AuthGuard: Generalizable Deepfake Detection via Language Guidance

Guangyu Shen[1]*, Zhihua Li[2], Xiang Xu[2], Tianchen Zhao[2], Zheng Zhang[2], Dongsheng An[2],
Zhuowen Tu[2], Yifan Xing[2], Qin Zhang[2]
[1]Purdue University [2]AWS DS3

Abstract

Existing deepfake detection techniques struggle to keep-up with the ever-evolving novel, unseen forgeries methods. This limitation stems from their reliance on statistical artifacts learned during training, which are often tied to specific generation processes that may not be representative of samples from new, unseen deepfake generation methods encountered at test time. We propose that incorporating language guidance can improve deepfake detection generalization by integrating human-like commonsense reasoning – such as recognizing logical inconsistencies and perceptual anomalies – alongside statistical cues. To achieve this, we train an expert deepfake vision encoder by combining discriminative classification with image-text contrastive learning, where the text is generated by generalist MLLMs using few-shot prompting. This allows the encoder to extract both language-describable, commonsense deepfake artifacts and statistical forgery artifacts from pixel-level distributions. To further enhance robustness, we integrate data uncertainty learning into vision-language contrastive learning, mitigating noise in image-text supervision. Our expert vision encoder seamlessly interfaces with an LLM, further enabling more generalized and interpretable deepfake detection while also boosting accuracy. The resulting framework, ***AuthGuard****, achieves state-of-the-art deepfake detection accuracy in both in-distribution and out-of-distribution settings, achieving AUC gains of 6.15% on the DFDC dataset and 16.68% on the DF40 dataset. Additionally,* ***AuthGuard*** *significantly enhances deepfake reasoning, improving performance by 24.69% on the DDVQA dataset.*

1. Introduction

Recent advances in generative AI [13, 16, 42, 44] have now enabled the creation of hyper-realistic facial manipulations, increasingly blurring the line between real and synthetic imagery. This raises significant risks, including misinformation, identity fraud, and the erosion of public trust in visual evidence. Imagine a breaking news featuring a world leader announcing a sudden policy change, sparking panic in financial markets – only to later be revealed as a deepfake fabricated to manipulate public opinion. Or picture a virtual job interview with a recruiter, but the person behind the screen is a deepfake avatar orchestrating a phishing scam. As digital media increasingly shapes both public discourse and personal interactions, the ability to detect and analyze deepfake images is becoming more and more essential for safeguarding authenticity [51].

However, existing deepfake detection methods – such as [23, 27, 32, 47, 53, 55, 61] – struggle to keep pace with the rapid advancements in generative AI. Most approaches focus either on enhancing the vision encoders [47, 53, 55]

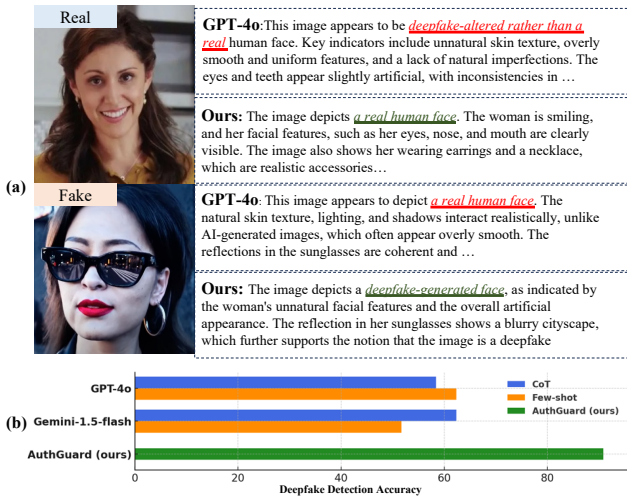


Figure 1. (a) Commercial models perform significantly worse than AuthGuard. For instance, GPT-4o frequently misclassifies low-resolution faces as fake, highlighting its limitations in distinguishing real from fake within general-purpose MLLMs. This underscores the need for a more accurate and lightweight deepfake-specific detection and reasoning model. (b) We leverage the VIC [31] to generate reasoning prompts for GPT-4o and Gemini-1.5 in real/fake classification. Few-shot prompting includes four labeled images (two real, two fake) for context. AuthGuard significantly outperforms commercial models on the DD-VQA [60] dataset.

*Work done during internship at AWS DS3.

or employing data-driven classification models trained on exemplar attack samples [9, 35, 52]. While these methods are effective against known manipulations, they often fail catastrophically when confronted with novel, unseen deepfakes, leaving critical vulnerabilities in real-world applications as new forgery methods rapidly emerge. This limitation arises because these methods rely heavily on recognizing ***statistical deepfake artifacts*** tied to specific generation processes – patterns that exist within their labeled training data [2, 34, 39] – which may not appear in the test data. In contrast, humans approach deepfake detection fundamentally differently: instead of relying on statistical cues, they apply commonsense reasoning and describe inconsistencies using natural language [14]. As also discussed in [14, 20], many cross-domain deepfake detection errors align with language describable patterns, indicating that existing methods overlook such key semantic insights.

To enhance generalization in deepfake detection, we propose **AuthGuard**, a unified deepfake detection and reasoning framework that captures both statistical deepfake artifacts and commonsense deepfake artifacts – human-interpretable, language-describable features that are independent of specific generative models. To achieve this, we develop an automatic data generation pipeline, leveraging state-of-the-art public Multimodal Large Language Models (MLLMs). By incorporating real or fake labels as contextual prompts, we generate 114k high-quality image-text pairs, with each text explaining why a face image appears fake or real. Using this dataset, we train an expert deepfake vision encoder by image-text combining contrastive learning with standard binary classification. This allows the trained vision encoder to capture broad semantic relationships for better generalization while enabling precise distinctions between real and deepfake images. Meanwhile, we mitigate text noise through probabilistic embedding [46], ensuring more robust cross-modal feature alignment. To effectively balance the contributions from statistical and commonsense deepfake artifacts, we introduce an adaptive aggregation mechanism which uses a light-weight adaptor network to generate input-dependent probabilistic weights, dynamically combining these artifacts into a single visual representation. The aggregated representation enhances generalization and addresses challenges in adapting to novel attacks. Finally, we integrate the aggregated visual representation into the LLaVA architecture [29] through instruction tuning on our curated dataset, creating a flexible framework where the LLM functions as a plug-and-play component for deepfake reasoning and explainability.

