Towards General Visual-Linguistic Face Forgery Detection

Supplementary Material

Overview of Supplementary Materials

This supplementary material provides additional details and experimental results to support our main paper. It is organized as follows:

- Section A details the FFTG algorithm's forgery type decision criteria and procedures.
- Section B presents additional experimental results on cross-manipulation and multi-source evaluation.
- Section C describes the dataset details and training protocols.
- Section D provides comprehensive visualizations including attention maps, annotation comparisons, and LLaVA responses.
- Section E explains the prompt design and implementation details.

A. Details of FFTG

This section mainly introduces the details of the forgery type decision in the FFTG algorithm.

Color Difference. This phenomenon occurs in the face swap when the color of the source and target face has a drastic difference. Inspired by the color transfer [9], we leverage the distance of the average channel-wise mean and variance of the real and fake regions in the Lab color space to determine whether there exists a color difference. The Lab color space minimizes correlation between channels, which helps reduce the impact of changes in a certain channel on the overall color. The pseudocode is shown in Alg. 1, $split$ represents dividing the channel of the image, Lab denotes converting the RGB color space into Lab space.

Blur. There exists local blurring in forgery faces due to the instability of the generated model or blending operation. To quantify such phenomena, we make use of the Laplacian image, which can reflect the sharpness of image edges. Specifically, as shown in Alg. 2, we compute the variance of the real and fake images of the selected region after the Laplacian operator, and if the value of the real is larger than the fake one and their difference is greater than a certain threshold, we define this part as blurred. The $Laplacian(.)$ represents the Laplacian operator, $var(.)$ means calculating the variance of the input image.

Structure Abnormal. We observed that compared with normal faces, some organs of fake faces will be obviously deformed. To metric such structure deformable, we use the Structural Similarity (SSIM) index difference between real and fake images of the selected region R_s to decide whether the chosen region has a structure abnormal or not, which details in Alg. 3.

Algorithm 1 Color Difference Decision

Input: Real image selected region $R_s(i_r)$, fake image selected region $R_s(i_f)$, mean threshold θ_c^m, standard deviation thresholdθ_c^s

$R_s(i_r)^{'}, R_s(i_f)^{'} = Lab(R_s(i_r)), Lab(R_s(i_f))$
$L_r, a_r, b_r = split(R_s(i_r)^{'})$
$L_f, a_f, b_f = split(R_s(i_f)^{'})$
$L^m = ||mean(L_r) - mean(L_f)||_2$
$a^m = ||mean(a_r) - mean(a_f)||_2$
$b^m = ||mean(b_r) - mean(b_f)||_2$
$L^s = ||std(L_r) - std(L_f)||_2$
$a^s = ||std(a_r) - std(a_f)||_2$
$b^s = ||std(b_r) - std(b_f)||_2$
m = (L^m + a^m + b^m) / 3
s = (L^s + a^s + b^s) / 3
if m $> \theta_c^m$ and s $> \theta_c^s$ **then**
    Return True
else
    Return False
end if

Algorithm 2 Blur Decision

Input: Real image selected region $R_s(i_r)$, fake image selected region $R_s(i_f)$, variance threshold θ_b^v

$r\_var = var(Laplacian(R_s(i_r)))$
$f\_var = var(Laplacian(R_s(i_f)))$
if $r\_var > f\_var$ and $(r\_var - f\_var) > \theta_b^v$ **then**
    Return True
else
    Return False
end if

Texture Abnormal. It has been proved that the generator typically correlates the values of nearby pixels and cannot generate as strong texture contrast as real data [8], leading to texture differences in some forgery regions. Similar to the Gram-Net [8], we leverage a texture analysis tool–the contrast of Gray-Level Co-occurrence Matrix (GLCM) [4], formed as C_d. Larger C_d reflects stronger texture contrast, sharper and clearer visual effects. Inversely, a low value C_d means the texture is blurred and unclear. We define a forgery region as texture abnormal when the C_d of the real is larger than the fake one beyond the threshold. The algorithm is shown in Alg. 4, where $GLCM$ represents the average Gray-Level Co-occurrence Matrix of the input from right, down, left, and upper four orthogonal directions.

Blend Boundary. Existing face manipulation methods of-

Algorithm 3 Structure Abnormal Decision

Input: Real image selected region $R_s(i_r)$, fake image selected region $R_s(i_f)$, ssim threshold θ_s

$s = ssim(R_s(i_r), R_s(i_f))$
if $s < \theta_s$ **then**
    Return True
else
    Return False
end if

Algorithm 4 Texture Abnormal Decision

Input: Real image selected region $R_s(i_r)$, fake image selected region $R_s(i_f)$, contrast threshold θ_t

Init: $N = 256 \times 256$

$P_r = GLCM(R_s(i_r))$
$P_f = GLCM(R_s(i_f))$
$C_d^r = \frac{1}{N}\sum_{i=0}^{255}\sum_{j=0}^{255}|i-j|^2 P_r(i,j)$
$C_d^f = \frac{1}{N}\sum_{i=0}^{255}\sum_{j=0}^{255}|i-j|^2 P_f(i,j)$
if $C_f^r > C_d^f$ and $(C_d^r - C_d^f) > \theta_t$ **then**
    Return True
else
    Return False
end if

