

Exploring Gemini 2.5 for Explainable Deepfake Detection under Black-Box Constraints

Hyunjune Kim
Department of Computer Science and
Engineering
Sungkyunkwan University
South Korea
hyunjune.kim@g.skku.edu

Hyeongjun Choi
Department of Artificial Intelligence
Sungkyunkwan University
South Korea
junhjun@g.skku.edu

Simon S. Woo
Department of Computer Science and
Engineering
Sungkyunkwan University
South Korea
swoo@g.skku.edu

Abstract

The rapid advancement of deepfake generation poses significant challenges for reliable media verification. Effective detection increasingly demands methods that are both accurate and interpretable, motivating the use of multimodal large language models (MLLMs) for transparency and human-aligned explainability. While prior work has primarily focused on open-source MLLMs, we investigate, for the first time, the potential of a closed-source model, Google Gemini 2.5, for deepfake detection and explanation. We systematically evaluate Gemini via zero-shot testing and adapter-based black-box fine-tuning using Google Vertex AI. On a simple binary dataset (FaceForensics++), zero-shot performance is low and fine-tuning yields only modest gains. Remarkably, on a vision-language benchmark (DD-VQA), even straightforward black-box fine-tuning enables Gemini to outperform existing state-of-the-art models, highlighting the dataset-dependent impact of fine-tuning on closed-source models. Our study empirically demonstrates the feasibility of explainable deepfake detection using closed-source MLLMs, revealing both their promise and current limitations.

Keywords

Deepfake, Explainable AI, Multimodal AI, Human-centric AI

1 Introduction

Over the past decade, advances in deep learning have enabled the creation of highly realistic synthetic media (deepfakes). Their increasing prevalence blurs the lines between authentic and manipulated content. This trend raises profound concerns for security, privacy, and trust in digital information. Consequently, the development of reliable detection methods has become urgent. To date, deepfake detection has been framed mainly as a binary classification problem and approached with machine learning models. Early methods, however, suffer from limited transparency. Their decision processes are often opaque, which reduces reliability and interpretability. Deployable forensic tools require a new paradigm that goes beyond simple classification. Such tools must provide trustworthy and interpretable predictions.

One promising direction is the use of multimodal large language models (MLLMs). These models combine visual understanding with natural language reasoning, offering a path toward explainable deepfake detection. Their capabilities have motivated the field to explore the potential of MLLMs-based forensic methods. Recent work shows that MLLMs, such as GPT-4V [1] and Gemini

1.0-pro-vision [14], can detect AI-generated images with carefully designed prompts [6]. Benchmarks have also emerged to evaluate both open- and closed-source MLLMs on synthetic data detection across diverse modalities [20]. However, these studies are limited to inference-only evaluations and do not examine more costly adaptation procedures, such as fine-tuning. Further studies have explored fine-tuning MLLMs on vision-language datasets. BLIP [7]-based models, for example, are adapted to perform reasoning over visual artifacts beyond deepfake detection [21]. Similarly, LLaVA [10]-style frameworks have been used for authenticity classification and artifact explanation tasks [8, 17]. Despite these advances, all these efforts are confined to open-source architectures. The potential of fine-tuning commercial, closed-source MLLMs remains largely unexplored. In summary, prior work leaves two gaps: (i) the performance of industrial-grade, closed-source MLLMs in explainable deepfake detection remains untested; and (ii) the effects of costly, black-box fine-tuning are largely unexplored.

To address the above gaps, we present the first systematic evaluation of Google Gemini 2.5 [4] for deepfake detection and explanation. We move beyond the common practice of evaluating or fine-tuning open-source MLLMs. Instead, we study Google’s commercial, closed-source MLLM, the Gemini 2.5 family. We examine its capacity for deepfake detection and explanation under two settings: zero-shot inference and black-box fine-tuning via Vertex AI¹. We first test Gemini on standard deepfake datasets designed primarily for binary classification. We craft targeted questions to probe Gemini’s zero-shot classification ability. We also design a simple question-answer format to evaluate its performance after fine-tuning. In these tasks, Gemini underperforms compared to existing vision-only detectors optimized for deepfake detection. Even with paid fine-tuning through a black-box service, the model struggles to adapt effectively or achieve strong generalization. We hypothesize that the source of this weakness lies in our dataset design. The binary format we construct, requiring only "yes" or "no" answers, does not align well with how MLLMs learn. Prior work [17] shows that limiting the models to binary outputs such as "real" or "fake" can impair both detection and explanatory ability. They show that reframing it as a visual question answering, where the model must classify and describe image artifacts, improves alignment between textual responses and visual content. Inspired by this, we explore Gemini’s potential using well-defined datasets designed for both classification and explanation. We extend the task by asking Gemini not only to detect deepfakes but

