REVEAL – Reasoning and Evaluation of Visual Evidence through Aligned Language

Ipsita Praharaj[1] Yukta Butala[1] Badrikanath Praharaj[2] Yash Butala[1]
Carnegie Mellon University[1] VIT Bhopal[2]

Abstract

The rapid advancement of generative models has intensified the challenge of detecting and interpreting visual forgeries, necessitating robust frameworks for image forgery detection while providing reasoning as well as localization. While existing works approach this problem using supervised training for specific manipulation or anomaly detection in the embedding space, generalization across domains remains a challenge. We frame this problem of forgery detection as a prompt-driven visual reasoning task, leveraging the semantic alignment capabilities of large vision-language models. We propose a framework, 'REVEAL' (Reasoning and Evaluation of Visual Evidence through Aligned Language), that incorporates generalized guidelines. We propose two tangential approaches - (1) Holistic Scene-level Evaluation that relies on the physics, semantics, perspective, and realism of the image as a whole and (2) Region-wise anomaly detection that splits the image into multiple regions and analyzes each of them. We conduct experiments over datasets from different domains (Photoshop, DeepFake and AIGC editing). We compare the Vision Language Models against competitive baselines and analyze the reasoning provided by them.

1. Introduction

The rapid evolution of deep generative models, such as Stable Diffusion [24], DALLE [7, 17] has enabled the creation of high-quality, realistic images. However, these tools pose societal risks—spreading misinformation, enabling cyberbullying, violating copyright through falsely claimed works [31] and also unfair outcomes in creative competitions [28]. These concerns underscore the urgent need for reliable image forgery detection to protect authenticity and intellectual property. As shown in 1, image forgery detection aim to identify manipulated images and explain the manipulations.

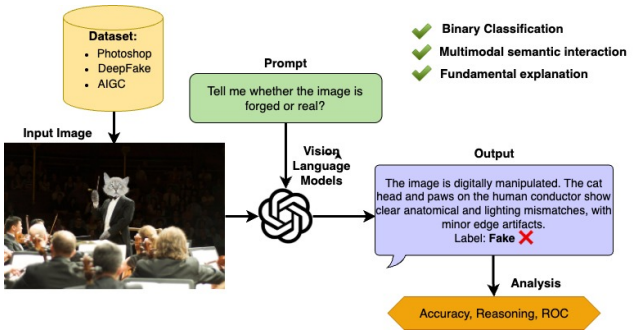
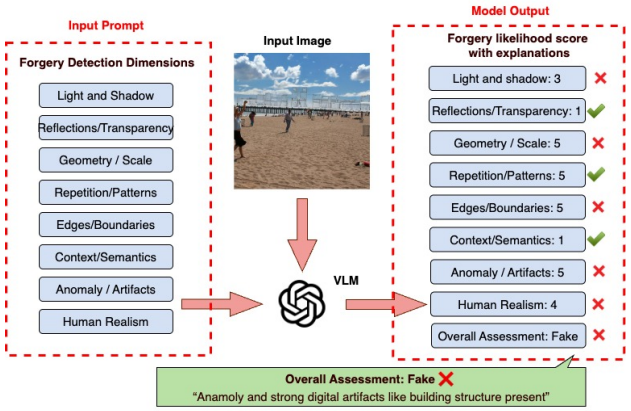


Figure 1. An overview of the image forgery detection task using multimodal large language models. We propose a structured prompt-based approach, REVEAL, to detect forgery and provide nuanced explanations.

