Shiohara & Yamasaki, 2022)	88.33	20.47	88.13	17.25	76.53	30.22	80.76	26.97
RECCE (Cao et al., 2022)	99.32	3.38	89.91	19.95	75.88	32.41	70.50	35.34
CLIP (Radford et al., 2021)	99.09	3.16	89.03	17.13	78.83	28.95	77.16	29.30
FFTG (Sun et al., 2025)+CLIP (Radford et al., 2021)	99.16	3.11	94.81	**15.22**	83.21	22.43	83.15	23.66
FiFa-Annotator+CLIP (Radford et al., 2021)	99.30	3.09	**95.04**	15.31	**83.65**	**21.04**	**83.70**	**23.42**

Table 7: Model design ablation for FiFa-VLM. Both GLaMM settings train the LLM. TOE, Loc., and AGE scores are averaged over image, region, and box levels.

		GLaMM		FiFa-VLM			
		Vision frozen	Vision trainable	Variant 1	Variant 2	Variant 3	**Full**
Vision Encoder	Face-pretrained	✗	✗	✓	✓	✓	✓
	Single encoder	✗	✗	✓	✓	✓	✓
	Trainable	✗	✓	✗	✓	✓	✓
Auxiliary Supervision	Det. and Cls.	✗	✗	✗	✗	✓	✓
	Region Mask	✗	✗	✗	✗	✗	✓
Det.	ACER ↓	6.97	6.33	5.21	5.07	4.79	**4.47**
Cls.	ACC ↑	90.48	90.64	90.74	91.97	92.72	**93.00**
TOE	METEOR ↑	20.4	20.8	19.8	21.1	21.3	**21.4**
Loc.	mIoU ↑	23.1	26.3	24.6	29.8	30.0	**30.9**
AGE	METEOR ↑	20.9	21.4	20.4	21.6	**21.7**	21.6
	mIoU ↑	24.9	26.5	21.0	29.9	30.0	**30.4**

Table 8: Data Ablation on FiFa-11

Training Data	Det.	Det.-TOE	Det.-TOE-Loc.	Det.-I	Det.-I-R	All
Det. (Intra-Domain) ACER ↓	0.90	0.70	0.35	0.61	0.33	**0.30**
Det. (Cross-Domain) ACER ↓	29.32	29.08	28.79	28.94	28.89	**28.57**

5.4 Analysis Experiment

Model Design Ablation for FiFa-VLM. We ablate FiFa-VLM's design in Table 7. (1) Variant 1 replaces GLaMM's (Rasheed et al., 2024) frozen CLIP-ViT-H/14 and SAM-ViT-H/16 with a frozen FaRL-ViT-B/16, saving 0.94B parameters and improving Det., Cls., and Loc. (2) Variant 2 unfreezes FaRL, outperforming GLaMM with both vision encoders unfrozen across all tasks. (3) Variant 3 adds task-specific query embeddings for Det. and Cls. auxiliary supervision, improving both tasks over Variant 2. (4) The full FiFa-VLM adds Region Mask Prediction supervision, improving all reported metrics over Variant 3 except AGE METEOR, which decreases slightly. We attribute the gains on global Det. and Cls. to improved capture of regional forgery clues.

Data Ablation on FiFa-11. In Table 8, we set the training data ratio between Det. and the other tasks to 1:1, keeping the relative ratios among other tasks as in the multi-task setting of Section 5.2 (Appendix B.1). Det. training uses only FFHQ and FaceAPP, which also provide intra-domain tests; other subsets provide cross-domain tests. Progressively adding TOE, Loc., and AGE, or extending their supervision from Image-Level to Region-Level and Box-Level, improves detection. Combining these three tasks at all three levels yields the best detection performance, demonstrating that fine-grained XDFA enhances forgery understanding.

Visualization Results. Qualitative visualizations of predicted Artifact-Existing Concepts and masks, together with output examples for the Loc. and AGE tasks, are provided in Appendix C.

6 Conclusion

We propose the **Fake-in-Facext (FiFa)** framework to enhance VLMs' fine-grained awareness for XDFA through Face Context (Facext) priors. First, we develop **FiFa-Annotator** to automatically generate precise forgery explanations with Regional Artifact Masks. These annotations can also improve detection generalization. Second, we introduce **FiFa-11**, including the novel Artifact-Grounding Explanation task and tasks with bounding box visual prompts. We contribute **FiFa-Instruct-1M** and **FiFa-Bench** as high-quality training and evaluation resources. Finally, we design and train **FiFa-VLM** to support all FiFa-11 tasks, substantially outperforming the strong baseline on almost all tasks.

AI USE STATEMENT

We used generative AI for data annotation, as described in Section 3.2, and for editing and polishing the manuscript.

REFERENCES

Junyi Cao, Chao Ma, Taiping Yao, Shen Chen, Shouhong Ding, and Xiaokang Yang. End-to-end reconstruction-classification learning for face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4113–4122, 2022.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV (1)*, volume 12346 of *Lecture Notes in Computer Science*, pp. 213–229. Springer, 2020. doi: 10.1007/978-3-030-58452-8_13.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *CoRR*, abs/2306.15195, 2023.

Shen Chen, Taiping Yao, Yang Chen, Shouhong Ding, Jilin Li, and Rongrong Ji. Local relation learning for face forgery detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 1081–1088, 2021.

Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *NeurIPS*, pp. 17864–17875, 2021.

François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil K. Jain. On the detection of digital face manipulation. In *CVPR*, pp. 5780–5789. Computer Vision Foundation / IEEE, 2020.

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

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*. OpenReview.net, 2021.

Xiao Guo, Xiufeng Song, Yue Zhang, Xiaohong Liu, and Xiaoming Liu. Rethinking vision-language model in face forensics: Multi-modal interpretable forged face detector. In *CVPR*, pp. 105–116. Computer Vision Foundation / IEEE, 2025.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

Zhengchao Huang, Bin Xia, Zicheng Lin, Zhun Mou, and Wenming Yang. FFAA: multimodal large language model based explainable open-world face forgery analysis assistant. *CoRR*, abs/2408.10072, 2024.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. LISA: reasoning segmentation via large language model. In *CVPR*, pp. 9579–9589. IEEE, 2024.

Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *CVPR*, pp. 5548–5557. Computer Vision Foundation / IEEE, 2020. doi: 10.1109/CVPR42600.2020.00559.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5001–5010, 2020a.

Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3207–3216, 2020b.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. Mediapipe: A framework for building perception pipelines. *CoRR*, abs/1906.08172, 2019.