Algorithm 5 Blend Boundary Decision

Input: Image region I, mask M, threshold set $\theta_g, \theta_e, \theta_f$

// Get boundary regions
$I_{inner}, I_{outer} = GetBoundaryRegion(M)$
// Check gradient discontinuity
$g_x = Sobel(I, x)$, $g_y = Sobel(I, y)$
$g_{mag} = \sqrt{g_x^2 + g_y^2}$
$s_g = |mean(g_{mag}[I_{inner}]) - mean(g_{mag}[I_{outer}])|$
// Check edge artifacts
$E = Canny(I)$
$s_e = sum(E*(I_{inner}+I_{outer}))/sum(I_{inner}+I_{outer})$
// Check frequency patterns
$F = DCT(I * (I_{inner} + I_{outer}))$
$s_f = sum(|F_{high}|)/sum(|F_{low}|)$
// Count evidence
$evidence = 0$
if $s_g > \theta_g$ **then** $evidence += 1$
end if
if $s_e > \theta_e$ **then** $evidence += 1$
end if
if $s_f > \theta_f$ **then** $evidence += 1$
end if
return $evidence \geq 2$

ten leave intrinsic cues at the blending boundaries when merging manipulated faces with original backgrounds. As detailed in Alg. 5, we first extract inner (I_{inner}) and outer (I_{outer}) boundary regions around the manipulation mask to analyze the transition area where artifacts typically occur. We then analyze three key characteristics: gradient discontinuity assessed by comparing mean gradient magnitudes between inner and outer regions using Sobel operators to identify abrupt changes in intensity transitions, edge artifacts detected through Canny detection on the combined boundary region where manipulation often creates abnormal edge densities and patterns at the interface between real and fake regions, and frequency domain abnormalities examined by analyzing the ratio of high to low frequency components in the DCT transform of the boundary area, as blending operations typically introduce unnatural high-frequency patterns that differ from smooth transitions in natural images. By analyzing the combined boundary region rather than separate inner and outer regions for edge and frequency analysis, we can better capture the complete transition patterns and avoid missing artifacts that occur exactly at the boundary interface. The detection combines these multiple evidence sources to ensure reliability, requiring at least two metrics to exceed their thresholds before classifying a region as containing significant blending artifacts, thus reducing false positives while maintaining sensitivity to various types of blending anomalies.

B. Additional Experimental Results

B.1. Cross-manipulation evaluation

To further validate the generalization capability of our FFTG-enhanced CLIP model, we conduct cross-manipulation experiments using the high-quality version of FF++ dataset. We train our model on one manipulation method and evaluate it on all four methods (DeepFakes (DF), Face2Face (F2F), FaceSwap (FS), and NeuralTextures (NT)) to assess detection performance on unseen manipulation types. As shown in Table 1, we compare our approach with three recent state-of-the-art methods: Multi-attentional (MAT), GFF, and DCL. The diagonal values represent intra-domain performance, while off-diagonal values indicate cross-manipulation generalization. Our method demonstrates superior performance in most scenarios, particularly in challenging cross-manipulation cases. For instance, when training on FaceSwap and testing on DeepFakes, our method achieves 87.55% AUC, surpassing DCL by 13%. The improvements can be attributed to the high-quality text annotations generated by FFTG and our three-branch training framework, which help the model capture manipulation patterns that are common across different forgery types.

Train	Method	DF	F2F	FS	NT
DF	MAT	*99.92*	75.23	40.61	71.08
	GFF	*99.87*	76.89	47.21	72.88
	DCL	***99.98***	77.13	61.01	75.01
	Ours	*99.91*	**85.41**	**75.34**	**77.19**
F2F	MAT	86.15	*99.13*	60.14	64.59
	GFF	89.23	*99.10*	61.30	64.77
	DCL	91.91	*99.21*	59.58	66.67
	Ours	**92.32**	***99.35***	**62.19**	**67.81**
FS	MAT	64.13	66.39	*99.67*	50.10
	GFF	70.21	68.72	*99.85*	49.91
	DCL	74.80	69.75	***99.90***	52.60
	Ours	**87.55**	**79.13**	*99.27*	**53.53**
NT	MAT	87.23	48.22	75.33	*98.66*
	GFF	88.49	49.81	74.31	*98.77*
	DCL	91.23	52.13	79.31	*98.97*
	Ours	**93.10**	**61.55**	**83.27**	***98.98***

Table 1. Cross-manipulation evaluation in terms of AUC. Diagonal results indicate the intra-domain performance.

B.2. Multi-source manipulation evaluation.

We evaluate the model's generalization capability through multi-source manipulation experiments, where we train on three manipulation methods and test on the remaining unknown method. This challenging protocol assesses the model's ability to detect previously unseen manipulation types. The experiments are conducted on both high-quality (HQ) and low-quality (LQ) versions of FF++ dataset to comprehensively evaluate robustness across different image qualities. As shown in Table 2, our method consistently outperforms existing approaches across all settings. On high-quality DeepFakes (DF-HQ), our method achieves 95.07% accuracy, surpassing the previous state-of-the-art UIA-ViT by 4.67%. Similar improvements are observed for Face2Face (F2F) detection, where we achieve 88.12% accuracy on HQ data. Notably, the performance advantage is maintained in low-quality scenarios, where compression artifacts make forgery detection particularly challenging. For instance, on DF-LQ and F2F-LQ, our method achieves 86.17% and 71.25% accuracy respectively, significantly outperforming previous methods like DCL and EN-B4. These results demonstrate that our FFTG-enhanced approach not only excels at detecting high-quality forgeries but also maintains robust performance when dealing with compressed, low-quality images, suggesting effective learning of manipulation-specific features that persist across different image qualities.