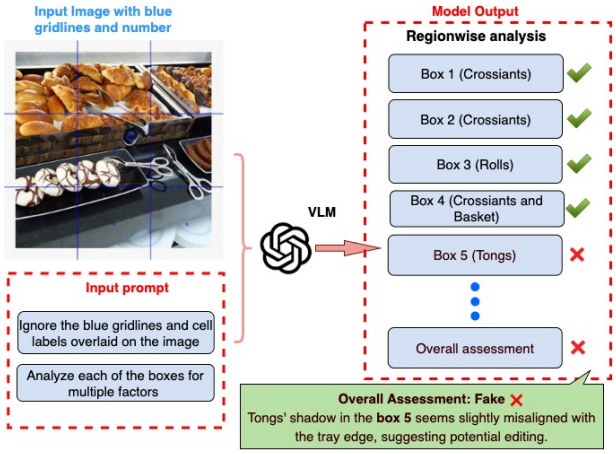
Increase in Realistic Image Forgeries Using Visual Generative Models: Early image generation began with Generative Adversarial Networks (GANs) [8], progressing to more stable variants like SA-GAN [34] and BigGAN [1]. StyleGAN [13] enabled fine-grained stylistic control, further improving image realism. More recently, diffusion-based models such as Stable Diffusion [24], DALLE-2 [17], and DALLE-3 [7] have advanced text-guided image synthesis, producing highly realistic forgeries. As a result, each new model increases the challenge of forgery detection.

Multimodal Large Language Models for Forgery Detection: Early models like CLIP [23] and BLIP [14] aligned visual and textual features for tasks such as image classification and captioning. Recent MLLMs, including InstructBLIP [3] and LLaVA [15], leverage instruction tuning for complex tasks like visual question answering. Unlike CNN-based approaches [5, 10, 27], MLLMs provide scene-level understanding and interpretable, text-driven analysis—capabilities critical for robust forgery detection.

Current gaps in Forgery Detection: Traditional methods, such as binary classifiers [5, 9] have relied on low-level visual cues, like noise patterns or texture inconsistencies. While effective for specific generative models like GANs [29], these approaches struggle with generalizability, particularly against advanced diffusion-based models like



(a) Holistic scene-level evaluation. The model reasons over the entire image for eight different dimensions as shown. It provides the Likert score [1-5] for 'authentic' to 'tampering' for each of the dimensions, followed by an overall assessment.



(b) Region-wise anomaly detection: The input image is divided into 9 labeled blocks. The model is expected to reason about forgery detection in each of the boxes before providing the final assessment.

Figure 2. The images show two kinds of prompt-strategies used in the REVEAL approach

Stable Diffusion [24] and DALLE-2 [17], due to overfitting on training data [27]. To address this, recent MLLM-based approaches like De-Fake[25], AntifakePrompt [2], ForgeryGPT [16], and FakeShield [32] integrate visual and linguistic reasoning for interpretable forgery detection and identifying the precise location of the forged detail. FakeShield [32] is a fine-tuned MLLM for the forgery classification task trained on real and fake images which also uses visual segmentation mask of the image. [12] explored the capability of ChatGPT in media forensics, highlighting MLLMs' potential in detecting subtle manipulations. However, their framework is limited to the forensics domain with only GAN generated images.

Key Contributions: We leverage the semantic and spatial reasoning abilities of vision-language models (VLMs) through structured prompting, as illustrated in Figure 2. Our work introduces two complementary zero-shot approaches for image forgery detection: (1) **Holistic Scene Evaluation**, which assesses realism, physics, and semantics at the image level; and (2) **Region-wise Anomaly Detection**, which analyzes labelled grid segments for localized inconsistencies. While these methods can be further improved with fine-tuning or segmentation pipelines, our focus is on evaluating zero-shot performance with carefully designed instructions. Our main contributions are:

1. We propose a dataset-agnostic prompt framework, REVEAL, that enables MLLMs to detect forgeries by reasoning from first principles. We evaluate two zero-shot prompting strategies: holistic and region-wise analysis.
2. We benchmark leading MLLMs against fine-tuned baselines and a simple binary prompt classifier across three image forgery domains from the Fakeshield dataset, and analyze the interpretability and localization of the reasoning produced by REVEAL.