Yuchen Luo, Yong Zhang, Junchi Yan, and Wei Liu. Generalizing face forgery detection with high-frequency features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16317–16326, 2021.

Kartik Narayan, Vibashan VS, Rama Chellappa, and Vishal M. Patel. Facexformer: A unified transformer for facial analysis. *CoRR*, abs/2403.12960, 2024.

OpenAI. Chatgpt. `https://openai.com/blog/chatgpt/`, 2023.

OpenAI. Hello gpt-4o, 2024. URL `https://openai.com/index/hello-gpt-4o/`.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *CoRR*, abs/2306.14824, 2023.

Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European conference on computer vision*, pp. 86–103. Springer, 2020.

Lixiong Qin, Mei Wang, Xuannan Liu, Yuhang Zhang, Wei Deng, Xiaoshuai Song, Weiran Xu, and Weihong Deng. Faceptor: A generalist model for face perception. In *ECCV (34)*, volume 15092 of *Lecture Notes in Computer Science*, pp. 240–260. Springer, 2024.

Lixiong Qin, Ning Jiang, Yang Zhang, Yuhan Qiu, Dingheng Zeng, Jiani Hu, and Weihong Deng. Towards interactive deepfake analysis. In *ICASSP*, pp. 1–5. IEEE, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Hanoona Abdul Rasheed, Muhammad Maaz, Sahal Shaji Mullappilly, Abdelrahman M. Shaker, Salman H. Khan, Hisham Cholakkal, Rao Muhammad Anwer, Eric P. Xing, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Glamm: Pixel grounding large multimodal model. In *CVPR*, pp. 13009–13018. IEEE, 2024.

Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *ICCV*, pp. 1–11. IEEE, 2019.

Kaede Shiohara and Toshihiko Yamasaki. Detecting deepfakes with self-blended images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18720–18729, 2022.

Haomiao Sun, Mingjie He, Shiguang Shan, Hu Han, and Xilin Chen. Task-adaptive q-face. *CoRR*, abs/2405.09059, 2024.

Ke Sun, Hong Liu, Qixiang Ye, Yue Gao, Jianzhuang Liu, Ling Shao, and Rongrong Ji. Domain general face forgery detection by learning to weight. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 2638–2646, 2021.

Ke Sun, Taiping Yao, Shen Chen, Shouhong Ding, Jilin Li, and Rongrong Ji. Dual contrastive learning for general face forgery detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 2316–2324, 2022.

Ke Sun, Shen Chen, Taiping Yao, Ziyin Zhou, Jiayi Ji, Xiaoshuai Sun, Chia-Wen Lin, and Rongrong Ji. Towards general visual-linguistic face forgery detection. In *CVPR*, pp. 19576–19586. Computer Vision Foundation / IEEE, 2025.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pp. 6105–6114. PMLR, 2019.

Ruben Tolosana, Rubén Vera-Rodríguez, Julian Fiérrez, Aythami Morales, and Javier Ortega-Garcia. Deepfakes and beyond: A survey of face manipulation and fake detection. *Inf. Fusion*, 64:131–148, 2020.

W Turton and A Martin. How deepfakes make disinformation more real than ever. bloomberg, 2020.

Zhipei Xu, Xuanyu Zhang, Runyi Li, Zecheng Tang, Qing Huang, and Jian Zhang. Fakeshield: Explainable image forgery detection and localization via multi-modal large language models. In *ICLR*. OpenReview.net, 2025.

Zhiyuan Yan, Yong Zhang, Yanbo Fan, and Baoyuan Wu. Ucf: Uncovering common features for generalizable deepfake detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 22412–22423, 2023.

Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Yu Liu, Kai Chen, and Ping Luo. Gpt4roi: Instruction tuning large language model on region-of-interest. In *ECCV Workshops (8)*, volume 15630 of *Lecture Notes in Computer Science*, pp. 52–70. Springer, 2024a.

Yue Zhang, Ben Colman, Xiao Guo, Ali Shahriyari, and Gaurav Bharaj. Common sense reasoning for deepfake detection. In *ECCV (88)*, volume 15146 of *Lecture Notes in Computer Science*, pp. 399–415. Springer, 2024b.

Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-attentional deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2185–2194, 2021a.

Tianchen Zhao, Xiang Xu, Mingze Xu, Hui Ding, Yuanjun Xiong, and Wei Xia. Learning self-consistency for deepfake detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 15023–15033, 2021b.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In *NeurIPS*, 2023.

Qingping Zheng, Jiankang Deng, Zheng Zhu, Ying Li, and Stefanos Zafeiriou. Decoupled multi-task learning with cyclical self-regulation for face parsing. In *CVPR*, pp. 4146–4155. IEEE, 2022a.

Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. General facial representation learning in a visual-linguistic manner. In *CVPR*, pp. 18676–18688. IEEE, 2022b. doi: 10.1109/CVPR52688.2022.01814.

Wanyi Zhuang, Qi Chu, Zhentao Tan, Qiankun Liu, Haojie Yuan, Changtao Miao, Zixiang Luo, and Nenghai Yu. Uia-vit: Unsupervised inconsistency-aware method based on vision transformer for face forgery detection. In *European conference on computer vision*, pp. 391–407. Springer, 2022.

A Additional Details for Data Annotation Pipeline

A.1 Facial Image Concept Tree

We have provided all the Atomic Concepts and Parent Concepts involved in the Facial Image Concept Tree in Tables 9 to 12.

Table 9: Facial Image Concept Tree (FICT). Due to page size limitations, we have divided FICT into four tables. This table presents concepts from Level 1 (L1) to Level 4 (L4). Atomic Concepts are highlighted in pink.

L1	L2	L3	L4
whole facial image	foreground	region around head	region around hair region around face
		region around neck	neck edges of neck
		region around adornment and clothing	region around adornments region around clothing
	background	left part of background right part of background	

Table 10: Sub-FICT under the Concept "region around hair (L4)". Atomic Concepts are highlighted in pink.

L5	L6	L7	L8
hair	hair near face	hair near left part of face	hair near left part of forehead hair near left temple hair near left ear hair near left cheek hair near left part of jaws
		hair near right part of face	hair near right part of forehead hair near right temple hair near right ear hair near right cheek hair near right part of jaws
	outer hair	left part of outer hair right part of outer hair	
edges of hair	inner edge of hair outer edge of hair		

Table 11: Sub-FICT under the Concept ”region around adornment and clothing (L3)”. Atomic Concepts are highlighted in pink.