Method	DF (HQ)	DF (LQ)	F2F (HQ)	F2F (LQ)
	ACC	ACC	ACC	ACC
EN-B4	82.40	67.60	63.32	61.41
Focalloss	81.33	67.47	60.80	61.00
Multi-task	70.30	66.76	58.74	56.50
MLDG	84.21	67.15	63.46	58.12
LTW	85.60	69.15	65.60	65.70
DCL	87.70	75.90	68.40	67.85
UIA-ViT	90.40	-	86.40	-
Ours	**95.07**	**86.17**	**88.12**	**71.25**

Table 2. Performance on multi-source manipulation evaluation, the protocols and the compaired results are from [10]. DF means traning on the other three manipulated methods of FFpp and test on deepfakes class. The same for the others.

B.3. Comparison with DFFD and Fakelocator

Our FFTG method demonstrates superior performance when compared to recent detection approaches DFFD [2] and Fakelocator [5]. We conducted comprehensive experiments on the FFpp dataset using ground-truth masks converted to text descriptions for evaluation. The results clearly show that FFTG achieves better localization performance across all metrics. Specifically, FFTG obtains 89.48% precision, 57.12% recall, and 64.96% F1 score, significantly outperforming both DFFD (which achieved 83.11% precision, 50.25% recall, and 58.93% F1 score) and Fakelocator (with 85.49% precision, 53.79% recall, and 60.63% F1 score). This substantial improvement demonstrates the effectiveness of our mask-guided annotation strategy in precisely identifying manipulated regions.

B.4. Scalability on latest AIGC methods

To thoroughly assess the scalability of our approach, we conducted extensive experiments on cutting-edge AIGC datasets including DiffSwap, SD-XL, and LDM from DiffusionFace [1]. The results demonstrate that our method consistently outperforms existing approaches across all three datasets. Specifically, on DiffSwap, our method achieves 95.21% accuracy, significantly surpassing UCF (85.17%) and CLIP (90.65%). For SD-XL, our approach reaches 96.77% accuracy compared to UCF's 83.77% and CLIP's 92.67%. Similarly, on LDM, we obtain 92.23% accuracy versus UCF's 83.92% and CLIP's 89.35%. These comprehensive results validate that our language supervision mechanism enables substantially better semantic generalization to the latest AIGC methods. Furthermore, when applying FFTG to annotate DiffSwap data, we achieved 90.12% region accuracy, which considerably outperforms GPT annotation (70.22%), thus demonstrating the strong generalization capabilities of our approach.

C. Dataset Details

C.1. Training and Test dataset.

To evaluate the generalization of our proposed annotation, we conduct our experiments on several challenging datasets: 1) FaceForensics++ [10]: a widely-used forgery dataset contains 1000 videos with four different manipulated approaches, including two deep learning based *DeepFakes* and *NeuralTextures* and two graphics-based methods *Face2Face* and *FaceSwap*. This dataset provides pairwise real and forgery data, enabling us to generate mixed forgery images with FFTG. 2) DFDC-P [3] dataset is a challenging dataset with 1133 real videos and 4080 fake videos, containing various manipulated methods and backgrounds. 3) DFD is a forgery dataset containing 363 real videos and 3068 fake videos, which is mostly generated by the Deepfake method. 4) Celeb-DF [7] is another high-quality Deepfake dataset that contains various scenarios. 5) Wild-Deepfake [12] is a forgery face dataset obtained from the internet, leading to a diversified distribution of scenarios. We use DSFD [6] to extract faces from each video.

C.2. Analysis of Text Annotations

To better understand the characteristics of FFTG annotations across different manipulation types, we visualize their word distributions through word clouds in Figure 1. In Deepfakes, the annotations concentrate on structural aspects, with "distortions" and "nose" being prominent, along with texture-related descriptions, reflecting the method's tendency to create geometric inconsistencies. For Face2Face, the word cloud reveals a focus on color inconsistencies and transitions, with terms like "lipcolor" and "particularly" frequently appearing, indicating the method's impact on local appearance details. In FaceSwap cases, FFTG identifies broader structural changes, with "facial" and "structure" being dominant terms, while also capturing clear signs of alterations in face contours. The NeuralTextures annotations emphasize blending-related artifacts, with "blending" and "surrounding" appearing prominently, along with specific attention to mouth regions and transitions. This visualization demonstrates FFTG's ability to generate precise, manipulation-specific annotations that capture the unique characteristics of each forgery type. The focused vocabulary and consistent emphasis on specific artifacts reflect the effectiveness of our mask-guided approach in identifying and describing relevant manipulation features.

D. Additional Visuallization

D.1. Visualizations on FFpp dataset.

To further validate the interpretability of our method, we visualized the attention heatmaps across different approaches on the test set of FFpp HQ dataset, comparing our method with a baseline (binary classification with CLIP pretrained image-encoder) and the state-of-the-art UIA-VIT [11]. The comparison in Figure 2 spans four manipulation methods: DeepFake, FaceSwap, Face2Face and NeuralTextures, with corresponding ground truth masks serving as references for manipulation regions. The baseline model shows diffused attention patterns that lack precise localization of manipulated regions. UIA-VIT demonstrates improved focus but still exhibits scattered attention that sometimes deviates from the actual manipulation areas. In contrast, our method achieves significantly more precise attention localization that closely aligns with the ground truth masks across all manipulation types. This is particularly evident in the NeuralTextures example, where our method accurately concentrates on the subtle mouth area manipulations while other methods show misplaced or dispersed attention. For Deepfake and FaceSwap cases, our attention maps precisely highlight the key manipulated facial regions, and in Face2Face examples, they effectively capture the structural modifications. This precise alignment between our attention maps and ground truth masks demonstrates that the fine-grained linguistic supervision from FFTG annotations effectively guides the model to focus on genuine manipulation artifacts, improving both detection accuracy and interpretability.

