AIti-FAct: Content-Based Image Distortion for Synthetic Dataset Generation

Ece Selin Böncü
Middle East Technical University
Ankara,Turkey
boncu@metu.edu.tr

Gözde Bozdagi Akar
Middle East Technical University
Ankara,Turkey
bozdagi@metu.edu.tr

Abstract

In this study, we propose a novel agent, AIti-FAct, that generates images based on self-written descriptions and autonomously produces an edited version by applying context-based distortions. AIti-FAct decomposes the end-to-end image editing process into simpler sub-tasks—description, generation, and distortion—enabling the use of multiple tools within the framework. Notably, the distortion task is implemented as a separate agent equipped with over 12 tools for applying content-aware image distortions.

By leveraging the reasoning capabilities of pre-trained large language models (LLMs) and their text-to-image generation abilities, this nested agent can both interpret prompts and determine realistic alterations based solely on image descriptions. With further development, the proposed framework aims to address the challenge of ground-truth data scarcity in image enhancement and restoration tasks.

1. Introduction

Due to the wide range of application areas for digital cameras, the representational quality of these devices has become an essential performance criterion. Either due to hardware limitations or momentary alterations in the scene, undesired artifacts may appear on the resulting image. Since the procedure is often irreversible and non-repeatable, image enhancement and restoration problems become ill-posed. The main objective of this study is to provide a method that enables the users to obtain artifact and artifact-free image pairs of the same scene in a fast, low-cost and effortless manner, and extend the operation to generate semantically diverse image artifact datasets. By involving three viewpoints of photography, *idea, scena, camera*[7], the proposed AIti-FAct framework mimics the procedure of image capture in order to fully incorporate the mechanism of how image artifacts are formed.

1.1. Overview

In this study, we propose AIti-FAct as an end-to-end visual data generation agent for context-based image distortion. Primarily, this framework is intended for automatic generation of synthetic image pairs illustrating the undistorted and content-aware distorted versions of the same scene setup. Owing to the complexity of the overall task, the framework adopts a layered solution. The complete structure of AIti-FAct comprises two nested ReACT[39] based agents both with utilities of multiple tool employment.

Our approach has following benefits: *i. interpretability:* as it is native to an agent, the progress in the execution is presented step by step.*ii. flexibility :* tool usage allows easy customization by simply including new artifact types in the distortion toolkit, so; it is quite straight-forward to increase the diversity. *iii. modularity:* nested structure of the model can also omit end-to-end image generation and apply only context-based distortion to a user-provided image. *iv. training-free:* the approach eliminates the necessity of training for new distortion effects. *v. affordability:* AIti-FAct omits capturing devices or the use of any other dataset as the images are generated synthetically. *rapidity:* by eliminating the need to set up a scene, an image pair is generated in measures of seconds.

Finally, the novelties of AIti-FAct can be listed as below:

- **Semantic Approach in Artifact Introduction:** AIti-FAct, to our knowledge, is the only method/agent to generate a variety of artifacts on images automatically through semantic reasoning.
- **End-to-End Generation:** The proposed model generates the distortion pairs along with the image caption, without the requirement of any kind of textual or visual input, by keeping each and every step synthetic, eliminating the need to capture any images.
- **Wide Range of Distortions:** Many distortions modelled and explored as the inverse problem in removal or restoration algorithms [12][25]; hence only few of them are addressed in a single work. Unlike the literature, this study covers up to 12 type of artifacts that may occur during image acquisition.

- **Novel Artifact Dataset Generation Method:** The base images being the ground truth, the generated image duos can be gathered as the expandable and dynamic AItImages Dataset to be used in training/fine-tuning/benchmarking of image restoration, reconstruction, inpainting algorithms.

1.2. Related Work

Due to the ever-growing interest on the use of LLMs, decision making methodology has been a hot topic. The pioneer of these group of works is Chain of Thoughts [33], where analogy is thought step-by-step to the controller LLM through single or few-shot learning. Further studies improve the model verifications on its reasoning through *reductio ad absurdum* and use of special symbols [14] for tasks involving complex spatial relations. Zhou et al. divide a larger though process into smaller chunks and proceed with summarization and examination on each, before the fine decision [44]. Making use of graphs [4], [40] and trees [38], [23] enables dynamic interplay and backtracking within the thought process.