L4	L5	L6
region around adornments	hat	
	edges of hat	
	eyeglasses	
	edges of eyeglasses	
	earrings	left earring
		right earring
	edges of earrings	
region around clothing	clothing	
	edges of clothing	

Table 12: Sub-FICT under the Concept ”region around face (L4)”. Atomic Concepts are highlighted in pink.

L5	**L6**	**L7**	**L8**
region around forehead	forehead	left part of forehead	upper left part of forehead
			lower left part of forehead
		right part of forehead	upper right part of forehead
			lower right part of forehead
	edges of forehead	top edge of forehead	
		bottom edge of forehead	
	temples	left temple	
		right temple	
	edges of temples	edges of left temple	
		edges of right temple	
region around eyebrows	eyebrows	left eyebrow	headstart of left eyebrow
			tail of left eyebrow
			middle part of left eyebrow
		right eyebrow	headstart of right eyebrow
			tail of right eyebrow
			middle part of right eyebrow
	edges of eyebrows	edges of left eyebrow	top edge of left eyebrow
			bottom edge of left eyebrow
		edges of right eyebrow	top edge of right eyebrow
			bottom edge of right eyebrow
	region between eyebrows		
region around eyes	eyes	left eye	iris of left eye
			sclera of left eye
		right eye	iris of right eye
			sclera of right eye
	edges of eyes	edges of left eye	top edge of left eye
			bottom edge of left eye
		edges of right eye	top edge of right eye
			bottom edge of right eye
	region between eyes		
	eyelids	eyelids of left eye	upper eyelid of left eye
			lower eyelid of left eye
		eyelids of right eye	upper eyelid of right eye
			lower eyelid of right eye
	eye bags	eye bag of left eye	
		eye bag of right eye	
	eye sockets	eye sockets of left eye	upper eye socket of left eye
			lower eye socket of left eye
		eye sockets of right eye	upper eye socket of right eye
			lower eye socket of right eye
region around ears	ears	left ear	upper part of left ear
			lower part of left ear
		right ear	upper part of right ear
			lower part of right ear
	edges of ears	edges of left ear	
		edges of right ear	
region around nose	nose	nasion	
		nose bridge	
		nasal ala	nasal ala of left side
			nasal ala of right side
		nose base	
	edges of nose		

L5	L6	L7	L8
region around cheeks	cheeks	left cheek	upper part of left cheek lower part of left cheek upper part of right cheek lower part of right cheek
		right cheek	
	edges of cheeks	edges of left cheek edges of right cheek	
region around mouth	mouth	lips	upper lip lower lip
	edges of mouth	oral cavity top edge of mouth bottom edge of mouth	
region around jaws	jaws	upper jaw	left part of upper jaw right part of upper jaw left part of lower jaw right part of lower jaw
		lower jaw	
	edges of jaws	top edge of jaws bottom edge of jaws left part of chin right part of chin	
	chin		
	edges of chin philtrum		
	nasolabial folds	nasolabial fold of left side nasolabial fold of right side	
face edge	left part of face edge	face edge near left part of forehead face edge near left temple face edge near left ear face edge near left cheek face edge near left part of jaws	
	right part of face edge	face edge near right part of forehead face edge near right temple face edge near right ear face edge near right cheek face edge near right part of jaws	

A.2 Tools of Facial Landmark Localization and Face Parsing

We briefly introduce the facial landmark localization and face parsing tools used in our FiFa-Annotator.

Facial Landmark Localization. We use MediaPipe (Lugaresi et al., 2019), developed and open-sourced by Google Research, for facial landmark localization in face images. MediaPipe supports the prediction of 478 dense facial landmarks.

Face Parsing. We employ the DML-CSR (Zheng et al., 2022a) model, trained on CelebA-Mask-HQ (Lee et al., 2020), to perform face region segmentation. CelebA-Mask-HQ provides 19 classes, including all facial components and accessories such as skin, nose, eyes, eyebrows, ears, mouth, lip, hair, hat, eyeglass, earring, necklace, neck, and cloth.

A.3 Prompts Used in FiFa-Annotator

We provide the prompts used in the FiFa-Annotator. The prompts used in Step 3 are shown in Table 13; the prompts used in Steps 4 and 5 are shown in Table 14. The prompts we used for generating detailed forgery explanations are referenced from the prompts in MMTD-SET (Xu et al., 2025) when setting the analysis perspective. In Section 5.3 of our paper, to demonstrate the superiority of the FiFa-Annotator, we directly use GPT-4o (OpenAI, 2024) to generate data similar to existing approaches (Xu et al., 2025; Huang et al., 2024; Qin et al., 2025), but with the same forgery sample sources as our FiFa-Instruct-1M. The prompt used is shown in Table 15.

Table 13: The prompt used for Step 3 of the FiFa-Annotator. "{ concept_list }" is the list of Artifact-Existing Atomic Concepts.

> You are an AI visual assistant that helps humans analyze some images of human faces that have been tampered with by deepfake. You will receive two images, the first is the image A of the face tampered by deepfake and the second is the binary mask image B of the tampered area (a value of 1 (white) indicates the tampered area and a value of 0 (black) indicates the untampered area).

Now, your task is to describe the visible details (artifacts) of the tampered area in the image, using the binary masks provided for the tampered regions of the deepfake tampered image A and image B. In your answer, please describe the tampered image based on the highlighted area of the binary mask, but do not mention the binary mask. Always assume that you are just looking at the tampered image. Responses should identify the tampered areas and corresponding detailed descriptions of the forgery in those areas. Use the format [tampered area]: [detailed forgery description].

All the following areas MUST be discussed:
{ concept_list }
Please analyze each area in one sentence.