In summary, we make the following contributions:

1. We propose **AuthGuard**, a unified vision-language model that seamlessly integrates deepfake detection, reasoning, and analysis with enhanced generalization. To the best of our knowledge, we are the first work to achieve such unification in the deepfake domain.
2. We develop a simple yet effective training strategy for learning an expert deepfake vision encoder that captures both commonsense and statistical deepfake artifacts. This strategy employs a text-regularized representation learning method that uses MLLM-generated text data to reduce overfitting to statistical patterns in the training data, and incorporates a small adaptor to combine contrastive and discriminative features dynamically.
3. Through thorough benchmarking, we demonstrate that AuthGuard outperforms state-of-the-art methods in accuracy, generalization and reasoning. Specifically, AuthGuard achieves a 6.15% improvement in deepfake detection accuracy on the DFDC dataset [12] in the cross-dataset setting and a 24.36% increase in reasoning accuracy on the DDVQA dataset [60].

2. Related Work

Deepfake Detection As deepfake generation technology progresses, datasets now encompass a wider variety of attack types to evaluate the accuracy, robustness, and generalization of detection methods [12, 24, 43, 56, 59]. Traditionally, deepfake detectors have relied on binary image classifiers [9, 23, 61]. Recent advancements aim to enhance both accuracy and generalizability by optimizing various components of these detectors [17, 26, 35, 48, 55]. One approach, as seen in [47, 61], enhances the training data by generating synthetic samples that combine source and target images, thus enriching the dataset and potentially improving detector robustness. Another line of work aims to improve representation learning by adopting more powerful backbones and introducing novel training mechanisms. For instance, [9] introduces additional blending mask learning in the training and [53] decomposes images to reveal common forgery features, enabling the detector to learn more generalized characteristics of deepfakes. Similarly, [55] demonstrates that using representations from a wider variety of forgeries helps create more generalizable decision boundaries, reducing overfitting to method-specific features.

Adapting Vision-Language Models for Deepfake Detection Recent works [7, 15, 60] have explored using instruction tuning of vision-language models [22, 37] to perform deepfake detection by framing it as a visual question answering (VQA) task. However, existing works focus either on improving reasoning benchmarks or on enhancing binary detection accuracy, without providing an integrated model that addresses both. One approach [7] tunes soft prompts to classify images as real or fake, but it only provides a binary yes/no answer, missing the opportunity for more informative, explainable outputs Zhang *et al*. [60] annotated the FaceForensics++ dataset [43] with explanations for fake images and fine-tuned a BLIP model [22] to generate reasoning about why an image appears fake. However, their

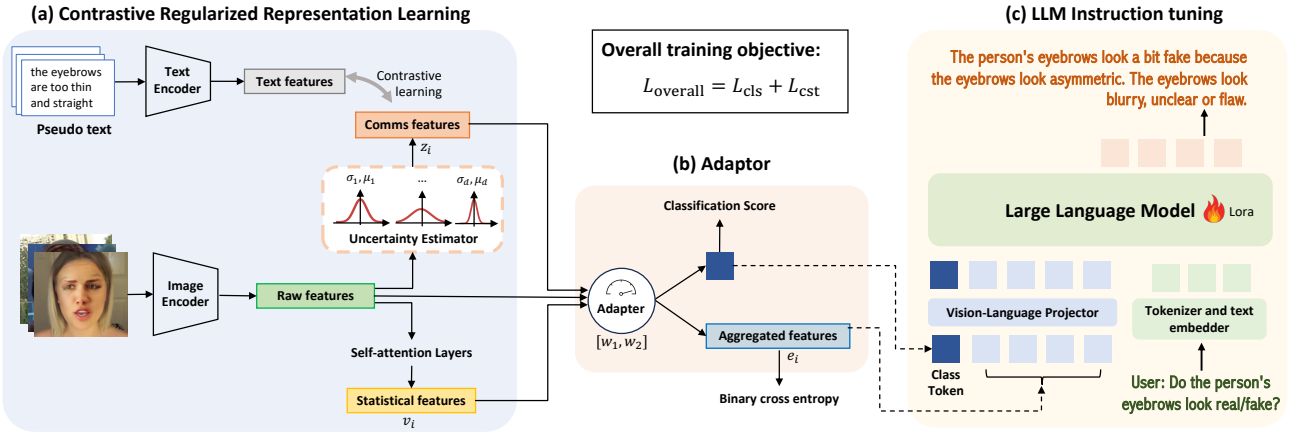


Figure 2. The framework of AuthGuard comprises two main components: (a) and (b) for expert vision representation learning, and (c) for LLM-based deepfake reasoning, where the LLM utilizes image tokens generated from the vision module. In the representation learning module, deepfake artifact learning is divided into commonsense and statistical artifacts. Commonsense artifacts are learned through vision contrastive learning and refined with probabilistic embedding to mitigate label noise. These are then combined with data-driven, forgery-specific artifacts via an adaptive router to extract expert deepfake features. Finally, the adapter's output class token is concatenated with patch-wise tokens from deeper layers and fed into the LLM for reasoning.

approach overlooks the critical role of statistical artifacts in deepfake detection, limiting its robustness when handling images with subtle discrepancies that cannot be captured by natural language.

3. Method

Fig. 2 illustrates the overall framework for **AuthGuard**. We first develop an expert deepfake encoder that captures both statistical and commonsense deepfake features by combining classification with vision-language contrastive learning. Specifically, commonsense deepfake features refer to language-describable and semantically meaningful artifacts learned through pseudo-text pairs, while statistical deepfake features include artifacts that are often imperceptible to humans, such as GAN fingerprints. An adaptor combines commonsense visual features and statistical deepfake features, creating a balanced representation for more generalized deepfake detection. After the deepfake vision encoder is trained, we integrate it with an LLM, enabling plug-and-play deepfake reasoning and interpretability. In the following sections, we describe our pseudo-text generation process in Sec. 3.1, detail the design and training of the vision encoder in Sec. 3.2 and Sec. 3.3, and discuss its integration with an LLM in Sec. 3.4.