D.2. Visualizations on unseen dataset.

We visualize attention maps from different models on various unseen datasets (WildDeepfake, DFDC, and Celeb-DF) along with real faces in Figure 3. The baseline model's attention appears scattered and unfocused, with activation spread across irrelevant facial regions, indicating its limited ability to identify manipulation-specific features. UIA-ViT shows improved attention patterns with better concentration on facial components, but still exhibits some dispersion and occasionally highlights unmanipulated areas. In contrast, our method demonstrates more precise attention localization that aligns well with actual forgery regions. For instance, in WildDeepfake samples, our model precisely concentrates on the manipulated facial features while maintaining minimal activation on unmodified areas. On DFDC and Celeb-DF, it effectively captures the subtle manipulation artifacts despite their varying characteristics. When processing real faces, our model maintains clean and evenly distributed attention patterns without false activations. These visualizations confirm that our FFTG-guided approach helps the model learn more accurate and interpretable features for face forgery detection, enabling better generalization across different domains and manipulation types.

Figure 1. Word cloud comparison of FFTG annotations on FFpp dataset.

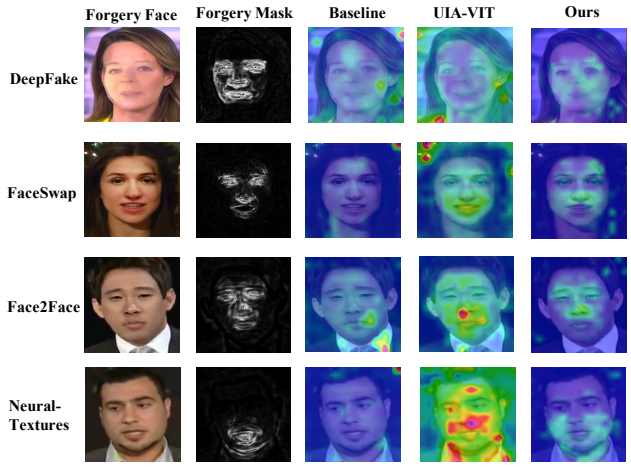


Figure 2. Visualization of attention heatmap on training dataset (FFpp) of the baseline, UIA-VIT, and our proposed method. The forgery Mask represents the ground truth manipulation mask generated by Eq. 1.

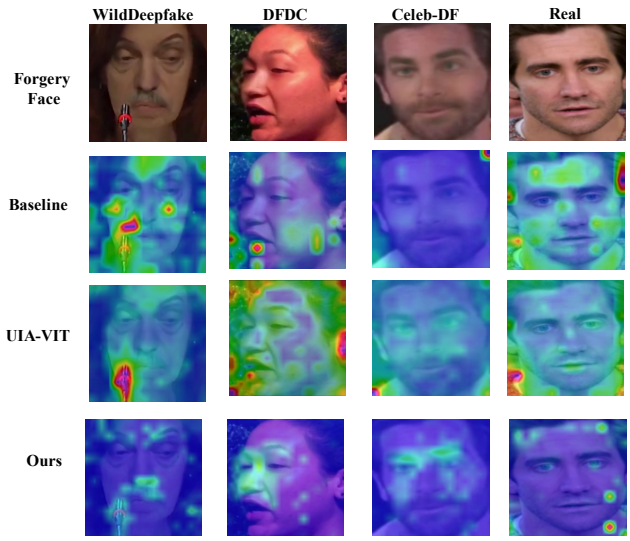


Figure 3. Attention heatmap visualization of the baseline, UIA-VIT, and our proposed method on the unseen dataset. The first row represents the original images that did not appear in the training set.

D.3. Visualizations of Annotation

To better understand the differences between annotation methods and demonstrate FFTG's advantages, we provide a detailed comparison of annotations generated by different approaches across four major manipulation types: Deepfakes, Face2Face, FaceSwap, and NeuralTextures. We present the manipulated image, forgery mask, real image, and corresponding annotations from human annotators, GPT-4o, DD-VQA, and our FFTG method, with key forgery-related terms highlighted in red to emphasize each

method's detection focus.

As shown in Figure 4, the Deepfake example reveals distinct differences in annotation approaches. Human annotations focus primarily on obvious visual cues like facial symmetry and cheek irregularities, but also incorrectly identify nose distortions. GPT-4o's description tends toward general stylistic observations about computer generation and animation-like qualities, lacking specific artifact identification. DD-VQA provides more structured observations about the eyes and mouth regions, correctly identifying texture patterns and blending artifacts, though still missing some key details. Our FFTG's raw annotation demonstrates superior accuracy by precisely identifying the manipulated regions indicated by the forgery mask. It correctly pinpoints unusual texture patterns in the eyes and highlights blending artifacts around the eyes and mouth, while also detecting color distribution inconsistencies. This mask-guided approach helps avoid the hallucination of non-existent artifacts and ensures descriptions align with actual manipulation evidence.