When providing detailed forgery descriptions, consider the visible details caused by tampering from these perspectives, but do not give an ambiguous, unclear description that is otherwise challenging:
1. Symmetrical Facial Features: Deepfake-generated faces may exhibit unnaturally perfect symmetry, lacking the subtle asymmetry typically found in real faces.
2. Blur or Distortion Around Edges: Deepfake manipulation may introduce blur or distortion around the edges of the face where the manipulation has taken place, especially if the face has been digitally overlaid onto another body.
3. Inconsistent Lighting and Shadows: Deepfake algorithms may struggle to accurately match the lighting and shadows in the original image, leading to discrepancies in lighting direction or intensity across the face.
4. Unnatural Facial Expressions: Deepfakes may produce facial expressions that appear unnatural, exaggerated, or out of sync with the rest of the image.
5. Mismatched Facial Proportions: Deepfake manipulation may result in facial proportions that are inconsistent with the person's gender or age, such as a man's face on a woman's body or vice versa.
6. Inconsistent Skin Texture and Tone: Deepfake-generated faces may exhibit unnatural skin texture or tone, such as overly smooth or pixelated skin, that differs from the surrounding areas.
7. Missing or Inconsistent Eye Reflections: Deepfake manipulation may result in missing or inconsistent reflections in the eyes, which can provide clues about the authenticity of the image.
8. Change hairstyle: Some deepfake algorithms only tamper with hair and may change hair color and hairstyle, or add long hair for boys and short hair for girls.
9. Irregularities in Makeup Application: Deepfake manipulation may introduce makeup styles that are inconsistent with the person's gender or age, or exhibit poor application quality.
10. Contextual Inconsistencies: Deepfake-generated images may contain inconsistencies in the overall context of the image, such as discrepancies in perspective, clothing, or surroundings, that suggest manipulation.
11. Unreasonable accessories: Some deepfake algorithms add glasses, earrings, hats, masks, etc. to the image, and the edges, lighting relationships, and perspective of these accessories may be incorrect.
12. If there are glasses or sunglasses in the picture, please pay special attention to whether the glasses or sunglasses have been tampered with or not, the rims, frames, temples, lenses, etc. of the glasses are very susceptible to imperfections and problems.

Table 14: The prompt used for Steps 4 and 5 of the FiFa-Annotator.

Summarize the descriptions of face forgery areas into a paragraph to make it as clear as possible. The face region words in the summary can be more diverse, but the semantic consistency must be guaranteed. An index number will be provided for each face forgery

area. Please indicate the index number in the summary.

Here is an example:

Input:
Descriptions
left part of chin: Displays inconsistency in skin texture, appearing artificially uniform.
right part of chin: Exhibits disproportionate shadowing inconsistent with light sources.
edges of chin: Subtle distortions suggest digital manipulation at the boundaries.
nasolabial fold of left side: Appears unnaturally softened, lacking typical depth.
nasolabial fold of right side: Exhibits lack of natural shading and depth variation.
Index Numbers
left part of chin: 0
right part of chin: 1
edges of chin: 2
nasolabial fold of left side: 3
nasolabial fold of right side: 4

Summary:
The <0> left side of the chin </0> reveals an unnaturally uniform skin texture, while the <1> right chin area </1> exhibits disproportionate shadowing that contradicts expected lighting. The <2> chin's edges </2> display subtle distortions, indicating possible digital manipulation. Additionally, the <3> left nasolabial fold </3> appears overly smoothed, lacking natural depth, whereas the <4> right nasolabial region </4> fails to show proper shading and depth variation.

Input: [input]

Table 15: The prompt used for GPT-4o (OpenAI, 2024).

You are an AI visual assistant that helps humans analyze some images of human faces that have been tampered with by deepfake. You will receive two images, the first is the image A of the face tampered by deepfake and the second is the binary mask image B of the tampered area (a value of 1 (white) indicates the tampered area and a value of 0 (black) indicates the untampered area).

Now, your task is to describe the visible details (artifacts) of the tampered area in the image, using the binary masks provided for the tampered regions of the deepfake tampered image A and image B. In your answer, please describe the tampered image based on the highlighted area of the binary mask, but do not mention the binary mask. Always assume that you are just looking at the tampered image. Responses should identify the tampered areas and corresponding detailed descriptions of the forgery in those areas. Use the format [tampered area]: [detailed forgery description].

Here are some phrases to describe tampered areas for reference: lower eye socket of left eye, tail of right eyebrow, eyeglasses, upper eyelid of left eye,edges of eyeglasses, region between eyes, lower eyelid of right eye, upper eye socket of right eye, top edge of left eye, lower eye socket of right eye, eye bag of right eye, upper eyelid of right eye, sclera of right eye, face edge near left temple,sclera of left eye, edges of right temple, face edge near right temple, face edge near left ear, iris of left eye, lower eyelid of left eye, nasion, upper eye socket of left eye, bottom edge of right eyebrow, edges of nose, bottom edge of right eye,bottom edge of left eye, edges of left cheek, eye bag of left eye, top edge of right eye,hair near left part of forehead, right temple, hair near left temple, inner edge of hair, face edge near left part of forehead, bottom edge of left eyebrow, face edge near right ear, face edge near left cheek, iris of right eye, top edge of right eyebrow,hair near left ear, hair near left part of jaws, edges of right cheek, upper part of left cheek, hair near left cheek, nose bridge, up-

per lip, oral cavity, edges of neck, upper part of right cheek, nose base, nasal ala of right side.

When providing detailed forgery descriptions, consider the visible details caused by tampering from these perspectives, but do not give an ambiguous, unclear description that is otherwise challenging:
1. Symmetrical Facial Features: Deepfake-generated faces may exhibit unnaturally perfect symmetry, lacking the subtle asymmetry typically found in real faces.
2. Blur or Distortion Around Edges: Deepfake manipulation may introduce blur or distortion around the edges of the face where the manipulation has taken place, especially if the face has been digitally overlaid onto another body.
3. Inconsistent Lighting and Shadows: Deepfake algorithms may struggle to accurately match the lighting and shadows in the original image, leading to discrepancies in lighting direction or intensity across the face.
4. Unnatural Facial Expressions: Deepfakes may produce facial expressions that appear unnatural, exaggerated, or out of sync with the rest of the image.
5. Mismatched Facial Proportions: Deepfake manipulation may result in facial proportions that are inconsistent with the person's gender or age, such as a man's face on a woman's body or vice versa.
6. Inconsistent Skin Texture and Tone: Deepfake-generated faces may exhibit unnatural skin texture or tone, such as overly smooth or pixelated skin, that differs from the surrounding areas.
7. Missing or Inconsistent Eye Reflections: Deepfake manipulation may result in missing or inconsistent reflections in the eyes, which can provide clues about the authenticity of the image.
8. Change hairstyle: Some deepfake algorithms only tamper with hair and may change hair color and hairstyle, or add long hair for boys and short hair for girls.
9. Irregularities in Makeup Application: Deepfake manipulation may introduce makeup styles that are inconsistent with the person's gender or age, or exhibit poor application quality.
10. Contextual Inconsistencies: Deepfake-generated images may contain inconsistencies in the overall context of the image, such as discrepancies in perspective, clothing, or surroundings, that suggest manipulation.
11. Unreasonable accessories: Some deepfake algorithms add glasses, earrings, hats, masks, etc. to the image, and the edges, lighting relationships, and perspective of these accessories may be incorrect.
12. If there are glasses or sunglasses in the picture, please pay special attention to whether the glasses or sunglasses have been tampered with or not, the rims, frames, temples, lenses, etc. of the glasses are very susceptible to imperfections and problems.