3.1. Automatic Image-Text Pair Data Generation

Given the time-consuming nature of human annotation and the large size of typical deepfake datasets, we use a publicly available MLLM[1] to scale the generation of pseudo-text pairs for all training images. We design a customized prompt that incorporates image labels as contextual information to prompt MLLM to generate the data: *Explain why the face attributes (e.g., eyes, mouth, chin, hair, nose, and others) make this image look* ***Type***. Here ***Type*** denotes the binary (real/fake) image ground truth label. This label-based prompting method helps reducing hallucinations, ensuring that fake images consistently receive negative descriptions. The generated paragraph is then split into individual sentences based on keywords like "mouth", "eyes", and other facial landmarks. This added context helps the model generate more relevant and accurate pseudo-text pairs. Examples of the generated data are shown in Fig. 3.

To enhance multimodal alignment and deepfake reasoning in LLMs, we further generate instruction tuning data from image captions to teach the model to analyze visual inputs step by step, enabling reasoning for detecting deepfake artifacts [28, 29, 49]. It also helps better alignment of image and language representations, ensuring structured and logical responses. We then use a LLM to generate diverse instructions and responses from image descriptions, creating a large dataset of image-centric, conversational training data for deepfake reasoning. We provide detailed examples of the generated instruction tuning data in Fig. 4.

3.2. Aligning Deepfake Detection with Language

We design a specialized loss function to train our expert vision encoder. To enhance generalization, we incorporate image-text contrastive regularization to capture commonsense deepfake artifacts—features that can be described in natural language—alongside standard classification loss,

[1] In this paper, we use Llama 3.2 [33], but any publicly available generalist MLLM could be used in principle.

➢ Prompt: Explain why the face attributes make this image **look fake**.
Llama 3.2: The face appears to be a composite of multiple faces, *with the eyes and mouth being from one person and the hair from another*. The skin tone is also unnatural, with a *bluish tint* that is not typical of human skin. Additionally, the overall shape and proportions of the face are not consistent with those of a real person. Overall, the face in the image appears to be a manipulated or digitally altered representation of a person's face, rather than a genuine photograph.

Prompt: Is this image real or fake?
Llama 3.2: The image appears to be a screenshot of a news broadcast, with the woman's face possibly being displayed on a television screen, *making it a real image*.

Figure 3. Automatic pseudo-text generation: comparing Llama 3.2's responses to prompts with and without labels.

which focuses on learning statistical deepfake artifacts. The overall loss function can be expressed as follows:

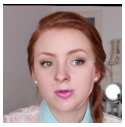
$$\mathcal{L}_{\text{overall}} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{cst}} \tag{1}$$

where L_{cls} is the the regular binary cross entropy loss and L_{cst} is the contrastive loss with details covered in Sec. 3.3.

A straightforward way to implementing this loss is to add a small adaptor that transforms the original vision embeddings for computing the classification loss while using the original vision embeddings for computing the contrastive loss. However, our experiments in Tab. 4 reveal that this design results in suboptimal performance. This occurs because the model overfits to easily learned statistical patterns from the classification loss, which hinders its ability to effectively capture the contrastive features. To address this, we propose an adaptive weighting mechanism that dynamically adjusts the contribution of each artifact type. For each image x_i, we branch its raw embedding from the vision encoder, denoted as h_i, into two sub-modules: one is used for contrastive learning to capture language-describable deepfake features, while the other passes through two self-attention layers to generate a transformed embedding that captures statistical deepfake features. We further introduce a trainable adapter, R^t, that dynamically aggregates the original and transformed embeddings for each image, similar to the mixture-of-experts routing mechanism [58]. For each image, denoting the contrastive feature as z_i and the statistical feature as v_i, the aggregated embedding is computed as $e_i = w_1 v_i + w_2 z_i$ where $\mathbf{w}_i = [w_1, w_2]$ represents the gating weights assigned by R^t to control the contribution of each representation. The gating weights are computed as $\mathbf{w}_i = \text{Softmax}\big((R^t(v_i))\big)$ where R^t projects v_i to a 1-D vector. The elements of $\mathbf{w}_i$ sum to 1, representing the relative importance of each deepfake artifact. Finally, we apply a binary cross-entropy loss $\mathcal{L}_{\text{cls}}$ on the e_i and the contrastive loss $\mathcal{L}_{\text{cst}}$ on z_i. As shown in our ablation in Tab. 4, this design yields 0.91% AUC improvement on DFDC.

3.3. Uncertainty Control for Text Noise

For contrastive learning, directly using the noisy pseudo-labels generated by a MLLM can introduce hallucinations and degrade detection performance. To tackle this, we draw



Question: "Is the image in question likely to be a real, unedited photograph or a digitally altered one?"
Answer: "Based on the description, it is likely that the image has been digitally altered or enhanced. The characteristics mentioned, such as the unrealistic smoothness and uniformity of the skin texture, overly perfect lighting, and …"

Question: "What specific features of the image suggest that it has been edited?"
Answer: "The features that suggest the image has been edited include the unnatural smoothness and uniformity of the skin, the overly perfect lighting without …"

Question: "Can you elaborate on why the smoothness and uniformity of the skin suggest editing?"
Answer: "In real, unedited photographs, skin texture often appears more natural and varied, with subtle imperfections such as pores, fine lines, and slight unevenness in tone. The smooth and uniform skin texture in the image, lacking …"

Question: "How does the lighting in the image contribute to the suspicion of editing?"
Answer: "The lighting in the image overly perfect, which implies that it lacks the natural imperfections and variations that are typically present in real-world lighting conditions. In most photographs, lighting can create subtle shadows, highlights …"

Question: "What role does symmetry play in suggesting that the image has been edited?"
Answer: "Symmetry in an image can be a natural occurrence, but when it is overly pronounced or perfect, it can be an indicator of digital editing …"

Figure 4. Example of instruction-tuning data generated from image captions (Sec. 3.1).