Prominent autonomous agents for vision-language tasks are based on Visual Programming [11], ViperGPT [30]. Similar domain works include [21], [37] [22] & [24], but also extend the problem to a more multi-modal approach rather than focusing solely on image tasks.

2. Proposed Method

The task of generating a dataset of image pairs of base and with-artifact images automatically calls for extensive leveraging of pre-trained LLMs, including tasks dedicated to scene description, text2image generation, decision making based on semantic content, task sequencing etc. The framework proposed to accomplish this complex objective is an autonomous agent, AIti-FAct, which is composed of two nested agents namely, the Sequencer Agent and the Distorter Agent.

2.1. Sequencer Agent

The outer and hierarchically superior agent is the Sequencer Agent, a reasoning and decision-making agent in charge of task scheduling, equipped with a number of complex tools. Briefly, this is the part of the framework where the language model is exhorted to pick up the task using input or the previous action left off to complete a *Description-Generation-Distortion* sequence to obtain a pair of images with a descriptive caption. This very duo consists of images, one that is generated at the middle phase of an adequately flowing AIti-FAct sequence, the Generation Step, and is referred to as the base or artifact-free image, and the other, the distorted image, as obtained in the third and final phase, Distortion Step, along with the caption obtained at the very first Description Step.

In order to maintain both efficacy with the least number of input tokens and foolproof functionality to prevent operational failure, regarding the complexity of the task, the method proposed by Yao et al. [39] is adapted to our approach as the agent architecture. In terms of the human interaction required to AIti-FAct, the only input needed from the user is a simple prompt. Without any external guidance, AIti-FAct is capable of attaining a vast diversity of scenes and its entities, however; merely a name of an object or setting and/or a simple phrase is enough to merely steer the LM towards a specific type of photography and/or situation if required. Yet, it is still the language model that designs the layout of the image to be generated depending on this input or an arbitrary token. In addition, prompting with a number is essential to determine the size of the dataset by extending the operation as *De-G-Di-De-G-Di...De-G-Di*.

AIti-FAct is an autonomous agent that employs multiple self-defined tools to achieve the aforementioned subtasks for the scheduled stages of operation. This toolkit along with the flow architecture is illustrated in Figure 1

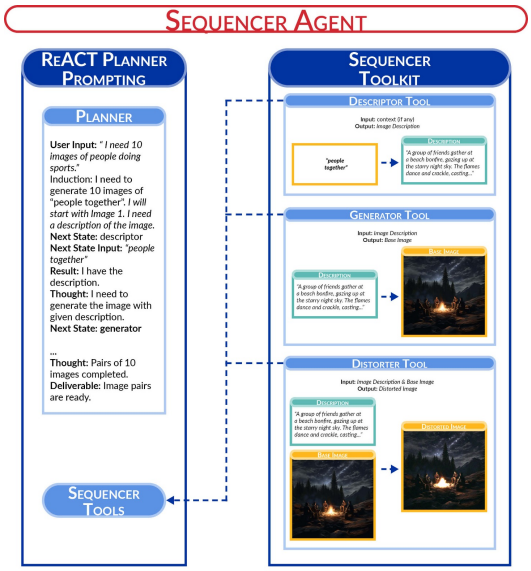

Figure 1. Overview of Sequencer Agent

2.1.1. Descriptor Tool

Descriptor is the tool associated with the Description step of an AIti-FAct operation which determines the entire content of the base image. Although this step is a language-only task, keeping the sequencing and description tasks united would result in performance drop due to memory interference; hence these two must be handled in different instances. Additionally, even if the language models are able to accomplish the task of of story telling[3][10][20], scene description in this task has further limitations. First of all, since an image is only perceived by sight, the inclusion of

the narrative of other senses is redundant. Also, pre-trained LLMs are limited by the number of output tokens, so are text2image (T2I) generators by finite memory. Experiments suggest that as an expression moves further in the entire input, the importance and visualization of the included semantic information diminishes significantly. Furthermore, certain T2I models produce better visuals with a pre-defined syntax [27]. The required format for pre-trained T2I method is model-dependent; therefore; a bare juxtaposition of all the desired details is very likely to give poor results. In short, the desired descriptor outputs are of optimal length, tailored into a specific syntax and contains a lexical choice that enwraps dense semantic information.