A.4 STATISTICS OF FIFA-INSTRUCT-1M AND FIFA-BENCH

This subsection provides the detailed statistics of FiFa-Instruct-1M and FiFa-Bench introduced in Section 3.4. Table 16 summarizes the data sources and the numbers of training, development, and test examples for each task.

Table 16: Statistics of FiFa-Instruct-1M and FiFa-Bench

Task	Data Source	Training	Dev.	Test
Det./Cls.	FFHQ/CelebA/DFFD	160000	15439	16000
I-Loc.	DFFD	10000	439	1000
R-Loc.		300791	1000	1000
B-Loc.		182170	1000	1000
I-TOE		9940	438	992
R-TOE		300744	1000	1000
B-TOE		182097	1000	1000
I-AGE		9796	431	976
R-AGE		124113	1000	1000
B-AGE		103561	1000	1000
Total		1383212	22747	24968

B Additional Experimental Details and Results

B.1 Training Configurations for FiFa-VLM

Primary Experiment for Multi-Task Setting. During training, the LLM parameters remain frozen, with fine-tuning performed using LoRA (rank set to 128). Parameters of all other modules are set as learnable. The auxiliary tasks (Det. and Cls.) employ cross-entropy loss; text generation uses autoregressive cross-entropy loss; mask prediction utilizes per-pixel binary cross-entropy loss combined with Dice loss. The weights of these four loss functions are empirically set to 0.2, 1.0, 0.5, and 2.0, respectively. Training is optimized using DeepSpeed Zero-2. The primary experiment (reported as FiFa-VLM multi-task learning) is conducted on 4 NVIDIA H800 GPUs with a batch size of 40 for 12,500 steps, employing a linear decay learning rate scheduler with a 100-step linear warm-up. The base learning rate is set to 3×10^{-4}. Table 17 details the data ratios and actual training data amounts. Configurations in experiments for Model Design Ablation for FiFa-VLM remained consistent with the primary experiment.

Single-Task Setting. To ensure fair comparison, all single-task models are trained using sample quantities identical to their corresponding subsets in the multi-task setting. Each experiment utilizes one NVIDIA H800 GPU with a batch size of 10. The number of training steps is specified in Table 17 for different tasks, while other configurations align with the primary experiment.

Experiments on DD-VQA. The training data ratio for Det. and I-TOE is set to 4:1. The experiment employs two NVIDIA H800 GPUs with a batch size of 20. The actual training data amount for Det. corresponded to 12 epochs. Other configurations remained consistent with the primary experiment.

Experiments on DFA-Bench. The training data ratio for Det., Cls., and I-TOE is set to 1:1:1. The experiment employs four NVIDIA H800 GPUs with a batch size of 40. The actual training data amount for Det. corresponded to 1 epoch. Other configurations remained consistent with the primary experiment.

Ablation Study for the Superiority of FiFa-Annotator. The data ratio between Det. and I-TOE is set to 1:1. The experiment employs one NVIDIA H800 GPU with a batch size of 10. The actual training data amount for Det. corresponded to 3 epochs. Other configurations remained consistent with the primary experiment.

Ablation Study for Data Ablation on FiFa-11. The training data ratio between Det. and other tasks is set to 1:1. The data ratios among the other tasks remain consistent with those used in the primary experiment, as detailed in Table 17. The experiment employs one NVIDIA H800 GPU with a batch size of 10. The actual training data amount for Det. corresponded to 3 epochs. Other configurations remained consistent with the primary experiment.

Table 17: The data ratios and actual training data amounts in the multi-task setting. The training steps for different tasks in the single-task setting.

Task	Data Ratio	Data Amount in FiFa-Instruct-1M	Training Data Amount for Multi-Task Model	Actual Epoch	Training Steps for Single-Task Models
Det.	4	160000	55556	0.35	5555
Cls.	4	160000	55556	0.35	5555
I-TOE	1	9940	13888	1.4	1388
I-Loc.	1	10000	13888	1.39	1388
I-AGE	2	9796	27776	2.84	2777
R-TOE	3	300744	41668	0.14	4166
R-Loc.	3	300791	41668	0.14	4166
R-AGE	6	124113	83332	0.67	8333
B-TOE	3	182097	41668	0.23	4166
B-Loc.	3	182170	41668	0.23	4166
B-AGE	6	103561	83332	0.8	8333

B.2 Complete Cross-Dataset Evaluation

Table 18 provides the complete comparison corresponding to Table 6 in Section 5.3, including all 17 methods. Table 6 presents a subset of these results.

Table 18: Frame-level cross-dataset evaluation from FF++(HQ) to DFD, DFDC-P, and Celeb-DF. The FF++ represents intra-domain, while others represent cross-domain.

Methods	FF++		DFD		DFDC-P		Celeb-DF	
	AUC ↑	EER ↓	AUC ↑	EER ↓	AUC ↑	EER ↓	AUC ↑	EER ↓
Xception (Chollet, 2017)	99.09	3.77	87.86	21.04	69.80	35.41	65.27	38.77
EN-b4 (Tan & Le, 2019)	99.22	3.36	87.37	21.99	70.12	34.54	68.52	35.61
Face X-ray (Li et al., 2020a)	87.40	-	85.60	-	70.00	-	74.20	-
F3-Net (Qian et al., 2020)	98.10	3.58	86.10	26.17	72.88	33.38	71.21	34.03
MAT (Zhao et al., 2021a)	99.27	3.35	87.58	21.73	67.34	38.31	70.65	35.83
GFF (Luo et al., 2021)	98.36	3.85	85.51	25.64	71.58	34.77	75.31	32.48
LTW (Sun et al., 2021)	99.17	3.32	88.56	20.57	74.58	33.81	77.14	29.34
LRL (Chen et al., 2021)	**99.46**	**3.01**	89.24	20.32	76.53	32.41	78.26	29.67
DCL (Sun et al., 2022)	99.30	3.26	91.66	16.63	76.71	31.97	82.30	26.53
PCL+I2G (Zhao et al., 2021b)	99.11	-	-	-	-	-	81.80	-
SBI (Shiohara & Yamasaki, 2022)	88.33	20.47	88.13	17.25	76.53	30.22	80.76	26.97
UIA-ViT (Zhuang et al., 2022)	-	-	94.68	-	75.80	-	82.41	-
RECCE (Cao et al., 2022)	99.32	3.38	89.91	19.95	75.88	32.41	70.50	35.34
UCF (Yan et al., 2023)	97.05	-	80.74	-	75.94	-	75.27	-
CLIP (Radford et al., 2021)	99.09	3.16	89.03	17.13	78.83	28.95	77.16	29.30
FFTG (Sun et al., 2025)+CLIP (Radford et al., 2021)	99.16	3.11	94.81	**15.22**	83.21	22.43	83.15	23.66
FiFa-Annotator+CLIP (Radford et al., 2021)	99.30	3.09	**95.04**	15.31	**83.65**	**21.04**	**83.70**	**23.42**

C Additional Qualitative Results

C.1 Visualization of Artifact-Existing Concepts and Masks

In Fig. 4, we present the Artifact-Existing Concepts predicted by our FiFa-VLM in the AGE task along with their corresponding masks. It can be observed that both the concepts and masks predicted by our model exhibit a substantial overlap with the ground truth (GT), thereby demonstrating the effectiveness of our method.