¹<https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini-use-supervised-tuning>

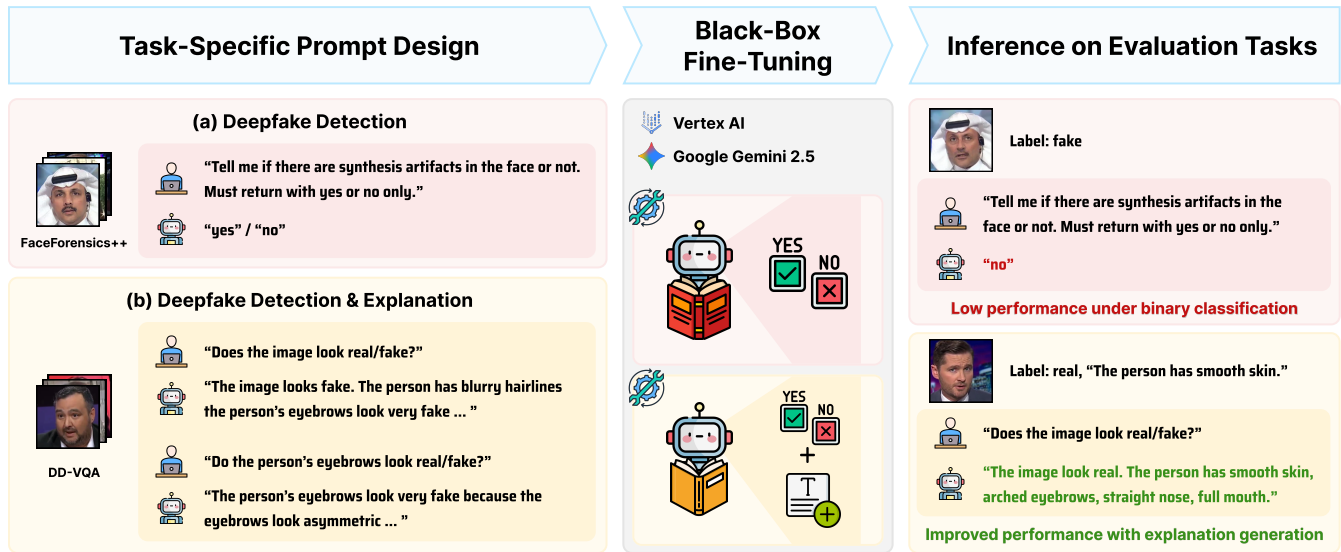


Figure 1: Illustration of the Gemini fine-tuning scenarios. The study is designed with two scenarios: (a) fine-tuning for deepfake detection on FF++ (c23), and (b) fine-tuning with DD-VQA for detection and explanation. Black-box fine-tuning is performed via Vertex AI, with costs depending on training tokens and configuration.

also to describe anomalies in facial components. Unfortunately, in zero-shot inference, both detection and explanation remain far below previous baseline levels. Gemini fails to recognize or explain manipulations that are obvious even to humans. In contrast, fine-tuned Gemini surpasses existing state-of-the-art models in both detection and explanation. Achieving such performance in a restricted fine-tuning setting is a surprising result. These findings show that commercial models offer capabilities beyond conventional model-centric approaches. We also share practical insights from our experiments, including dataset construction, hyperparameter choices, and cost-performance trade-offs.

Our main contributions are summarized as follows:

- We present the first evaluation of the commercial Gemini 2.5 family on deepfake detection and explanation. This extends the community’s focus beyond open-source MLLMs.
- We analyze Gemini’s adaptation to deepfake tasks under zero-shot inference and black-box-based fine-tuning. We show that with well-designed image-text pairs, commercial MLLMs can achieve and even surpass state-of-the-art performance.
- Using both detection and explanation-oriented datasets, we highlight considerations for data construction, hyperparameter settings, and cost trade-offs. These offer guidance for future use of commercial MLLMs in deepfake research.