For Face2Face manipulation (Figure 5), the human annotation correctly identifies the unnatural contouring and lighting around the face, particularly noting mouth region abnormalities. GPT-4o mentions various facial features including eyebrows and skin texture, but seems scattered in its focus. DD-VQA provides a more concise description focusing specifically on the structural distortion and blending artifacts in the mouth region. Our FFTG raw annotation shows the highest precision by accurately identifying structural distortions in the mouth area and highlighting specific artifacts like color inconsistencies and blending anomalies at region boundaries, which aligns well with the forgery mask's indication.

In the FaceSwap example (Figure 6), human annotation identifies unnatural brightness in the eyes and mouth distortions, along with skin smoothing effects. GPT-4o's description is notably limited, only mentioning curved nose and eyebrow asymmetry. DD-VQA provides more comprehensive detection, identifying structural distortions across eyes, nose, and mouth regions, with proper attention to blending artifacts. FFTG's raw annotation demonstrates superior precision by accurately capturing both the structural distortions and texture abnormalities in the eyes and nose regions, while also detailing the blending artifacts around the mouth, closely matching the forgery mask's indications.

In the NeuralTextures example (Figure 7), human annotation focuses on skin texture and asymmetry issues, particularly noting abnormalities in the mouth and lipstick regions. GPT-4o provides minimal observation, only mentioning eye and nose irregularities without specific details. DD-VQA maintains a focused description of the mouth region's structural distortions and blending artifacts. FFTG's raw annotation demonstrates the most precise detection by identifying specific texture abnormalities in the mouth region and structural distortions in the lip area, matching the forgery mask's indication of manipulation. The annotation particularly emphasizes unnatural texture patterns and deviations from natural curves, providing detailed evidence of manipulation.

Across all four manipulation types, FFTG consistently demonstrates superior accuracy in identifying and describing forgery artifacts, with its annotations closely aligning with the ground truth masks while providing detailed, artifact-specific descriptions that avoid hallucination.

D.4. Visualizations of LLaVA Responses

We demonstrate the effectiveness of FFTG annotations in improving multimodal language models' forgery detection capabilities through both quantitative evaluation and qualitative analysis. As shown in Table 1, our FFTG-enhanced LLaVA achieves superior performance across all metrics, with 95.84% accuracy on FFpp and 75.00% on the challenging Celeb-DF dataset, significantly outperforming models trained with DD-VQA annotations. More importantly, our model demonstrates higher precision (88.07%) and recall (55.30%) in identifying manipulation regions, indicating more accurate and reliable detection capabilities.

This quantitative improvement is further illustrated through example dialogues in Figure 8. When presented with a challenging fake image, DD-VQA-trained LLaVA relies heavily on general stylistic observations about computer generation and animation-like qualities, focusing on superficial features like eye asymmetry and nose curvature. In contrast, our FFTG-trained LLaVA provides more precise and artifact-focused analysis, accurately identifying specific texture patterns in the mouth region and structural distortions that deviate from natural appearances. More importantly, when analyzing real images, while DD-VQA-trained LLaVA exhibits bias toward forgery detection with false positives, our model demonstrates better discrimination ability by correctly identifying authentic images and providing detailed natural features as supporting evidence. These qualitative examples, supported by the strong numerical results, demonstrate that FFTG's precise annotation guidance helps LLaVA develop more reliable and interpretable forgery detection capabilities.

E. Prompt Details

E.1. Connectives of Raw Annotation

To enhance the naturalness and readability of raw annotations, we design specific connective phrases for each forgery type, as shown in Figure 9. These connectives are used in conjunction with a region token (e.g., eyes, nose, mouth) to form complete, natural descriptions. For example, when blur is detected in the eye region, the annotation

would read "the eyes appears blurry compared to natural faces". For blending artifacts, the base connective "shows blending artifacts characterized" is further enhanced with specific evidence phrases based on our detection metrics: "sharp changes in image gradients at the boundaries" when gradient discontinuity is detected, "unnatural edge patterns" for edge artifacts, and "unusual frequency patterns at the boundaries" for frequency domain abnormalities. These detailed characterizations help specify the exact nature of the blending artifacts detected. This structured approach helps guide GPT in generating more accurate and contextually appropriate refined annotations while maintaining consistent terminology across different forgery types.

E.2. Annotation Refinement Prompt

To guide GPT in generating accurate and natural language annotations, we design four complementary prompts as shown in Figure 10. The *Visual Prompt* pairs fake and real images to enable direct visual comparison, helping GPT identify manipulation artifacts through contrast. For each case, we provide dynamically generated raw annotations that combine detected regions with corresponding connective phrases as initial guidance. The *Guide Prompt* explains the FFTG detection process, including mask generation, region analysis, and specific criteria for detecting texture abnormalities, structural deformations, color inconsistencies, and blending artifacts, helping GPT understand the technical basis. The *Task Description Prompt* establishes the expert analysis context and provides step-by-step instructions for comparing images and generating comprehensive descriptions. Finally, the *Pre-defined Prompt* specifies the required JSON output format and key requirements to ensure consistent and focused annotations. This multi-faceted prompting strategy enables GPT to generate detailed, accurate descriptions while maintaining natural language expression and avoiding hallucination.