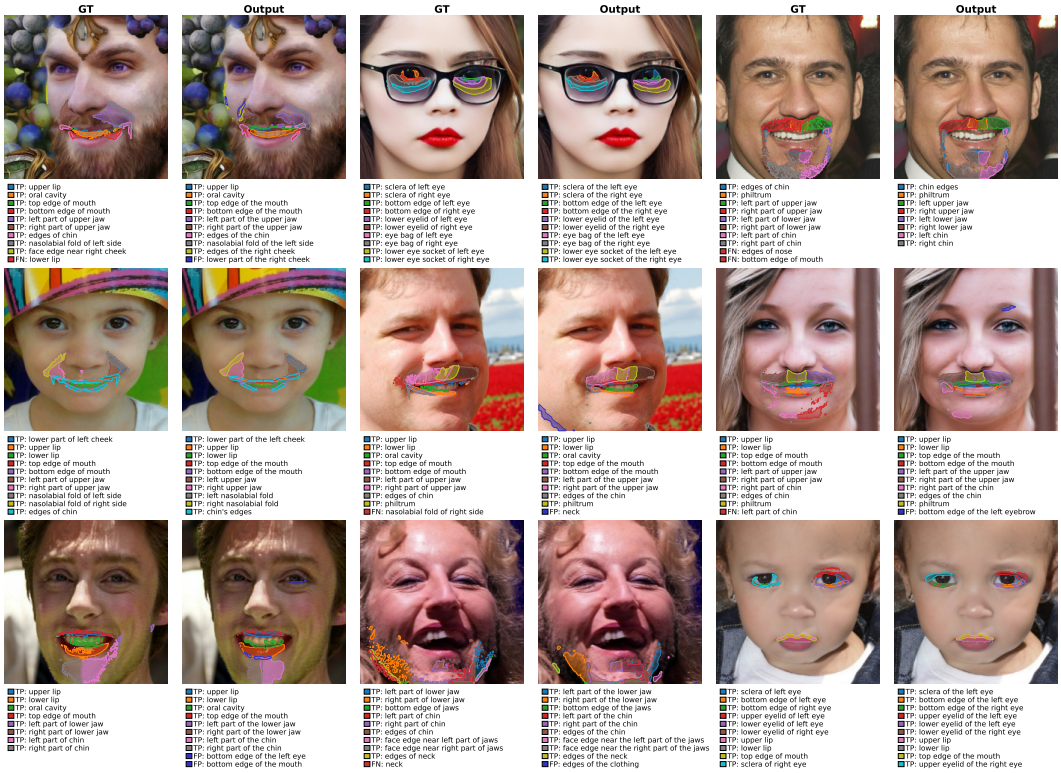

Figure 4: Visualization results

C.2 Qualitative Examples for Loc. and AGE Tasks

We provide qualitative illustrations for the Loc. and AGE tasks as shown in Figures 5-10.

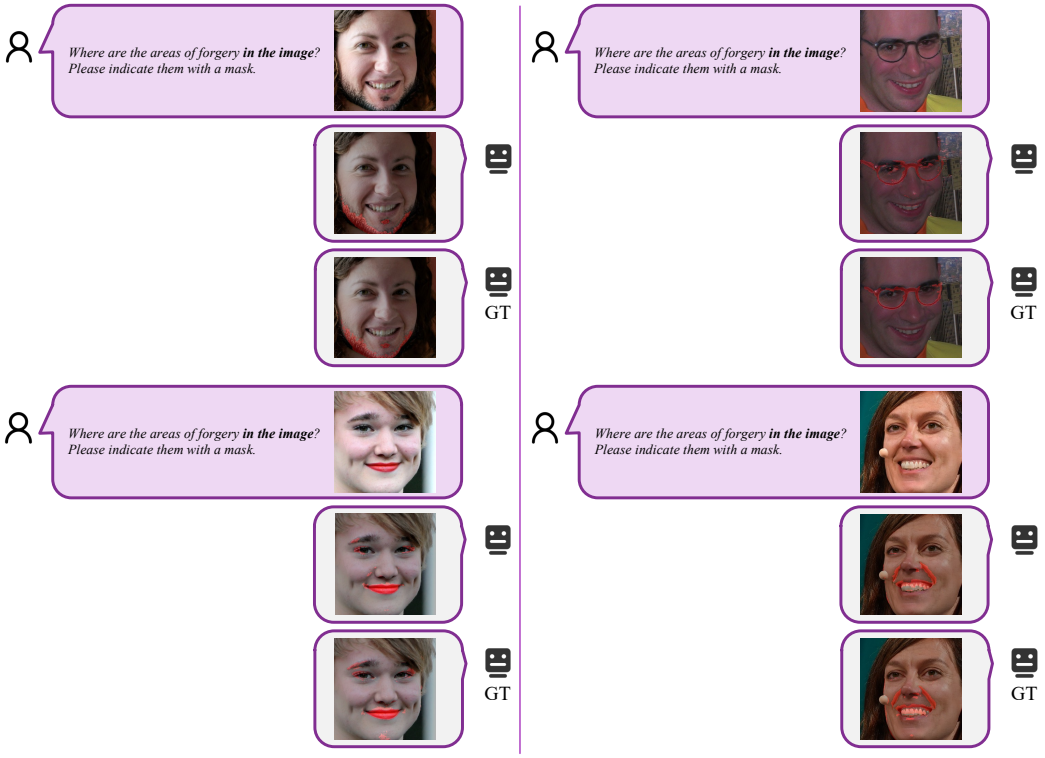

Figure 5: Output samples of our FiFa-VLM (multi-task setting) for I-Loc. We provide the Ground Truth (GT) for comparison.