2. Methodology

2.1. Model and Dataset Details

Similar to the finetuned approach in FakeShield [32], we select several challenging public benchmark datasets including PhotoShop tampering (CASIA1+ [6], Columbia [18], IMD2020 [19], Coverage [19], DeepFake tampering (e.g., FFHQ, FaceApp [4], Seq-DeepFake [26], and AIGC-Editing data from FakeShield [32]. We select this scope due to the prevalence of these forgery types in real-world scenarios, their semantic complexity, and the availability of established visual cues for benchmarking. We employ open MLLMs, including LLaVA [15], GPT-4o [21], and GPT-4.1 [20], Gemini [22].

2.2. Prompting Strategies

We systematically evaluate on the baseline prompts and two prompts specialized prompts (provided in Appendix [1], each targeting a different aspect of VLM reasoning:

- **Baseline Prompt:** A simple binary classification prompt ("Is this image real or fake?") to assess the VLM's out-of-the-box ability without explicit reasoning or spatial cues.
- **Holistic Scene-Level Prompt (Image Semantics):** A detailed checklist-based prompt requiring the model to assess global scene plausibility across eight forensic dimensions: *lighting/shadow, reflections/transparency, perspective/geometry, repetition/patterns, edge/boundary,*

[1] We upload the appendix to Google Drive to adhere to the page limit. Link

Table 1. The table shows the performance of benchmarks and prompt-based baselines over different tasks for image forgery detection.

Method	Photoshop								DeepFake		AIGC-Editing	
	CASIA1+		IMD2020		Columbia		Coverage					
	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1
Finetuned Baselines												
SPAN	0.60	0.44	0.70	0.81	0.87	0.93	0.24	0.39	0.78	0.78	0.47	0.05
ManTraNet	0.52	0.68	0.75	0.85	0.95	0.97	0.95	0.97	0.50	0.67	0.50	0.67
HiFi-Net	0.46	0.44	0.62	0.75	0.68	0.81	0.34	0.51	0.56	0.61	0.49	0.42
PSCC-Net	0.90	0.89	0.67	0.78	0.78	0.87	0.84	0.91	0.48	0.58	0.49	0.65
CAT-Net	0.88	0.87	0.68	0.79	0.89	0.94	0.23	0.37	0.85	0.84	0.82	0.81
MVSS-Net	0.62	0.76	0.75	0.85	0.94	0.97	0.65	0.79	0.84	0.91	0.44	0.24
FakeShield	**0.95**	**0.95**	**0.83**	**0.90**	**0.98**	**0.99**	**0.97**	**0.98**	**0.93**	**0.93**	**0.98**	**0.99**
Baseline Prompt												
LLaVA-1.5-7b-hf	0.5	0	0.49	0	0.5	0	0.44	0	0.50	0	0.50	0
Gemini 2.0 Flash	0.67	0.5	0.69	0.58	0.62	0.39	0.47	0.08	0.83	0.8	0.54	0.23
GPT-4o	0.5	0.4	0.68	0.58	0.79	0.73	0.42	0	0.57	0.25	0.53	0.18
GPT-4.1	0.83	0.8	**0.74**	**0.70**	0.97	0.97	0.51	0.27	0.73	0.64	0.62	0.46
Holistic Scene level evaluation prompt												
LLaVA-1.5-7b-hf	0.58	0.62	0.56	0.60	0.73	0.75	0.60	0.70	0.43	0.41	0.59	0.65
Gemini 2.0 Flash	0.83	0.80	0.72	0.67	0.87	0.85	0.47	0.08	**0.93**	**0.93**	**0.66**	**0.53**
GPT-4o	0.75	0.67	0.69	0.62	0.94	0.94	0.49	0.26	0.62	0.39	0.55	0.26
GPT-4.1	**0.92**	**0.92**	0.71	0.64	**0.97**	**0.97**	0.47	0.08	0.66	0.47	0.62	0.42
Region-wise anomaly detection prompt												
LLaVA-1.5-7b-hf	0.58	0.62	0.48	0.41	0.55	0.51	0.51	0.52	0.45	0.41	0.60	0.44
Gemini 2.0 Flash	0.83	0.86	0.65	0.67	0.82	0.80	0.49	0.15	0.62	0.68	0.60	0.44
GPT-4o	0.67	0.60	0.60	0.59	0.77	0.76	0.47	0.20	0.56	0.46	0.51	0.22
GPT-4.1	**0.92**	**0.91**	0.67	0.59	0.83	0.81	0.44	0.07	0.60	0.33	0.60	0.44