2 Method

2.1 Study Design

We investigate the ability of Gemini 2.5 family to detect and explain deepfakes. Our core question is whether a commercial model, with only limited user control, can adapt effectively to this task. We focus on two experimental scenarios: zero-shot testing and black-box fine-tuning. The overall workflow is illustrated in Figure 1. Our

study proceeds in two main steps. In the first step, we evaluate Gemini on a standard **real vs. fake binary dataset**. For *zero-shot inference*, we test whether Gemini can distinguish real and fake images without any task-specific training. Due to the cost of inference, we design prompts carefully to elicit the model’s semantic knowledge. Following prior work [6], we use the prompt: "Tell me if there are synthesis artifacts in the face or not. Must return with yes or no only." This design reduces rejection rates and avoids overly complex context that may lower accuracy. We then parse outputs into "yes" or "no" to measure detection performance. For *black-box fine-tuning*, we use the Google Vertex AI’s service to access the Gemini 2.5. This service allows users to adjust only limited settings such as epochs, adapter size, and learning rate multiplier. *Epoch* denotes how many passes are made over the training data, with the default automatically adjusted to the dataset size. *Adapter size* controls the number of trainable parameters, where larger sizes offer higher capacity but require more data and training steps. The *learning rate multiplier* scales the default rate, enabling either faster convergence or more conservative updates. Unless otherwise noted, we set *epoch* = 3, *learning rate multiplier* = 5, and *adapter size* = 4. Results without explicit mention correspond to this default configuration. During fine-tuning, we pair each image with the aforementioned question prompt as the question and assign "yes" or "no" as the corresponding answer, thereby training the model to generate strictly binary responses.

In the second step, we move beyond binary classification datasets by using a **question-answer (VQA-style) dataset** that supports both detection and explanation. This tests whether Gemini can not only classify real vs. fake but also justify its decisions in natural language. For *zero-shot inference*, we use question-answer based deepfake dataset (e.g., "Does the person’s skin look real or fake?") and analyze Gemini’s responses (e.g., "The person’s

skin looks real. The skin is naturally illuminated.".) We evaluate both detection accuracy and the quality of generated explanation. To fine-tune Gemini, we used the *black-box fine-tuning* provided by Vertex AI. This time we use structured question-answer pairs from VQA-based dataset to adapt Gemini for joint detection and explanation. In addition, we analyze the impact of *adapter size* and *model size* on performance, highlighting how these components interact within the restricted fine-tuning regime.

2.2 Datasets

We use two widely adopted benchmark datasets: FaceForensics++ (FF++) [13] and DD-VQA [21]. The FF++ dataset, widely used benchmark for facial forgery detection, containing videos generated by four manipulation methods: Deepfakes, Face2Face, FaceSwap, and NeuralTextures. We use the commonly employed c23 versions. Video pre-processing follows the DeepfakeBench [19] protocol, extracting frames for analysis. DD-VQA is a dataset designed for artifact explanation in faces. It extends FF++ (c40 compression) with human-annotated question-answer pairs about visual artifacts such as blurred hairlines, mismatched eyebrows, rigid pupils, and unnatural shadows. DD-VQA supports both binary classification and explanation tasks.

2.3 Baselines

We compare Gemini’s performance against representative baselines for each benchmark. For FF++, we follow the common setup of training on the c23 dataset and evaluating inter-dataset detection. We select F3-Net [12], RFM [16], MultiAtt [22], RECCE [3], and TALL [18] as baselines. These models define the typical performance range and provide a reference for comparison with Gemini. For DD-VQA, we use the baselines reported in the original work [21]: BLIP, BLIP-T, BLIP-I, and BLIP-TI. BLIP-T applies their proposed text contrastive loss, BLIP-I applies their image contrastive loss, and BLIP-TI combines both. BLIP-TI is presented as the state-of-the-art. We adopt these results as reference points to assess Gemini’s relative performance.

2.4 Evaluation Metrics

We evaluate Gemini on two benchmarks. For FF++ (c23), we report accuracy (ACC). For DD-VQA, we report detection metrics including accuracy (ACC), Precision, Recall, and F1-score (F1). To assess the quality of generated responses, we use BLEU-4 [11], CIDEr [15], ROUGE-L [9], METEOR [5], and SPICE [2], following [21].