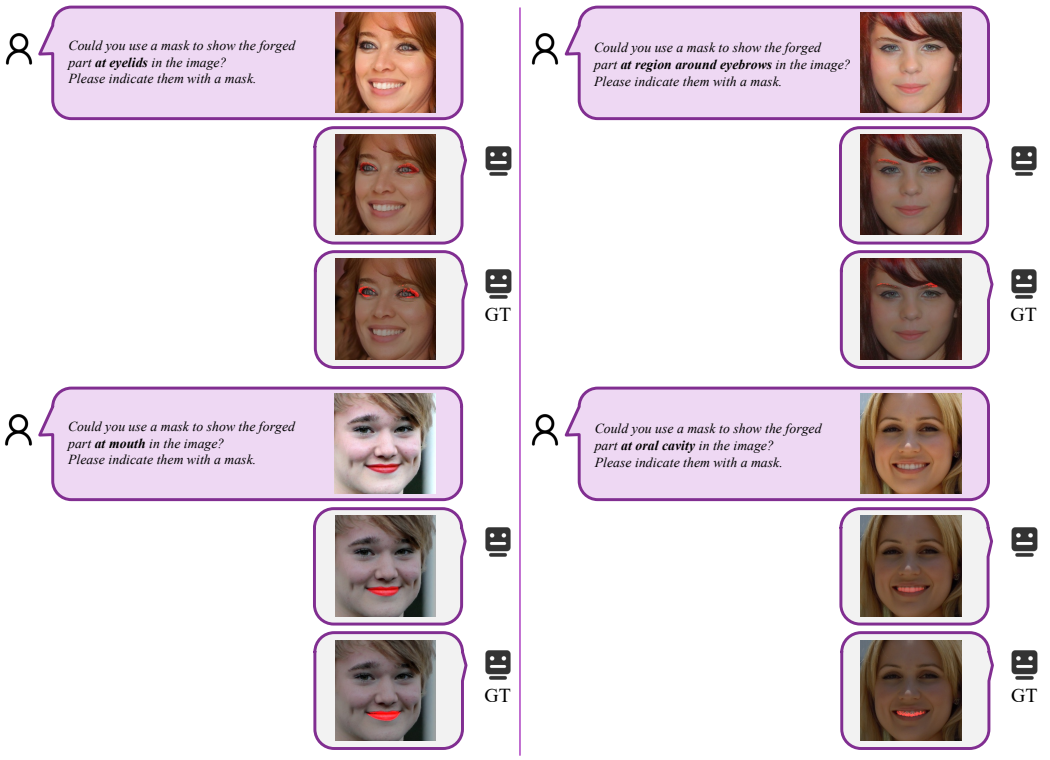

Figure 6: Output samples of our FiFa-VLM (multi-task setting) for R-Loc. We provide the Ground Truth (GT) for comparison.

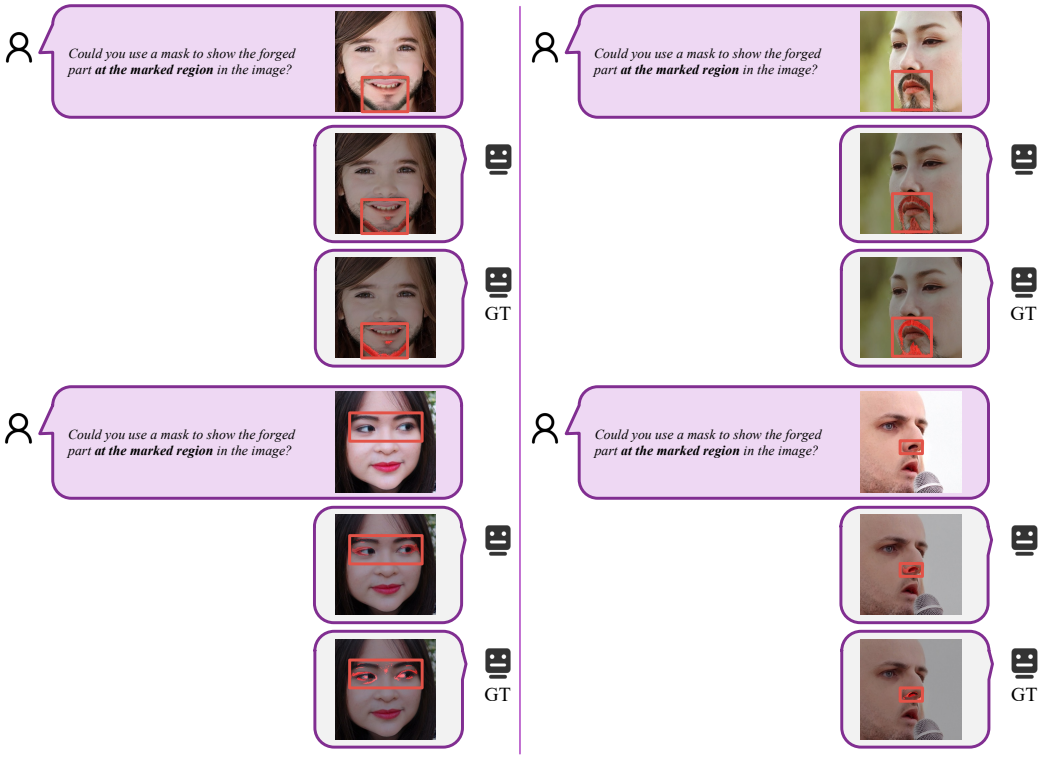

Figure 7: Output samples of our FiFa-VLM (multi-task setting) for B-Loc. We provide the Ground Truth (GT) for comparison.



Figure 8: Output samples of our FiFa-VLM (multi-task setting) for I-AGE. We provide the Ground Truth (GT) for comparison. In the textual explanations, the recalled Atomic Concepts are highlighted in **bold**.