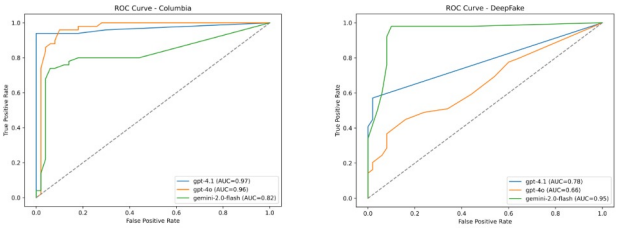


Figure 3. We plot the ROC curve using average of factor-scores for the Holistic-eval approach as a proxy to the probability. We find that GPT-4.1 performs well over the Columbia dataset while Gemini is the best amongst the baselines over DeepFake dataset.

contextual/semantic consistency, anomaly/artifact detection, and human/object realism based on [11]. For each factor, the model provides a Likert score from 1 to 5 (authentic to forged) and a concise factor-wise justification, followed by a global label and a brief major reasoning of the most decisive evidence from the 8 factors.

- **Region-wise Anomaly Detection Prompt (Image Detailing):** Inspired by recent advances in visual grounding - Set of Mark [33], this prompt overlays a 3x3 labelled grid over the image and instructs the model to first perform a holistic assessment, then analyze each grid cell for local anomalies only if no holistic cues are found. The model integrates both overall and localized (cell-specific) findings into a single, concise explanation, referencing cell numbers for local issues.

2.3. Metrics

Using the predicted labels from each of the approaches (baseline, holistic, region-wise), we report accuracy and F1 score over the classification task of image forgery detection. Additionally, for the Holistic Scene-level evaluation, we average the Likert scores for each of the factors to get a tampering score between 0 to 1. It is used to obtain the ROC curve for threshold-independent comparison between different MLLMs, as shown in figure 3.

3. Results

3.1. Quantitative Results

The main findings as shown in the table 1 are:

- Structured prompts (holistic and region-wise) in general significantly improve the performance over baseline prompts. For example, GPT-4.1's accuracy on CASIA1+ increases from 0.83 (baseline) to 0.92 (holistic/region-wise), and Gemini's F1 on Columbia [18] rises from 0.39 (baseline) to 0.85 (holistic) and 0.80 (region-wise).
- Structured prompts have the greatest impact on datasets where the baseline model struggles—typically those with subtle, complex, or localized manipulations (e.g., DeepFakes, AIGC-Editing). For example, Gemini and GPT-4o show significant improvement on DeepFake and AIGC datasets when using structured prompts, as these prompts help the models focus on specific forensic cues. Similarly, for the smaller model LLaVA, structured prompts improve performance consistently and significantly, where accuracy and F1 scores rise from near-zero to moderate,