References

[1] Zhongxi Chen, Ke Sun, Ziyin Zhou, Xianming Lin, Xiaoshuai Sun, Liujuan Cao, and Rongrong Ji. Diffusionface: Towards a comprehensive dataset for diffusion-based face forgery analysis. *arXiv preprint arXiv:2403.18471*, 2024. 3

[2] Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil K Jain. On the detection of digital face manipulation. In *CVPR*, pages 5781–5790, 2020. 3

[3] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge dataset. *arXiv preprint arXiv:2006.07397*, 2020. 4

[4] Robert M Haralick, Karthikeyan Shanmugam, and Its' Hak Dinstein. Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, (6):610–621, 1973. 1

[5] Yihao Huang, Felix Juefei-Xu, Qing Guo, Yang Liu, and Geguang Pu. Fakelocator: Robust localization of gan-based face manipulations. *IEEE Transactions on Information Forensics and Security*, 17:2657–2672, 2022. 3

[6] Jian Li, Yabiao Wang, Changan Wang, Ying Tai, Jianjun Qian, Jian Yang, Chengjie Wang, Jilin Li, and Feiyue Huang. Dsfd: dual shot face detector. In *CVPR*, pages 5060–5069, 2019. 4

[7] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A new dataset for deepfake forensics. *arXiv preprint arXiv:1909.12962*, 2019. 4

[8] Zhengzhe Liu, Xiaojuan Qi, and Philip HS Torr. Global texture enhancement for fake face detection in the wild. In *CVPR*, pages 8060–8069, 2020. 1

[9] Erik Reinhard, Michael Adhikhmin, Bruce Gooch, and Peter Shirley. Color transfer between images. *IEEE Computer graphics and applications*, 21(5):34–41, 2001. 1

[10] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *ICCV*, pages 1–11, 2019. 3, 4

[11] Wanyi Zhuang, Qi Chu, Zhentao Tan, Qiankun Liu, Haojie Yuan, Changtao Miao, Zixiang Luo, and Nenghai Yu. Uiavit: Unsupervised inconsistency-aware method based on vision transformer for face forgery detection. *ECCV*, 2022. 4

[12] Bojia Zi, Minghao Chang, Jingjing Chen, Xingjun Ma, and Yu-Gang Jiang. Wilddeepfake: A challenging real-world dataset for deepfake detection. In *ACM MM*, pages 2382–2390, 2020. 4

This is a fake face. Additionally, the facial symmetry seems exaggerated, particularly around the cheeks, leading to an unnatural overall look. These discrepancies indicate potential forgery in the eye and nose regions. The subject's nose appears slightly distorted or over-processed, losing the realistic shadows and contours evident in the real face. The overall skin texture is overly uniform, giving it a plastic-like appearance.

GPT-4o Annotation

The image looks fake. It looks like a computer-generated artist styled image for animation, carton or films. The person‘s skin looks a bit fake because the skin is overly smooth and lacks of details, such as skin texture. It looks like a computer-generated artist styled image for animation, carton or films. The person's skin looks a bit fake because the skin is overly smooth and lacks of details, such as skin texture there are stains/flaws on the person's face there are areas looking unnaturally brighter there area of looking unnaturally darker the person has mismatched bangs. The person's eyebrows look a bit fake because the eyebrows look asymmetric.

DD-VQA Annotation

This image appears to be manipulated. The eyes region shows the eyes exhibits unusual texture patterns, and the eyes appears structurally distorted, and the eyes shows blending artifacts characterized by sharp changes in image gradients at the boundaries, unnatural edge patterns. The mouth region shows the mouth exhibits unusual texture patterns, and the mouth appears structurally distorted, and the mouth shows blending artifacts by sharp changes in image gradients at the boundaries, unnatural edge patterns. The face region shows unnatural color variations.

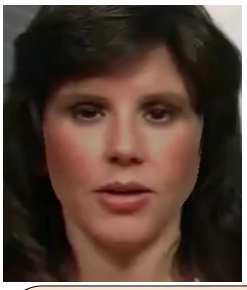
Raw Annotation

This is a fake face. The eyes in this image exhibit unusual texture patterns, with a lack of fine detail that indicates manipulation, along with noticeable structural distortions in their shape. There are blending artifacts around the eyes and mouth, characterized by sharp transitions at the boundaries, causing unnatural edge patterns and disruptions in continuity. The overall color distribution appears inconsistent, particularly in the LAB color space, suggesting unnatural adjustments that detract from a realistic appearance.

FFTG Annotation

Figure 4. Comparison of different annotation approaches on a Deepfake example.

Face2Face Image | **Mask Image** | **Real Image**



This is a fake face. The image shows noticeable inconsistencies in the contours of the face. The eyes appear unnaturally dark and lack detail, creating a lifeless effect. The mouth is also affected, showing a lack of natural contouring that is present in the real image.

GPT-4o Annotation

The image looks fake. It looks like an ai-generated image although with human appearance but with unrealistic texture or lighting. The person‘s eyebrows look very fake because the eyebrows do not match the face’s structure. Right broken eyebrows. The person‘s eyes look very fake because the eyes looks blurry and lack of details. The person has mismatched bangs. It is an image with manipulated face regions. The person's nose looks a bit fake because the nose is placed at the wrong place on the face. The person's skin looks a bit fake because there are boundaries on the person's face the person has mismatched bangs.

DD-VQA Annotation

This image appears to be manipulated. The mouth region shows the mouth appears structurally distorted, and the mouth shows blending artifacts characterized by sharp changes in image gradients at the boundaries, unnatural edge patterns.