Initially, Descriptor Tool is implemented as a few-shot text-generator focused on visual descriptions of a scene, limited to 4-5 sentences, as illustrated in Figure 2. It utilizes GPT-3.5-turbo[6] to write 10 descriptions to be manually refined into a desired form, then fed back into the model to generate a bank of 100 example descriptions. At inference, the AIti-FAct Agent (Sequencer) randomly selects 6 descriptions to add to the simple prompt. AIti-FAct Descriptor Tool has a buffer memory that takes up to a number of previously generated descriptions and as the agent operates, it is updated in a First-In-First-Out fashion. This allows the model to preserve the description syntax and style, as well as maintaining scene content information. The prompt that generates the descriptions is constructed in such a way that diversity in the layout, setting, objects and entities is encouraged. Eventually, using few-shot learning grants the model what is already generated and lets it envision novel scenes.

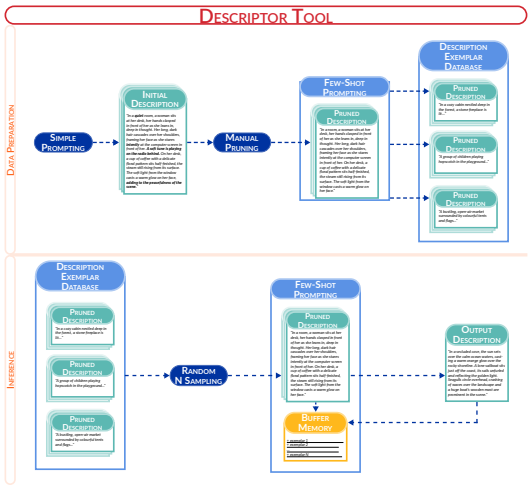

Figure 2. Descriptor Tool: The data preparation and flow of operation during interference

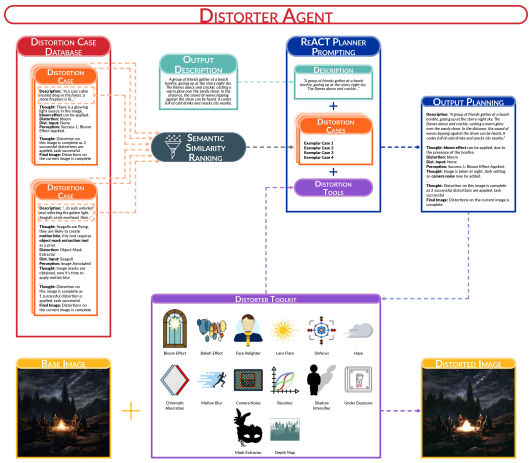

Figure 3. Overview of the AIti-FAct Distorter Agent

2.1.2. Generator Tool & Distortion Tool

Generator Tool encorporates a pre-trained text-to-image model which converts the scene description into an image. Selected model for our framework is developed by Midjourney [2], achieving high-quality results [41]. On the other, The Distorter Tool, another autonomous LLM agent, is described in 2.2.

2.2. Distorter Agent

AIti-Fact Distorter Agent is the inner LLM agent of the proposed framework where content-aware image distortion is applied on the Generator Tool output. Even though it is a tool with respect to the higher order Sequencer Agent, is indeed another another ReACT-based[39] agent with a toolkit and few-shot similarity-based example retrieval mechanism, as seen in Figure 3.