inspiration from data uncertainty learning in face recognition [6] and incorporate uncertainty control into contrastive learning [41]. Specifically, instead of using deterministic image embeddings, we adopt probabilistic embeddings [46], where each embedding is modeled as a distribution parameterized by a mean and variance. This allows for adaptive alignment and effectively mitigates the adverse effects of noisy pairs. To account for inconsistencies in image-text pairs, we model the raw vision embedding h_i as a Gaussian distribution conditioned on the text description T_i:

$$p(h_i|T_i) = \mathcal{N}(h_i; \mu_i, \sigma_i^2 I), \tag{2}$$

where the mean and variance are predicted in an input-dependent manner by the backbone using a two-layer self-attention mechanism: $\mu_i = f_\mu(h_i)$, $\sigma_i = f_\sigma(h_i)$, where $f_{(\cdot)}$ denotes the network parameters. Thus, the representation of each sample is no longer a deterministic embedding but a distribution. To allow the model to take gradients as usual, we apply the reparameterization trick [19]. Specifically, we first sample random Gaussian noise $\epsilon \sim \mathcal{N}(0, 1)$, and then generate the equivalent sampling representation as $z_i = \mu_i + \sigma_i \cdot \epsilon$. Finally, a contrastive loss is applied to the resampled visual embeddings z_i and textual embeddings $t_i = G(T_i)$ in a batch manner, where G is the text encoder. This can be formally expressed as:

$$\mathcal{L}_{\text{cst}} = -\frac{1}{2B}\sum_{i=1}^{B}\left[\log\frac{e^{w\cdot\tilde{z}_i\cdot\tilde{t}_i}}{\sum_{k=1}^{B} e^{w\cdot\tilde{z}_i\cdot\tilde{t}_k}} + \log\frac{e^{w\cdot\tilde{t}_i\cdot\tilde{z}_i}}{\sum_{k=1}^{B} e^{w\cdot\tilde{t}_i\cdot\tilde{z}_k}}\right] \tag{3}$$

In the equation, $\tilde{z}_i$, $\tilde{t}_i$ represent the normalized versions of z_i and t_i, and ω is the temperature parameter.

3.4. Deepfake Reasoning with a LLM

To integrate deepfake detection with reasoning, we adopt a LLaVA-like [29] architecture for its superior accuracy in tasks like interpreting facial expressions and identifying object properties. We replace the original LLaVA vision encoder with our specialist vision encoder. Also, unlike the original implementation, which projects only image patch tokens to the LLM, we incorporate both a refined, discriminative class embedding and the original patch-wise tokens. This approach mitigates hallucinations and reduces inconsistencies between VQA detection and VQA reasoning, as the LLM is modulated with label information indicating whether the input image is real or fake. Formally, let f_{proj} denote this projection function:

$$e'_i = f_{\text{proj}}(v_i, e_i), \tag{4}$$

where $e'_i \in \mathbb{R}^{d_l}$ and d_l is the dimension of the language model's input. The projection f_{proj} aligns e'_i with the language model's token space. Once aligned, these visual tokens, along with accompanying text tokens, are processed by a LLM (i.e., Vicuna [11]) to perform reasoning and generate responses based on both visual and textual inputs.

To train both the projection layer and the LLM, we employ instruction tuning using the auto-regressive loss. Given a triplet (x, q, y) consisting of an image x, question q, and response y with L tokens, the model factorizes the probability of generating the sequence using the chain rule:

$$p(y \mid x, q) = \prod_{i=1}^{L} p_\theta(y_i \mid x, q, y_{<i}), \tag{5}$$

Here, $\theta = \{W, \phi\}$ represents the parameters of the projector and the LLM. The vision encoder weights are kept frozen.

4. Experiments

4.1. Experimental Settings

Datasets We evaluate the effectiveness of our proposed framework in terms of generalizability and interpretability across multiple deepfake datasets:

1. **FF++** dataset [43] consists of 1,000 real and 4,000 fake videos from various sources. Four deepfake techniques are employed to generate the corresponding fake videos. We train all models over FF++ dataset. For evaluation, we adopt the c23 version of FF++.
2. **DFDC** dataset [12] is a more challenging and larger deepfake detection dataset. Consistent with existing literature, we train the detector on the FF++ dataset and evaluate its performance on the DFDC dataset.
3. **DF40** [56] is a recent dataset with 40 distinct forgery methods. We select 8 unseen faceswapping methods generated from the FF++ real samples for evaluation.
4. **DD-VQA** dataset [56] is a recently established deepfake visual question answering dataset. It includes 14,782 question-answer pairs, with the images collected from the FF++ and human annotated text.

Baselines For the binary deepfake classification task, we compare our proposed framework with 5 state-of-the-art binary deepfake detectors, including UCF [53], SRM [32], Face-X-Ray [23], SPSL [27], and LSDA [55]. We use the implementation and pre-trained weights from the third-party evaluation toolbox, DeepfakeBench [54]. For the deepfake reasoning task, we compare our method with BLIP-TI [60], a state-of-the-art multi-modal VQA model specialized in deepfake detection and reasoning tasks.

Metrics For the binary deepfake detection task, following exsisting worsk [53], we report the AUC score on the FF++, DFDC, and DF40 datasets and compare them with state-of-the-art methods. To assess our model's reasoning capability against baselines on DD-VQA, we follow the methodology outlined in [60]. This includes using detection accuracy and four natural language processing metrics: BLUE-4[38], CIDEr [50], ROUGE_L [25], and METEOR [4].

Implementation Details Our training process consists of two stages. In the first stage, we train an expert deepfake vision encoder using ViT-L/14 [41] for the vision branch and RoBERTa [30] for the text branch. Both ViT-L/14 and RoBERTa are pretrained on the LAION dataset [45]. During fine-tuning, the text encoder remains frozen, and only the vision encoder's weights are updated. The learning rate is set to 5e-6, with 1,000 warmup steps. We employ the Adam optimizer with a cosine learning rate scheduler and train for 5 epochs. The α and β are set to 0.05 and 1 respectively. In the second stage, we perform instruction tuning for deepfake reasoning.