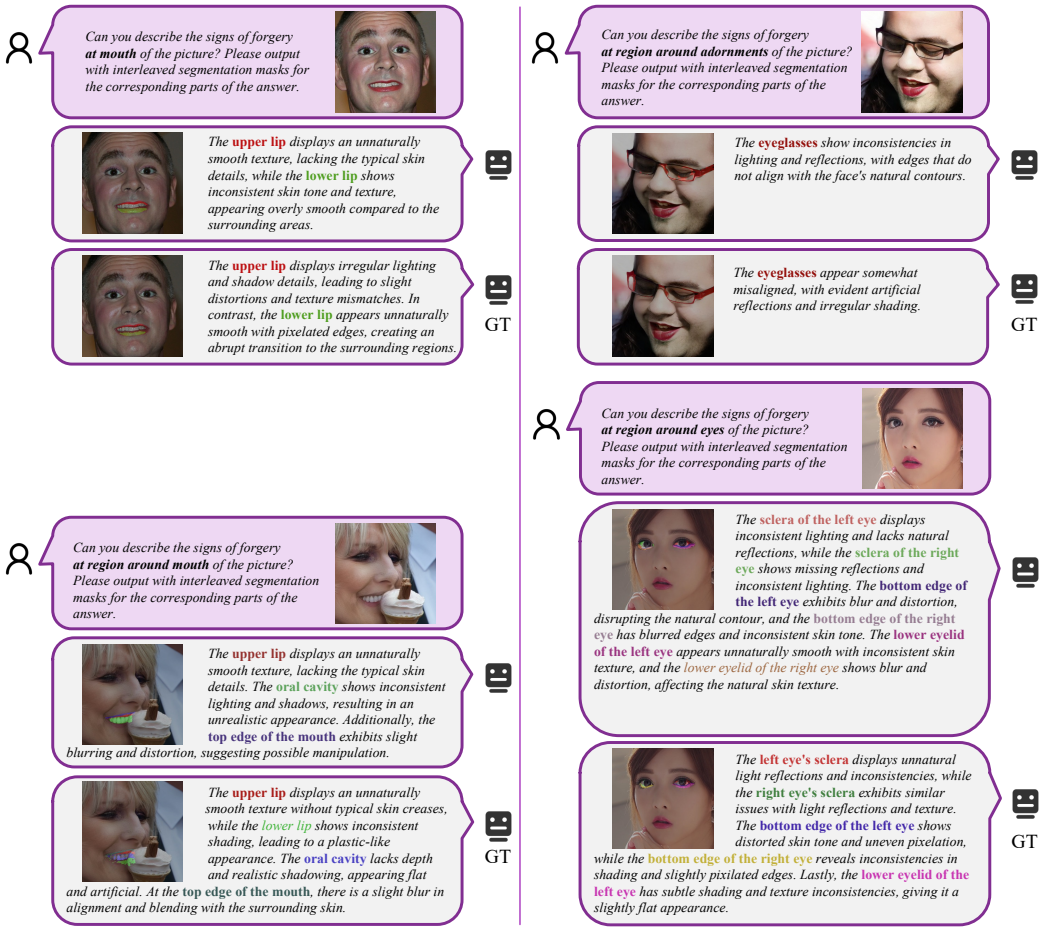

Figure 9: Output samples of our FiFa-VLM (multi-task setting) for R-AGE. We provide the Ground Truth (GT) for comparison. In the textual explanations, the recalled Atomic Concepts are highlighted in **bold**.

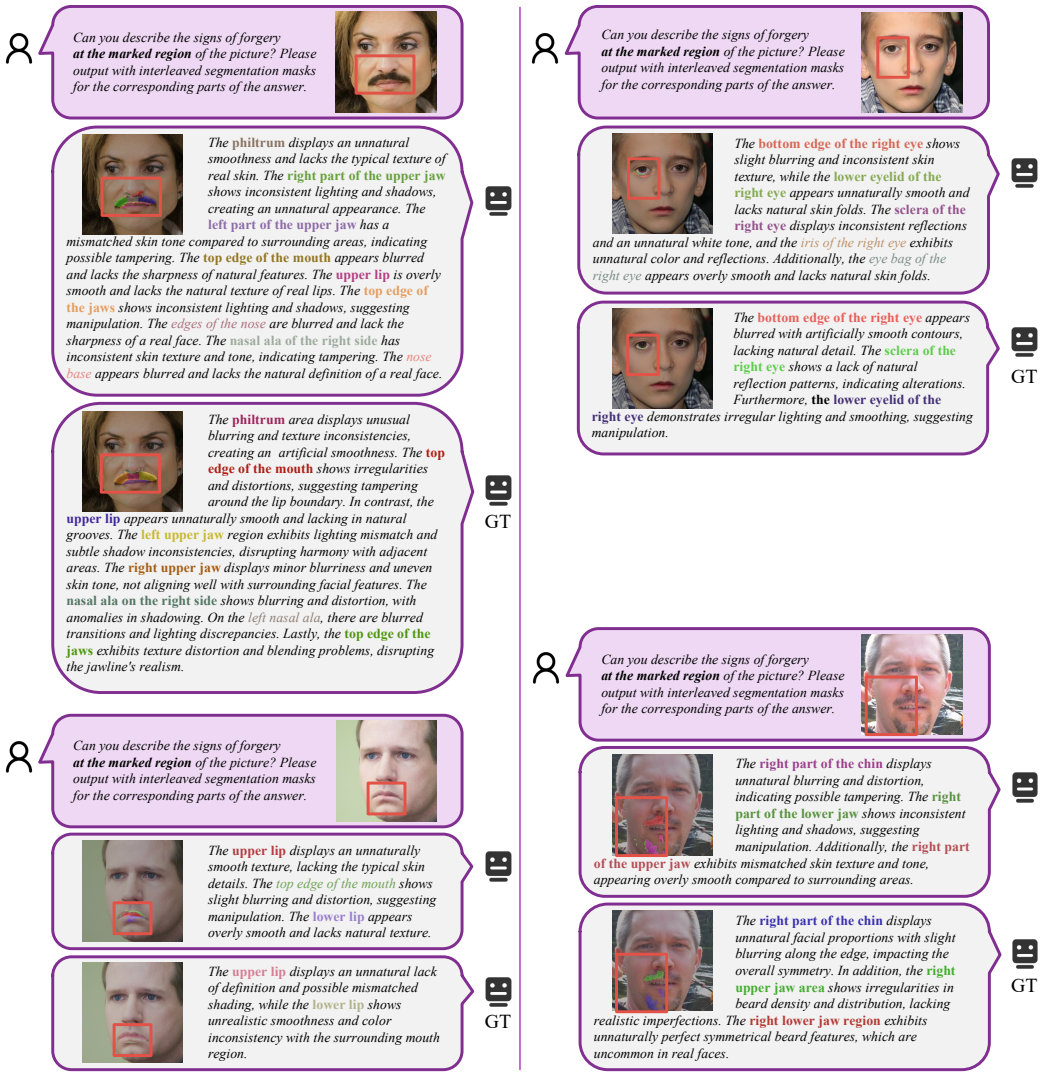

Figure 10: Output samples of our FiFa-VLM (multi-task setting) for B-AGE. We provide the Ground Truth (GT) for comparison. In the textual explanations, the recalled Atomic Concepts are highlighted in **bold**.