3 Results

3.1 Binary Classification

We first present intra-testing results on the FF++ [13] dataset in Table 1. In the *zero-shot* setting, we evaluate both the Gemini-flash and Gemini-pro models. The detection accuracy (ACC) is 44.96% and 46.73%, respectively. These values are far below deepfake-specialized baselines (row 1–5). The results indicate that commercial models remain poorly adapted to the deepfake domain. For *fine-tuning*, we use only the flash model, since it offers a better cost-performance trade-off. After fine-tuning, the Gemini-flash model reached 84.99% ACC, representing a large improvement. However,

Table 1: Intra-testing results on FF++ (c23).

Methods	Venue	ACC (%)
F3-Net [12]	ECCV20	97.52
RFM [16]	CVPR21	95.69
MultiAtt [22]	CVPR21	97.60
RECCE [3]	CVPR22	97.06
TALL [18]	ICCV23	98.65
Gemini-flash (zero-shot)	-	44.96
Gemini-pro (zero-shot)	-	46.73
Gemini-flash (fine-tuned)	-	84.99

Table 2: Intra-testing accuracy (%) on FF++ (c23) by category.

Methods	DF	F2F	FS	NT	Orig.
Gemini-flash (zero-shot)	58.70	23.39	41.32	4.24	97.14
Gemini-pro (zero-shot)	57.85	29.44	44.51	4.80	97.05
Gemini-flash (fine-tuned)	98.03	93.30	95.06	82.60	55.95
Gemini-flash (bal. & fine-tuned)	93.28	85.39	85.92	63.54	88.88

the score is still lower than prior baselines, revealing clear limits. Taken together, these results show that even with commercial models and costly fine-tuning, deepfake detection remains unsolved, even on FF++, where forgeries are easier for humans to spot.

Moreover, we conduct deeper analysis. Table 2 reports per-class ACC for five types in the FF++ dataset: Deepfakes (*DF*), Face2Face (*F2F*), FaceSwap (*FS*), NeuralTextures (*NT*), and Original (*Orig.*). In the *zero-shot* case, Gemini-flash and Gemini-pro predict the real class (e.g., *Orig.*) with about 97% accuracy, while failing on fake classes. This reflects a strong conservative bias, defaulting to real predictions. Interestingly, *fine-tuning* reverses the trend. ACC for the four fake classes increases sharply, but ACC for the real class drops to 55.95%. We suspect this effect arises from class imbalance in FF++. The dataset contains 1000 real and 1000 fake videos for each of four manipulation types, yielding 1:4 imbalance in binary form. Training on this split likely biases the model toward fake detection. To test this, we duplicate real samples to match the number of fake samples. The last row of Table 2 reports results. The balanced setting yields similar overall ACC but achieves better trade-off between real and fake accuracy. These results imply that commercial models, even with paid fine-tuning, are highly sensitive to class imbalance in training data.

3.2 Binary Classification with Explanation

We next test on DD-VQA [21] dataset, which supports both binary classification and natural language explanation. We hypothesize that constraining MLLMs to binary detection only, while ignoring

Table 3: Performance of baselines and Gemini variants on DD-VQA deepfake detection (%) and answer-generation.

Methods	Deepfake Detection				Answer Generation				
	ACC	Precision	Recall	F1	BLUE-4	CIDeR	ROUGE_L	METEOR	SPICE
BLIP [21]	81.68	78.61	95.96	86.42	0.3569	1.8177	0.5664	0.3301	0.6658
BLIP-T [21]	83.65	81.31	94.89	87.58	0.3714	1.8715	0.5774	0.3349	0.6710
BLIP-I [21]	84.87	82.98	94.48	88.36	0.3800	0.3800	0.5882	0.3419	0.6788
BLIP-TI [21]	87.49	86.97	93.41	90.07	0.4075	2.0567	0.6085	0.3463	0.6915
Gemini-flash (zero-shot)	57.56	88.41	39.61	54.71	0.0088	0.0058	0.1385	0.2094	0.1278
Gemini-pro (zero-shot)	69.04	53.47	91.67	67.54	0.0081	0.0000	0.1255	0.2356	0.1049
Gemini-flash (fine-tuned, <i>adapter</i> = 1)	88.24	88.46	92.66	90.51	0.5552	3.1076	0.7182	0.7081	0.6801
Gemini-flash (fine-tuned, <i>adapter</i> = 4)	79.36	83.17	82.61	82.89	0.5081	2.9452	0.6878	0.6812	0.6698
Gemini-flash (fine-tuned, <i>adapter</i> = 16)	81.25	83.33	86.28	84.78	0.5175	2.8762	0.6893	0.6801	0.6648
Gemini-pro (fine-tuned, <i>adapter</i> = 1)	80.02	84.96	81.39	83.14	0.3196	1.9603	0.5822	0.5819	0.5839