4.2. Evaluation Results

Evaluation on Seen Attacks We perform in-distribution evaluation with seen attacks using the FF++ dataset [43] and demonstrate that our vision encoder training by combining commonsense and statistical artifacts would not sacrifice the performance in this setting. As shown in Tab. 1, when evaluating the in-distribution performance, our approach excels with an AUC of 98.87% on the FF++ dataset [43], outperforming the best baseline UCF (98.12%) [53] by a margin of 0.76%. When analyzing the performance on the individual subsets within FF++, our method exhibits comparable results to UCF on the FF-DF, FF-F2F, and FF-FS subsets. However, our approach significantly outperforms UCF on the FF-NT subset, achieving an AUC of 98.13%, which exceeds UCF's 95.27% by a substantial 3.00%.

Evaluation on Unseen Attacks We compare our method with the recent state-of-the-arts on out-of-distribution test using unseen attacks from the DFDC dataset [12] in Tab. 1 and selected unseen face-swapping attacks from DF40 [56]

Method	Venues	In-distribution test (AUC (%)↑)					Out-distribution test (AUC (%)↑)	Out-distribution test (ACC (%)↑)
		FF-DF	FF-F2F	FF-FS	FF-NT	FF++	DFDC	DFDC
UCF [53]	ICCV 2023	99.05	**99.01**	**99.18**	95.27	98.12	73.15	65.75
SRM [32]	CVPR 2021	97.82	97.08	97.17	94.01	96.52	68.44	62.83
Face-X-Ray [23]	CVPR 2020	97.94	98.72	98.71	92.90	95.92	63.26	-
SPSL [27]	CVPR 2021	97.81	97.54	98.29	92.99	96.10	70.40	62.35
LSDA [55]	CVPR 2024	96.94	96.43	95.11	94.92	95.38	73.60	60.73
FreqDebias [18]	CVPR 2025	-	-	-	-	97.50	74.10	-
AuthGuard (ours)	-	**99.65**	98.83	98.85	**98.13**	**98.87**	**78.13**	**71.93**

Table 1. Comparison on AUC (%) is performed using a standard evaluation setup, where the models are trained on the FF++ dataset [43] and tested for in-distribution performance on FF++ and out-of-distribution performance on DFDC [12].

Method	Venues	Out-distribution test (AUC (%) ↑)							Out-distribution test (ACC (%) ↑)
		UniFace	E4S	FaceDancer	FS-GAN	InSwap	SimSwap	**Average**	**Average**
RECCE [5]	CVPR 2022	84.25	65.20	78.32	88.45	79.51	73.04	78.13	69.40
CORE [36]	CVPR 2022	81.69	63.39	71.69	91.06	79.37	69.34	76.09	68.74
SRM [32]	CVPR 2021	78.24	66.73	77.43	84.52	76.15	65.96	74.84	68.11
UCF [53]	ICCV 2023	78.67	69.17	80.06	88.09	76.85	64.92	76.29	68.26
LSDA [55]	CVPR 2024	84.25	65.19	78.32	88.45	79.37	69.34	77.49	65.37
AuthGuard (ours)	-	**95.54**	**89.65**	**85.56**	**95.37**	**95.06**	**85.78**	**91.16**	**83.81**

Table 2. Accuracy comparison for several recent methods on six representative face-swapping techniques from the DF40 dataset [56]. Our method outperforms the recent approaches by a substantial margin, leading to an 16.68% improvement in performance.

Figure 5. To qualitatively demonstrate the alignment between the vision representations from our expert vision encoder and the paired text, we visualize attention maps from our vision-language contrastive learning using the approach from [8]. For the same image, different descriptions of commonsense artifacts activate the corresponding facial regions.

in Tab. 2. In the out-of-distribution setting, our method achieves an AUC of 78.13% on the DFDC dataset [12] and 93.20% on the DF40 dataset [56]. These results represent significant improvements of 6.15% and 16.68%, respectively, compared to the best corresponding baselines [53]. Based on the superior performance over state-of-the-art methods depicted in Tab. 1 and Tab. 2, we can conclude that our method surpasses existing methods not only in the in-distribution setting but also in the out-of-distribution setting. This validates the effectiveness of combining statistical and commonsense deepfake features, enhancing generalization and enabling robust detection of both seen and unseen deepfakes.

Improving Explanation and Reasoning In addition of detection accuracy, we evaluate deepfake explanation and reasoning capabilities of our method on the DD-VQA dataset [60], comparing it with general MLLMs and the expert MLLM, BLIP-TI [60], as shown in Tab. 3. Firstly, we benchmark the performance of GPT-4 [3], LLaVA-1.5-7B [29], InternVL [10], Phi-3.5-Vision [1], InternVL2-8B [10], and LLaVA-OB-7B [21] for the deepfake detection task without any fine-tuning on the DD-VQA dataset. We prompt the MLLMs with "*Is this image real or fake?*" and evaluate accuracy by comparing the model's output (real/fake) with the ground truth labels. By this approach, these MLLMs obtain accuracy of 75.00%, 61.83%, 61.07%, 61.07%, and 70.23%, respectively. Additionally, we provide the detection and reasoning performance of the standard LLaVA model, trained on general vision data. While better than random guessing, its accuracy remains suboptimal, even for the supervised fine-tuned LLaVA on the same instruction tuning dataset. These limitations in standard MLLMs highlight the need to enhance the vision encoder with deepfake-specific knowledge. On the DD-VQA dataset, BLIP-TI [60] is the only deepfake domain expert model providing both capabilities on binary detection (yes/no) and reason explanation. Compared to BLIP-TI [60], our method can produce probabilistic output and achieve a detection accuracy of 90.84%, outperforming BLIP-TI by 3.83% on deepfake detection accuracy. This improvement in detection accuracy validates the effectiveness of our approach in accurately identifying deepfake artifacts within the DD-VQA dataset [60]. In reasoning, our method significantly outperforms BLIP-TI across all four

Method	Deepfake Detection		Visual Question Answering (VQA)				
	AUC (%)↑	Accuracy (%)↑	BLEU-4 ↑	CIDEr ↑	ROUGH_L ↑	METEOR ↑	Average ↑
General MLLMs (GPT-4o and [1, 3, 10, 21])	-	58.43-75.00	-	-	-	-	-
BLIP-TI [60]	-	87.49	0.4075	2.0567	0.6085	0.3463	0.9823
LLaVA (off-the-shelf) [29]	-	61.83	0.1422	1.4231	0.2314	0.2489	0.5114
LLaVA (fine-tuned)	-	70.60	0.3880	1.7320	0.5470	0.3212	0.7470
AuthGuard (ours)	**96.81**	**90.84**	**0.4980**	**3.3050**	**0.6950**	0.4010	**1.2248**

Table 3. Comparison of various methods on the DD-VQA dataset [60] for both deepfake detection and reasoning tasks. Our method achieved a 3.83% improvement on deepfake detection and a 24.69% improvement on the average quality metric on deepfake VQA.