Raw Annotation

This is a fake face. The mouth region exhibits significant structural distortion, with unnatural curves that diverge from typical human anatomy, contributing to an overall sense of artificiality. Blending artifacts are evident around the mouth, where sharp transitions in gradients disrupt the seamlessness usually seen in real images. Color inconsistencies are noticeable, particularly on the cheeks and lips, where the distribution of tones does not match typical human variations. The mouth shows pronounced edge distortions, reinforcing signs of manipulation that detracts from the overall impression of authenticity.

FFTG Annotation

Figure 5. Comparison of different annotation approaches on a Face2Face example.

FaceSwap Image | **Mask Image** | **Real Image**

This is a fake face. The eyes are unnaturally bright and glossy, with one iris appearing larger than the other, leading to an unbalanced expression. The mouth appears slightly distorted, missing natural shadows and highlights that would ordinarily define the lips. There is an evident smoothing effect across the skin, making it look overly uniform, unlike the textured skin in the real image. The shadows around the nose and mouth are inconsistent, creating an artificial appearance.

GPT-4o Annotation

The image looks fake. The person's nose looks a bit fake because the nose lookes curved. The person has mismatched bangs. The person's eyebrows look a bit fake because the eyebrows look asymmetric.

DD-VQA Annotation

This image appears to be manipulated. The eyes region shows the eyes appears structurally distorted. The nose region shows the nose exhibits unusual texture patterns, and the nose appears structurally distorted. The mouth region shows the mouth exhibits unusual texture patterns, and the mouth appears structurally distorted, and the mouth shows blending artifacts characterized by sharp changes in image gradients at the boundaries, unnatural edge patterns.

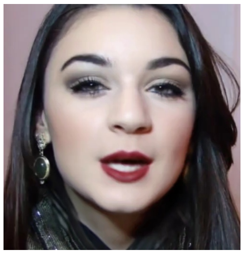
Raw Annotation

This is a fake face. The eyes appear structurally distorted, with misalignment and unusual reflections that deviate from natural human features. Textural abnormalities are evident, particularly around the iris, indicating manipulation. The nose presents unnatural texture patterns and structural deformations, creating an unrealistic appearance. This is compounded by discrepancies in shading, leading to color inconsistencies that detract from realism. The mouth exhibits blending artifacts, characterized by sharp transitions and unnatural edge patterns, suggesting digital alterations.

FFTG Annotation

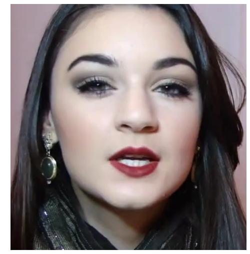
Figure 6. Comparison of different annotation approaches on a FaceSwap example.

NeuralTextures



Mask Image

Real Image



This is a fake face. The image shows a woman whose skin appears overly smooth and lacks natural texture, indicating potential digital manipulation. The left side of her mouth seems less defined, giving an unnatural asymmetry compared to the real image. Additionally, the lipstick shade appears more uniform, lacking the natural variation seen in the real image. Her eyes exhibit slight discrepancies in brightness, suggesting inconsistent reflections, especially in the right pupil.

GPT-4o Annotation

The image looks fake. The person's eyes look a bit fake because the eyes appear overly large to the face the person's nose looks a bit fake because the nose do not align properly with other facial features. The person has mismatched bangs.

DD-VQA Annotation

This image appears to be manipulated. The mouth region shows the mouth appears structurally distorted, and the mouth shows blending artifacts characterized by sharp changes in image gradients at the boundaries, unnatural edge patterns.

Raw Annotation

This is a fake face. The mouth region exhibits noticeable texture abnormalities, appearing overly smooth and lacking the natural variation typically present in human skin. The lips also show structural distortion, creating an unnatural pout-like appearance, diverging from the lip alignment in the real image. The lips are shaped in a way that seems artificial, with clear signs of distortion, deviating from the natural curves observed in the genuine image.

FFTG Annotation

Figure 7. Comparison of different annotation approaches on a NeuralTextures example.