its explanatory capacity, leads to suboptimal results. Table 3 summarizes findings. In the *zero-shot* case, Gemini-flash and Gemini-pro achieve 57.56% and 69.04% ACC, respectively. Both are below the BLIP-based baselines (row 1–4). The two Gemini variants also obtain the lowest scores across all four explanation metrics, including BLEU-4, CIDeR, ROUGE-L, and METEOR. These results show that commercial models without domain training struggle both in detection and in generating proper explanations. After *fine-tuning*, Gemini-flash with adapter size 1 achieves substantial gains. It outperforms the previous state-of-the-art model (BLIP-TI) on both detection and explanation. Prior baselines are optimized for deepfake detection with specialized objectives, such as contrastive learning. In contrast, Gemini reaches superior performances with simple black-box fine-tuning. This highlights its strong potential.

We then analyze the effect of **adapter size** when fine-tuning Gemini-flash. Results on DD-VQA reveal that using the adapter size 1 yields the best performance across both detection and explanation. In contrast, larger adapter sizes (4 and 16) consistently degrade performance. Increasing the adapter size enlarges the number of trainable parameters, which we expected to improve adaptation to the new task. However, the results show the opposite trend. We hypothesize that larger adapters may introduce over-parameterization under limited training signals, leading to reduced generalization. This suggests that, for closed-source models fine-tuned through restricted environment, smaller adapters can provide a better trade-off between adaptation capacity and stability.

Finally, we examine the **model capacity**. Gemini-pro is expected to have higher capacity than Gemini-flash. It also uses an internal reasoning process² to enhance multi-step planning during generation, unlike Gemini-flash. Surprisingly, fine-tuning results show that Gemini-flash actually outperforms Gemini-pro. This suggests that success in deepfake detection and explanation may not depend on model capacity or advanced reasoning features, but rather on the efficient adaptation under black-box constraints. Figure 2 illustrates the effect of fine-tuning on Gemini-flash, showing that the zero-shot model relies on lighting cues, whereas the fine-tuned one detects fakeness from facial inconsistencies.

²<https://ai.google.dev/gemini-api/docs/thinking>

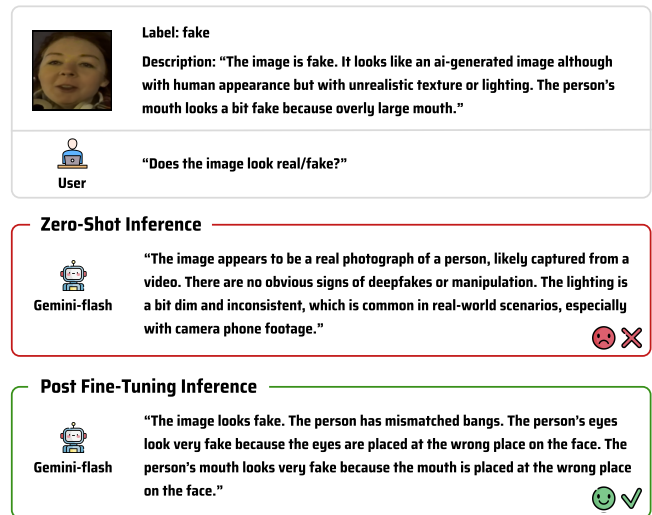


Figure 2: Qualitative comparison of zero-shot and post fine-tuning inferences generated by Gemini 2.5 flash on one sample from the test set.