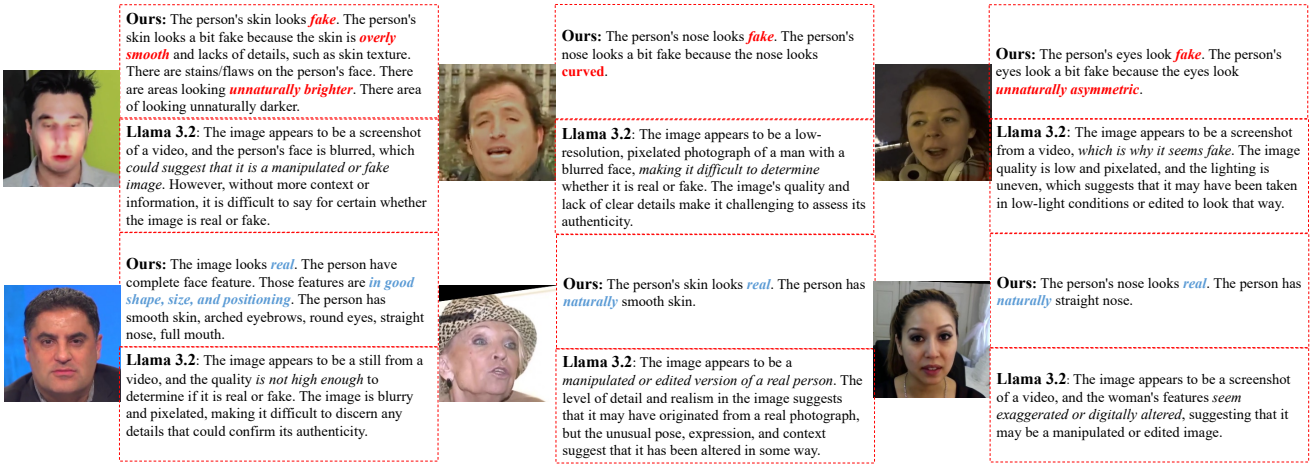


Figure 6. Reasoning examples of our method on DD-VQA. Highlighted text indicates the accurate descriptions of facial features.

Modules			Datasets	
Semantic Artifacts	Uncertainty Est.	Adapter	FF++	DFDC
			98.35	75.22
✓			98.69	76.29
✓	✓		98.62	77.20
✓	✓	✓	**98.87**	**78.13**

Table 4. Ablation study on AUC (%), evaluating the improvement of the proposed modules on the FF++ and DFDC datasets.

metrics and 24.69% improvement on average. Specifically, we achieve scores of 0.4980 for BLEU-4, 3.3050 for CIDEr, 0.6950 for ROUGE_L, and 0.4010 for METEOR, while BLIP-TI scores 0.4075, 2.0567, 0.6085, and 0.3463, respectively. These higher scores indicate that the responses generated by our method are better aligned with human annotators' assessments, and that the binary token produced by our method shows greater consistency with the labels.

4.3. Ablation Study and Visualization

Impact of Various Image Encoder Improvements To evaluate the impact of each proposed module, we performed ablation studies by incrementally adding each component to the image encoder training process depicted in Tab. 4. First, incorporating semantic learning through image-text contrastive learning led to an AUC increase from 98.35% to 98.69% on FF++ and from 75.22% to 76.29% on DFDC, suggesting that semantic alignment enhances performance even in the presence of noisy labels. Next, we introduced an uncertainty estimation (Uncertainty Est.) module to address the unreliability of pseudo-text annotations, further improving the AUC from 76.29% to 77.20%. This result indicates that probabilistic embeddings can strengthen the effectiveness of semantic learning by providing adaptive vision-language alignment. Finally, with the integration of adaptive balancing between commonsense and forgery-specific features, the AUC rose from 75.22% to 78.13%, achieving the highest scores across all configurations. These findings demonstrate that both commonsense and statistical artifacts are crucial for deepfake detection, and adaptively combining these artifacts offers an optimal solution.

Interpretation and Reasoning For reasoning training, we incrementally adapt the off-the-shelf LLaVA model to our specific requirements, assessing its performance on the DD-VQA dataset [60]. Tab. 3 reveal that with the default fine-tuning approach where the CLIP [41] vision encoder is fixed and only the vision-language projector and LLM are tuned, the model achieves a detection accuracy of just 70.60% and

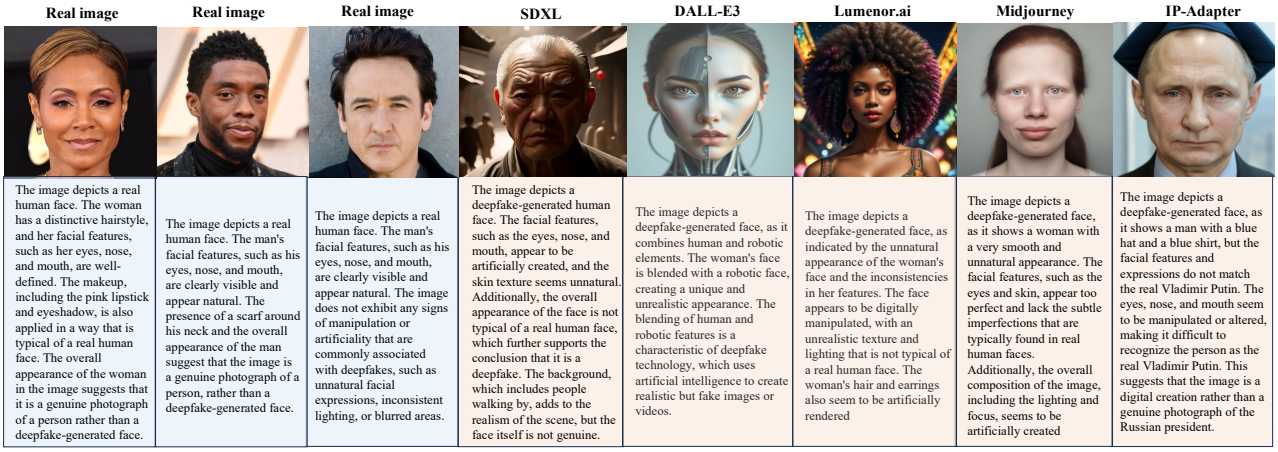


Figure 7. Evaluating deepfake images beyond FF++: the images with blue-shaded texts are real; the remaining images are AI-generated.