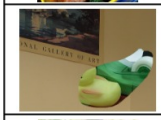
Image	Dataset	Gold Label	Baseline Prompt Response	Holistic Scene Evaluation Prompt	Region-wise Evaluation Prompt
	CASIA++	Forged	Natural colors and appropriate textures suggest this lunchbox meal is Authentic	The image depicts a plausible lunchbox scene with natural lighting and consistent container angles, though minor AI artifacts, slight plastic glare, and some odd edges and textures suggest moderate digital generation.	Distributed color saturation, harsh shadows, and unnaturally smooth textures suggest digital manipulation across the image.
	Columbia	Forged	The inserted duck and tape image segment is poorly blended and mismatched with the background.	The image shows clear signs of digital collage, including abrupt lighting mismatches, misaligned reflections, mismatched perspective, and visible pasted boundaries, with implausible elements like a misplaced rubber duck indicating GPT-4.1 generation.	All grid cells exhibit appropriate lighting and texture without signs of manipulation.
	DeepFake	Forged	Background, hair, and ear edges are consistent with natural lighting and image composition.	The image appears authentic, with no visible artifacts or inconsistencies, natural outdoor lighting, clean edges, consistent depth, and a plausible context, with natural-looking hair and ears.	Background and hairline textures are consistent; lighting and object boundaries appear natural overall.
	AIGC	Forged	Lighting, shadows, and object edges are consistent; no obvious digital artifacts present.	The scene is contextually plausible with consistent geometry, soft lighting, and no repetition, but the face and eyes appear artificial, with unnatural details and expression suggesting digital generation.	Lighting, shadows, and details of the plush toy and room are consistent across all grid cells.
	Coverage	Forged	Consistent lighting, natural shadows, no visible artifacts, text and products coherent across image.	Motion blur present, but no artifacts. Lighting and object proportions consistent. Minor glare and edge blur. Scene and items appear natural and plausible.	There is an unnatural shadow transition beneath the vases, especially across cells 8 and 9, and odd blending along the handles and pedestal boundaries indicating digital manipulation. *(**blue grid-label used for region-wise eval approach)*
	IMD2020	Authentic	Two plush toys are naturally positioned in a well-lit, ordinary home environment.	The image presents a realistic drone scene with no artifacts or glitches, consistent lighting and shadows, plausible perspective and geometry, sharp edges with minor blur, and a logical, coherent context.	No visual anomalies in all boxes drone, wind turbines, and shadows consistent throughout image.

Figure 4. We pick examples for qualitative analysis. The outputs for "Holistic Scene-Evaluation" and "Region-wise Evaluation" prompt are made concise using GPT.

showing LLaVA can use scene-level cues when prompted with explicit guidance. Thus, structured prompting is most beneficial for challenging datasets and weaker models that can follow the prompt. However, in contrast, on datasets like Photoshop forgeries (CASIA1+, Columbia), where GPT-4.1 already performs well, the benefit of structured prompts is less pronounced.

- Prompt-based methods achieve low F1-score over the Coverage [30] data due to subtle and distributed edits.

3.2. Qualitative Analysis

Figure 4 shows sample GPT-4.1 results. Key findings:

- **Region-wise prompts excel at local manipulations:** Grid-based prompts effectively detect localized edits (e.g., AIGC-Editing, Photoshop splicing), but may miss manipulations spanning multiple cells if inter-cell consistency is not considered. For example, in Figure 4 (row 2), the model finds each cell authentic but misses the global inconsistency.
- **Labelled grids outperform imagined grids:** Aligning with the observation in [33], overlaying grids and numbers on images significantly improved detection compared to asking the model to imagine the regions.
- **Holistic prompts capture global cues:** Holistic prompts are best for scene-wide manipulations (e.g., lighting, context), but struggle with localized edits. Example: in Row 3 (DeepFake) and Row 5 (Coverage) the model fails to detect forgery.
- **Factor-label correlation aids interpretability:** Likert scores for each of the factors help deduce the cues that matter for each dataset. We find that Human/Object Realism and Anomaly/ Artifacts are most important for AIGC Editing and DeepFakes. Edges/Boundaries and Reflections are key to detecting Photoshop Forgeries.
- **ROC curve using Likert scores:** The AUC score is less affected by imbalanced data and is helpful for threshold agnostic comparison between MLLMs. Figure 3 plots the ROC curves using the Likert scores as explained in section 2.3. They show quantified differences in various MLLM's classification. performance

4. Conclusion

We introduced REVEAL, a prompt-driven framework for image forgery detection with vision-language models. By leveraging holistic and region-wise evaluation, REVEAL enables interpretable, zero-shot reasoning across diverse forgery types. Our experiments show that structured prompts substantially improve both detection accuracy and explanation quality over baseline prompts. As generative models advance, prompt-based methods offer a scalable alternative to fine-tuning, adapting quickly to new manipulation techniques. Future work will integrate REVEAL with segmentation models to further enhance localization and region-specific analysis.