4 Conclusion and Discussion

Motivated by curiosity about whether state-of-the-art commercial models can address deepfake detection and explanation, we explore the potential of the closed-source MLLM, Google Gemini 2.5. Zero-shot settings scenario shows a bias toward real classifications. While FF++ fine-tuning scenario offers limited gains, DD-VQA fine-tuning scenario significantly improves accuracy and explanation quality, showing that models perform better when the instructions are tailored to the task. A key limitation of this study is the restricted access to hyperparameters and internal modifications due to the black-box nature of Vertex AI fine-tuning. The cost is also significant, with hyperparameter tuning leading to approximately \$5,000 (FF++) and \$300 (DD-VQA) in expenses, including trial-and-error efforts. We encourage industry investment in deepfake detection and note that open-source advanced MLLMs, such as the Gemini family, can be effectively leveraged for this task.

Acknowledgments

This work was partly supported by Institute for Information & communication Technology Planning & evaluation (IITP) grants funded by the Korean government MSIT: (RS-2022-II221199, RS-2022-II220688, RS-2019-II190421, RS-2023-00230337, RS-2024-004378 49, RS-2021-II212068, and RS-2025-02263841). Also, this work was supported by the Cyber Investigation Support Technology Development Program (No.RS-2025-02304983) of the Korea Institute of Police Technology (KIPoT), funded by the Korean National Police Agency. Lastly, this work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2024-00356293).

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).
- [2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *European conference on computer vision*. Springer, 382–398.
- [3] Junyi Cao, Chao Ma, Taiping Yao, Shen Chen, Shouhong Ding, and Xiaokang Yang. 2022. End-to-end reconstruction-classification learning for face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4113–4122.
- [4] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261* (2025).
- [5] Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*. 376–380.
- [6] Shan Jia, Reilin Lyu, Kangran Zhao, Yize Chen, Zhiyuan Yan, Yan Ju, Chuanbo Hu, Xin Li, Baoyuan Wu, and Siwei Lyu. 2024. Can chatgpt detect deepfakes? a study of using multimodal large language models for media forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4324–4333.
- [7] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*. PMLR, 12888–12900.
- [8] Yixuan Li, Yu Tian, Yipo Huang, Wei Lu, Shiqi Wang, Weisi Lin, and Anderson Rocha. 2025. Fakescope: Large multimodal expert model for transparent ai-generated image forensics. *arXiv preprint arXiv:2503.24267* (2025).
- [9] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.
- [10] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems* 36 (2023), 34892–34916.
- [11] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 311–318.
- [12] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. 2020. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European conference on computer vision*. Springer, 86–103.
- [13] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. 2019. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*. 1–11.
- [14] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* (2023).
- [15] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4566–4575.
- [16] Chengrui Wang and Weihong Deng. 2021. Representative forgery mining for fake face detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 14923–14932.
- [17] Siwei Wen, Junyan Ye, Peilin Feng, Hengrui Kang, Zichen Wen, Yize Chen, Jiang Wu, Wenjun Wu, Conghui He, and Weijia Li. 2025. Spot the fake: Large multimodal model-based synthetic image detection with artifact explanation. *arXiv preprint arXiv:2503.14905* (2025).
- [18] Yuting Xu, Jian Liang, Gengyun Jia, Ziming Yang, Yanhao Zhang, and Ran He. 2023. Tall: Thumbnail layout for deepfake video detection. In *Proceedings of the IEEE/CVF international conference on computer vision*. 22658–22668.
- [19] Zhiyuan Yan, Yong Zhang, Xinhang Yuan, Siwei Lyu, and Baoyuan Wu. 2023. Deepfakebench: A comprehensive benchmark of deepfake detection. *arXiv preprint arXiv:2307.01426* (2023).
- [20] Junyan Ye, Baichuan Zhou, Zilong Huang, Junan Zhang, Tianyi Bai, Hengrui Kang, Jun He, Honglin Lin, Zihao Wang, Tong Wu, et al. 2024. Loki: A comprehensive synthetic data detection benchmark using large multimodal models. *arXiv preprint arXiv:2410.09732* (2024).
- [21] Yue Zhang, Ben Colman, Xiao Guo, Ali Shahriyari, and Gaurav Bharaj. 2024. Common sense reasoning for deepfake detection. In *European conference on computer vision*. Springer, 399–415.
- [22] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. 2021. Multi-attentional deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2185–2194.