2.2.1. Context-Aware Distortion

Context-aware distortion is our proposed approach to decide on the type of imaging artifact that may appear on the captured scenes with respect to semantic information. Distorter Agent is aims to recreate logically convincing artifacts that naturally occur in the real-life counterparts of the image scenes. Image artifacts arise from a variety of factors inherent to the scene. Primarily, the global setting is affected by the lighting, the weather and the presence of a light sources, causing intense shadows, non-homogeneous illumination in any form or at extremes whiteout blobs. Additionally, the presence of multiple objects may result in lack of a single concentrate, and eventually inferior focus or none at all. Moreover, since the objects in motion cannot be contained in their own boundaries, creating isthmuses and protrusions within their neighbourhood. Lastly, material characteristics such as reflective and refractive properties of the objects are likely to cause aberrations and flare

"A kitchen with a wooden table in the foreground, featuring yellow-orange fruits, a glass vase with green leafy branches, and stacked bowls and plates. Two wooden chairs with woven seats are positioned at the table. Open wooden shelves on the wall hold various dishes, jars. A large window in the background allows natural light to illuminate the space, with green foliage visible outside. Potted plants are placed on the windowsill and counter, and a ceiling-hung light fixture is centered above the table."

Figure 4. Example output of AIti-FAct operation

formations, lowering the quality of the output.

To accomplish the challenging objective of mimicking the faulty image capture, the use of reasoning on scene description is required. In other words, LLM underlying the agent is expected to decide on likely artifact(s) by making analyses on the entire scene both globally and locally, so that the picked distortions are applied to the image through the use of tools. As before, we again adopt ReAct model [39] and modify it to observe image entities and conditions and force the model to use the matching effect from the Distorter Toolkit, which contains *bloom effect[16], bokeh effect[34], camera noise[9], chromatic aberration[8],[29], defocus[34], face relighting [13], haze generator[43], lens flare[26], motion blur[5], shadow intensifier[35], recolouring[17], under exposure effect[1]* tools along with two helper tools for depth map[36] & object mask extraction [19][42][18].

2.2.2. Case Database & Few-Shot Example Retrieval

The toolkit associated with the AIti-FAct Distorter Agent offers a wide variety of effect generators to simulate the real-life imaging artifacts. Presumably, the more the number of possibilities raises, the more complicated the analysis becomes with finite examples. With regard to the complexity of the analyses, the problem of hallucination is significant with some LLMs causing the aqent o terminate due to tool failure even with the reasoning mechanism. Hence; to ensure a guaranteed performance, few-shot learning is also utilized in our approach. Unfortunately, each pre-trained LLM has limitations on the input tokens that they are capable of processing in a single run and concatenation of too many exemplars is not an efficient and robust way of prompting. So to tackle this problem, a set of 200 example cases of proper end-to-end operation is prepared manually, to form a Distortion Cases Database and a strategic retrieval algorithm is utilized. 4 top-ranking exemplars are selected at runtime, based on the vector-embedding similarities of the exemplar descriptions to the input description. In prior research, resembling techniques are applied for retrival, but with the objective of eliminating irrelevant and under-performing tools[32], [45]. Unlike those works, the tools of the entire toolkit are maintained but instead at inference, whilst the agent steers the language model towards a correct viewpoint and the analysis converges to a more accurate form. It is to be remarked here that, employing semantic-similarity in reasoning is valid since comparable scenes are predisposed to have artifacts akin to each other.

Model	Token	Divs.	Non-Vis.	Story
GPT-4-Turbo-0[28]	86.21	0.38	20	12
GPT-4-Turbo-1	79.6	0.42	14	9
GPT-4-Turbo-3	75.15	0.43	13	8
GPT-4-Turbo-6	73.82	**0.45**	**11**	**6**
GPT-4-Turbo-9	**73.65**	**0.45**	**11**	**6**
Mixtral-8x7b-Instr.-v0.1-0[15]	84.05	0.35	18	11
Mixtral-8x7b-Instr.-v0.1-1	81.84	0.35	16	10
Mixtral-8x7b-Instr.-v0.1-3	78.58	.0.35	16	10
Mixtral-8x7b-Instr.-v0.1-6	77.08	0.37	15	10
Mixtral-8x7b-Instr.-v0.1-9	76.34	0.39	14	9
GPT-3-5-Turbo-0613-0[6]	85.38	0.3	37	23
GPT-3-5-Turbo-0613-1	80.03	0.32	29	20
GPT-3-5-Turbo-0613-3	79.72	0.34	26	18
GPT-3-5-Turbo-0613-6	78.32	0.35	24	15
GPT-3-5-Turbo-0613-9	78.67	0.38	22	13
Lllama-2-70B-chat-0[31]	92.8	0.16	51	46
Lllama-2-70B-chat-1	90.75	0.25	42	35
Lllama-2-70B-chat-3	88.28	0.28	35	24
Lllama-2-70B-chat-6	85.64	0.32	27	15
Lllama-2-70B-chat-9	85.43	0.32	27	14