References

[1] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis, 2019. 1

[2] You-Ming Chang, Chen Yeh, Wei-Chen Chiu, and Ning Yu. Antifakeprompt: Prompt-tuned vision-language models are fake image detectors, 2024. 2

[3] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 1

[4] Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil Jain. On the detection of digital face manipulation, 2020. 2

[5] Oscar de Lima, Sean Franklin, Shreshtha Basu, Blake Karwoski, and Annet George. Deepfake detection using spatiotemporal convolutional networks, 2020. 1

[6] Jing Dong, Wei Wang, and Tieniu Tan. Casia image tampering detection evaluation database. In *2013 IEEE China Summit and International Conference on Signal and Information Processing*, pages 422–426, 2013. 2

[7] Gabriel Goh, James Betker, Li Jing, Aditya Ramesh, et al. Dall·e 3 understands significantly more nuance and detail than our previous systems, allowing you to easily translate your ideas into exceptionally accurate images., 2023. 1

[8] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. 1

[9] Luca Guarnera, Oliver Giudice, and Sebastiano Battiato. Deepfake detection by analyzing convolutional traces, 2020. 1

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 1

[11] Sonja Hohloch. Homoclinic floer homology via direct limits, 2024. 3

[12] Shan Jia, Reilin Lyu, Kangran Zhao, Yize Chen, Zhiyuan Yan, Yan Ju, Chuanbo Hu, Xin Li, Baoyuan Wu, and Siwei Lyu. Can chatgpt detect deepfakes? a study of using multimodal large language models for media forensics, 2024. 2

[13] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks, 2019. 1

[14] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022. 1

[15] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 1, 2

[16] Jiawei Liu, Fanrui Zhang, Jiaying Zhu, Esther Sun, Qiang Zhang, and Zheng-Jun Zha. Forgerygpt: Multimodal large language model for explainable image forgery detection and localization, 2025. 2

[17] Gary Marcus, Ernest Davis, and Scott Aaronson. A very preliminary analysis of dall-e 2, 2022. 1, 2

[18] Tian-Tsong Ng, Jessie Hsu, , and Shih-Fu Chang. Columbia image splicing detection evaluation, 2004. 2, 3

[19] Adam Novozamsky, Babak Mahdian, and Stanislav Saic. Imd2020: A large-scale annotated dataset tailored for detecting manipulated images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision workshops*, pages 71–80, 2020. 2

[20] OpenAI. Gpt-4 technical report, 2024. 2

[21] OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Madry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schul-

man, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. Gpt-4o system card, 2024. 2

[22] Sundar Pichai, Demis Hassabis, Koray Kavukcuoglu, et al. Introducing gemini 2.0: our new ai model for the agentic era, 2024. 2

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 1

[24] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 1, 2

[25] Zeyang Sha, Zheng Li, Ning Yu, and Yang Zhang. De-fake: Detection and attribution of fake images generated by text-to-image generation models, 2023. 2

[26] Rui Shao, Tianxing Wu, and Ziwei Liu. Detecting and recovering sequential deepfake manipulation, 2022. 2

[27] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Frequency-aware deepfake detection: Improving generalizability through frequency space learning, 2024. 1, 2

[28] Chris Vallance. "art is dead dude" - the rise of the ai artists stirs debate, 2022. 1

[29] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. Cnn-generated images are surprisingly easy to spot... for now, 2020. 1

[30] Bihan Wen, Ye Zhu, Ramanathan Subramanian, Tian-Tsong Ng, Xuanjing Shen, and Stefan Winkler. Coverage—a novel database for copy-move forgery detection. In *2016 IEEE international conference on image processing (ICIP)*, pages 161–165. IEEE, 2016. 4

[31] Kyle Wiggers. This startup is setting a dall-e 2-like ai free, consequences be damned, 2022. 1

[32] Zhipei Xu, Xuanyu Zhang, Runyi Li, Zecheng Tang, Qing Huang, and Jian Zhang. Fakeshield: Explainable image forgery detection and localization via multi-modal large language models, 2025. 2

[33] Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v, 2023. 3, 4

[34] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks, 2019. 1