a BLEU-4 score of 0.3880. This shows limitations of using a vision encoder pre-trained on natural images, supporting our hypothesis that such encoders are insufficient for extracting deepfake-relevant features. To address this issue, we fine-tune the standard vision encoder with our specialized deepfake detection encoder, which is trained to capture both commonsense and statistical artifacts. After this modification, we observe significant improvements in both detection and reasoning performance. The model achieves an accuracy of 90.84% for deepfake detection and a BLEU-4 score of 0.4980 for reasoning, outperforming the baseline LLaVA model by a substantial margin. These results demonstrate the efficacy of our expert vision encoder in capturing deepfake-specific artifacts, enabling more accurate detection and improved reasoning capabilities.

Attention Map Visualization To further validate the efficacy of our proposed semantic learning module, we present a visual analysis of the attention maps [8] generated by our model, as depicted in Fig. 5. The text prompts displayed are pseudo-labels derived from the Llama 3.2 model [33], and all four images utilized in this analysis are synthetic deepfakes. As illustrated, by providing distinct text prompts for the same deepfake image, our model successfully guides the encoder to concentrate on the targeted regions, aligning with our objective of encouraging the model to learn semantic artifacts directed by pseudo-text descriptions. An examination of Fig. 5 reveals a strong correspondence between the attention masks and the associated text descriptions, indicating that our vision encoder effectively captures the relevant semantic artifacts as guided by the textual input.

Qualitative Examples of Deepfake Explanation Fig. 6 presents qualitative results on the DD-VQA dataset [60], comparing our model's responses with Llama 3.2. Our model accurately classifies all samples (top three: deepfake, bottom three: real) with precise explanations. While Llama 3.2 often expresses uncertainty in detection and provides only high-level explanations, our expert encoder-enhanced approach delivers both accurate classification and detailed semantic reasoning, particularly in identifying specific facial artifacts and structural inconsistencies commonly found in deepfakes. To evaluate our model's generalization, we test its performance on facial images from diverse generation methods (e.g., SDXL [40] and IP-Adapter [57]). As shown in Fig. 7, our model provides reasonable answers to queries, demonstrating its effectiveness beyond FF++ and its adaptability to recent image synthesis methods.

5. Limitations

AuthGuard cannot guarantee strictly causal explanations: subtle or hidden cues may yield plausible but non-causal rationales, a common limitation in explainable deepfake detection. While semantic guidance reduces reliance on low-level noise, causal faithfulness remains an open challenge. Nonetheless, explanations are still valuable in practice, as AuthGuard is designed to support rather than replace human judgment by highlighting why an image may appear suspicious.

6. Conclusion

This paper presents **AuthGuard**, a unified deepfake detection and reasoning framework that boosts both the generalization and interpretability of deepfake detection by combining statistical and commonsense deepfake artifacts. Extensive evaluations show that our approach outperforms existing methods in detection accuracy, generalization (6.15%), and interpretability (24.69%). By bridging specialized deepfake detection with multi-modal large language models, we take a step toward more transparent and generalizable deepfake detection systems. We hope our work contributes to combating misinformation, protecting digital identity, and fostering trust in media authenticity.

References

[1] Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024. 6, 7

[2] Sifat Muhammad Abdullah, Aravind Cheruvu, Shravya Kanchi, Taejoong Chung, Peng Gao, Murtuza Jadliwala, and Bimal Viswanath. An analysis of recent advances in deepfake image detection in an evolving threat landscape. *arXiv preprint arXiv:2404.16212*, 2024. 2

[3] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 6, 7

[4] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005. 5

[5] 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*, pages 4113–4122, 2022. 6

[6] Jie Chang, Zhonghao Lan, Changmao Cheng, and Yichen Wei. Data uncertainty learning in face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5710–5719, 2020. 4

[7] You-Ming Chang, Chen Yeh, Wei-Chen Chiu, and Ning Yu. Antifakeprompt: Prompt-tuned vision-language models are fake image detectors. *arXiv preprint arXiv:2310.17419*, 2023. 2

[8] Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 397–406, 2021. 6, 8

[9] Liang Chen, Yong Zhang, Yibing Song, Lingqiao Liu, and Jue Wang. Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18710–18719, 2022. 2

[10] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 6, 7

[11] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. 5

[12] 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. 2, 5, 6

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 1

[14] Matthew Groh, Ziv Epstein, Chaz Firestone, and Rosalind Picard. Deepfake detection by human crowds, machines, and machine-informed crowds. *Proceedings of the National Academy of Sciences*, 119(1):e2110013119, 2022. 2

[15] 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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 105–116, 2025. 2

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. pages 6840–6851, 2020. 1

[17] Baojin Huang, Zhongyuan Wang, Jifan Yang, Jiaxin Ai, Qin Zou, Qian Wang, and Dengpan Ye. Implicit identity driven deepfake face swapping detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4490–4499, 2023. 2

[18] Hossein Kashiani, Niloufar Alipour Talemi, and Fatemeh Afghah. Freqdebias: Towards generalizable deepfake detection via consistency-driven frequency debiasing. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8775–8785. IEEE, 2025. 6

[19] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 4

[20] Pavel Korshunov and Sébastien Marcel. Deepfake detection: humans vs. machines. *arXiv preprint arXiv:2009.03155*, 2020. 2