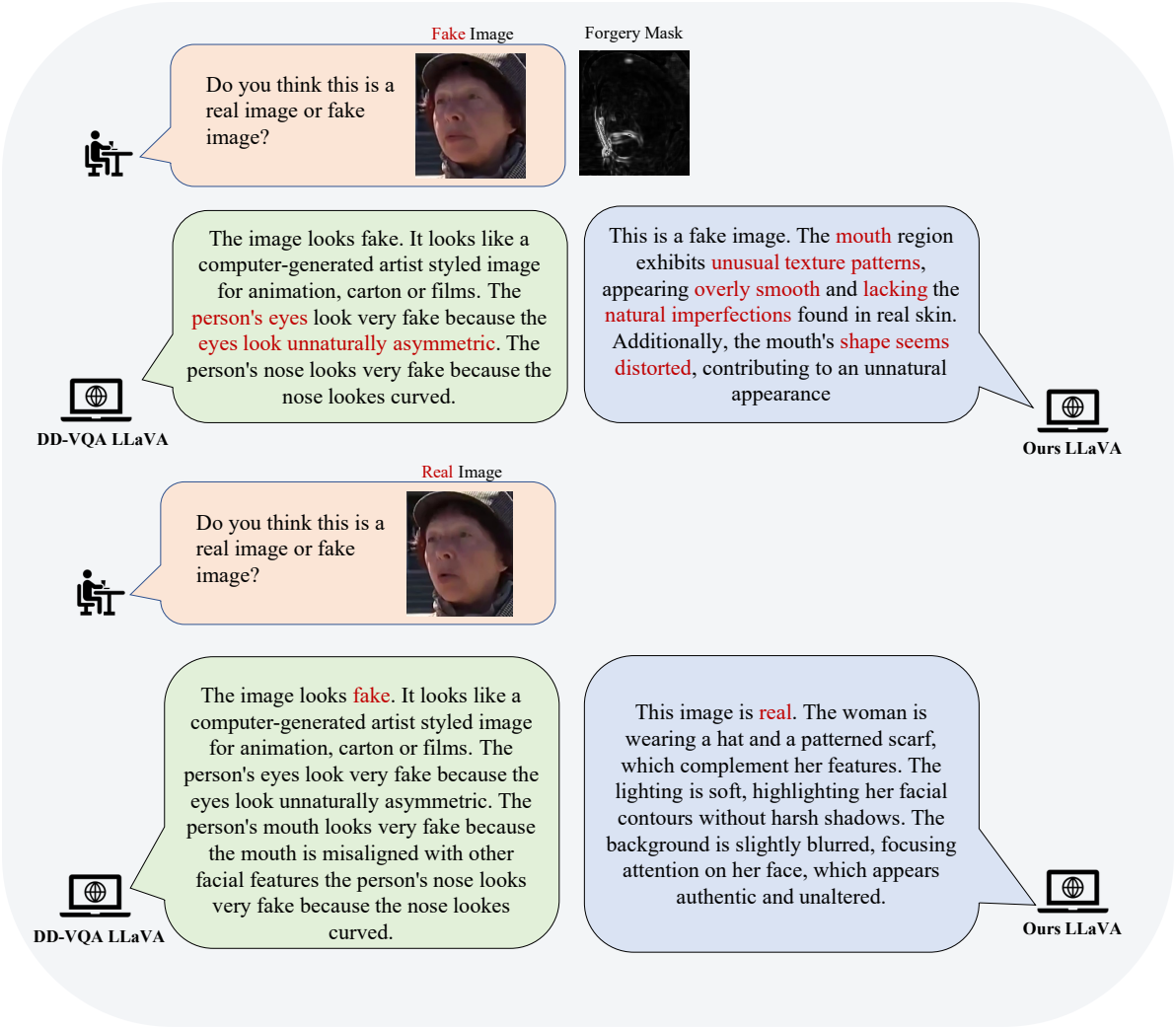


Figure 8. Interactive dialogue comparison between DD-VQA-trained and FFTG-trained LLaVA models on both fake (top) and real (bottom) images.

Forgery Type	Connectives
Blur	appears blurry compared to natural faces
Color Difference	shows unnatural color variations
Texture Abnormal	exhibits unusual texture patterns
Structure Abnormal	appears structurally distorted
Blend Boundary	shows blending artifacts characterized…

Figure 9. Connective phrases used for different forgery types in raw annotation generation. Each phrase starts with a specific region token region (e.g., eyes, nose, mouth) followed by these connectives to form natural descriptions of detected artifacts.

Visual Prompt

Fake Image Real Image

Raw Annotation

This image appears to be manipulated. The mouth region shows the mouth appears structurally distorted, and the mouth shows blending artifacts characterized by sharp changes in image gradients at the boundaries, unnatural edge patterns.

Guide Prompt

The FFTG Detection result is computed by:
1. First computing the difference mask between real and fake images to locate manipulated regions
2. Using facial landmarks to analyze specific regions (eyes, nose, mouth, whole face)
3. Detecting various artifacts in each region:
- Texture abnormality: Using GLCM (Gray-Level Co-occurrence Matrix) to detect unnatural texture patterns
- Structural deformation: Using SSIM (Structural Similarity) to detect shape distortions
- Color inconsistency: Analyzing color distributions in LAB color space
- Blending artifacts: Detecting unnatural transitions at region boundaries using gradient and frequency analysis

Task Description Prompt

You are an expert in analyzing facial images and detecting manipulations. Here are the images and initial FFTG analysis:
[Fake Image][Real Image]
PFIG Detection: {Raw Annotation}
Your task:
1. Analyze and compare the real and fake images.
2. For each forgery type described in the prompt, identify whether it is evident in the fake image.
3. Generate three different detailed captions for both the real and fake images, focusing on providing diverse but accurate descriptions of the visual discrepancies.
4. Make sure the captions are fine-grained, providing as much detail about the visual discrepancies as possible.

Pre-defined Prompt

The expected output should be in JSON format, structured as follows:
{{"real_image_captions": [
"Detailed description of the real image (caption 1).",
"Detailed description of the real image (caption 2).",
],"fake_image_captions": [
"Detailed description of the fake image (caption 1), highlighting the differences compared to the real image, with particular emphasis on the forgery regions and forgery types.",
"Detailed description of the fake image (caption 2), highlighting the differences compared to the real image, with particular emphasis on the forgery regions and forgery types.",
],}}
Important requirements:
1. EVERY caption MUST start with either "This is a real face." or "This is a fake face." exactly as shown
2. Provide exactly two distinct captions for each fake image and two captions for each real image
3. Use FFTG findings as reference but add your own observations
4. Be specific about visual details but avoid repetition across captions
5. Maintain objective, evidence-based descriptions
6. Focus on visible differences and anomalies
7. Try to show the counterfeit areas and types you think are counterfeit

Figure 10. Overview of FFTG prompting strategy for annotation refinement, consisting of Visual Prompt with paired images, Raw Annotation with dynamic descriptions, Guide Prompt explaining detection process, Task Description Prompt for analysis guidance, and Pre-defined Prompt for output format.