[21] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 6, 7

[22] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 2

[23] 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*, pages 5001–5010, 2020. 1, 2, 5, 6

[24] 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*, pages 3207–3216, 2020. 2

[25] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 5

[26] Yuzhen Lin, Wentang Song, Bin Li, Yuezun Li, Jiangqun Ni, Han Chen, and Qiushi Li. Fake it till you make it: Curricular dynamic forgery augmentations towards general deepfake

detection. In *European conference on computer vision*, pages 104–122. Springer, 2024. 2

[27] Honggu Liu, Xiaodan Li, Wenbo Zhou, Yuefeng Chen, Yuan He, Hui Xue, Weiming Zhang, and Nenghai Yu. Spatial-phase shallow learning: rethinking face forgery detection in frequency domain. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 772–781, 2021. 1, 5, 6

[28] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 3

[29] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 2, 3, 5, 6, 7

[30] Y Liu. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 5

[31] Ruilin Luo, Zhuofan Zheng, Yifan Wang, Yiyao Yu, Xinzhe Ni, Zicheng Lin, Jin Zeng, and Yujiu Yang. Ursa: Understanding and verifying chain-of-thought reasoning in multimodal mathematics. *arXiv preprint arXiv:2501.04686*, 2025. 1

[32] 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*, pages 16317–16326, 2021. 1, 5, 6

[33] Meta AI Research. Llama 3.2: Revolutionizing edge AI and vision with open, customizable models, 2024. Accessed: 2024-11-05. 3, 8

[34] MIT Media Lab. Detect fakes: Overview, n.d. Accessed: [2024-11-05]. 2

[35] Dat Nguyen, Nesryne Mejri, Inder Pal Singh, Polina Kuleshova, Marcella Astrid, Anis Kacem, Enjie Ghorbel, and Djamila Aouada. Laa-net: Localized artifact attention network for quality-agnostic and generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17395–17405, 2024. 2

[36] Yunsheng Ni, Depu Meng, Changqian Yu, Chengbin Quan, Dongchun Ren, and Youjian Zhao. Core: Consistent representation learning for face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12–21, 2022. 6

[37] Artemis Panagopoulou, Le Xue, Ning Yu, Junnan Li, Dongxu Li, Shafiq Joty, Ran Xu, Silvio Savarese, Caiming Xiong, and Juan Carlos Niebles. X-instructblip: A framework for aligning x-modal instruction-aware representations to llms and emergent cross-modal reasoning. *arXiv preprint arXiv:2311.18799*, 2023. 2

[38] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 5

[39] Gan Pei, Jiangning Zhang, Menghan Hu, Zhenyu Zhang, Chengjie Wang, Yunsheng Wu, Guangtao Zhai, Jian Yang, Chunhua Shen, and Dacheng Tao. Deepfake generation and detection: A benchmark and survey. *arXiv preprint arXiv:2403.17881*, 2024. 2

[40] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 8

[41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 4, 5, 7

[42] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1

[43] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics: A large-scale video dataset for forgery detection in human faces. *arXiv preprint arXiv:1803.09179*, 2018. 2, 5, 6

[44] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 1

[45] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 5

[46] Yichun Shi and Anil K Jain. Probabilistic face embeddings. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6902–6911, 2019. 2, 4

[47] Kaede Shiohara and Toshihiko Yamasaki. Detecting deepfakes with self-blended images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18720–18729, 2022. 1, 2

[48] Jiahe Tian, Cai Yu, Xi Wang, Peng Chen, Zihao Xiao, Jiao Dai, Jizhong Han, and Yesheng Chai. Real appearance modeling for more general deepfake detection. In *European Conference on Computer Vision*, pages 402–419. Springer, 2024. 2

[49] Emily Vaillancourt and Christopher Thompson. Instruction tuning on large language models to improve reasoning performance. *Authorea Preprints*, 2024. 3

[50] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 5

[51] Xiang Xu, Tianchen Zhao, Zheng Zhang, Zhihua Li, Jon Wu, Alessandro Achille, and Mani Srivastava. Principles of

designing robust remote face anti-spoofing systems. *arXiv preprint arXiv:2406.03684*, 2024. 1

[52] Yuting Xu, Jian Liang, Gengyun Jia, Ziming Yang, Yanhao Zhang, and Ran He. Tall: Thumbnail layout for deepfake video detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22658–22668, 2023. 2

[53] 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*, pages 22412–22423, 2023. 1, 2, 5, 6

[54] Zhiyuan Yan, Yong Zhang, Xinhang Yuan, Siwei Lyu, and Baoyuan Wu. Deepfakebench: A comprehensive benchmark of deepfake detection. *arXiv preprint arXiv:2307.01426*, 2023. 5

[55] Zhiyuan Yan, Yuhao Luo, Siwei Lyu, Qingshan Liu, and Baoyuan Wu. Transcending forgery specificity with latent space augmentation for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8984–8994, 2024. 1, 2, 5, 6

[56] Zhiyuan Yan, Taiping Yao, Shen Chen, Yandan Zhao, Xinghe Fu, Junwei Zhu, Donghao Luo, Li Yuan, Chengjie Wang, Shouhong Ding, et al. Df40: Toward next-generation deepfake detection. *arXiv preprint arXiv:2406.13495*, 2024. 2, 5, 6

[57] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 8

[58] Jiazuo Yu, Yunzhi Zhuge, Lu Zhang, Ping Hu, Dong Wang, Huchuan Lu, and You He. Boosting continual learning of vision-language models via mixture-of-experts adapters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23219–23230, 2024. 4

[59] Haocheng Yuan, Ajian Liu, Junze Zheng, Jun Wan, Jiankang Deng, Sergio Escalera, Hugo Jair Escalante, Isabelle Guyon, and Zhen Lei. Unified physical-digital attack detection challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 919–929, 2024. 2

[60] Yue Zhang, Ben Colman, Ali Shahriyari, and Gaurav Bharaj. Common sense reasoning for deep fake detection. *arXiv preprint arXiv:2402.00126*, 2024. 1, 2, 5, 6, 7, 8

[61] 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*, pages 15023–15033, 2021. 1, 2