Table 1. Description Generation Statistics of LLMs

3. Experimental Results

In this section, we present the numerical performance of various LLMs with different numbers of few-shot exemplars in generating desired scene descriptions in Table 1. The best performing model is GPT-4-Turbo-6, on the average of 100 generations, achieving minimum number of non-visual elements and story telling. 9Shot counterpart is also on par with these results, but utilizes more memory compared to the former. GPT-4-Turbo-6 also achieves maximum diversity in terms of the L2 distance of embeddings. Also, an example visual results is provided in Figure 4.

4. Conclusion

This study is preliminary work that provides an autonomous, efficient, and low-cost approach to address the scarcity of image artifact datasets with ground truth. Initial results are promising and motivate further work on tailoring the model for full dataset generation. Future work will include numerical results on failure cases across different LLMs and architectures, presented as an ablation study. Additionally, subjective test results regarding output quality and decision appropriateness will be provided. Finally, we aim to demonstrate the effectiveness of the generated AIti-FAct dataset in improving the performance of image enhancement and editing models on benchmark datasets.

References

[1] Mahmoud Afifi, Konstantinos G Derpanis, Bjorn Ommer, and Michael S Brown. Learning multi-scale photo exposure correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9157–9167, 2021. 4

[2] Midjourney AI. Midjourney, 2023. 3

[3] Prithviraj Ammanabrolu, Wesley Cheung, Dan Tu, William Broniec, and Mark O. Riedl. Bringing stories alive: Generating interactive fiction worlds, 2020. 2

[4] Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. Graph of thoughts: Solving elaborate problems with large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17682–17690, 2024. 2

[5] Tim Brooks and Jonathan T. Barron. Learning to synthesize motion blur. *CoRR*, abs/1811.11745, 2018. 4

[6] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. 3, 4

[7] Ece Selin Böncü. *Context-based visual data generation for image enhancement and automatic colour calibration*. Phd thesis, Middle East Technical University, Ankara, Turkey, 2024. 1

[8] Thomas Eboli. Fast Chromatic Aberration Correction with 1D Filters. *Image Processing On Line*, 13:198–214, 2023. https://doi.org/10.5201/ipol.2023.443. 4

[9] M.D. Grossberg and S.K. Nayar. Modeling the Space of Camera Response Functions. pages 1272–1282, 2004. 4

[10] Jian Guan, Fei Huang, Zhihao Zhao, Xiaoyan Zhu, and Minlie Huang. A knowledge-enhanced pretraining model for commonsense story generation. *Transactions of the Association for Computational Linguistics*, 8:93–108, 2020. 2

[11] Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training, 2022. 2

[12] Bernardo Henz, Eduardo S. L. Gastal, and Manuel M. Oliveira. Synthesizing camera noise using generative adversarial networks. *IEEE Transactions on Visualization and Computer Graphics*, 2020. 1

[13] Andrew Hou, Ze Zhang, Michel Sarkis, Ning Bi, Yiying Tong, and Xiaoming Liu. Towards high fidelity face relighting with realistic shadows. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 4

[14] Hanxu Hu, Hongyuan Lu, Huajian Zhang, Yun-Ze Song, Wai Lam, and Yue Zhang. Chain-of-symbol prompting elicits planning in large langauge models, 2023. 2

[15] Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of experts, 2024. 4

[16] CHARILAOS KALOGIROU. How to do good bloom for hdr rendering. https://kalogirou.net/2006/05/20/how-to-do-good-bloom-for-hdr-rendering, May 30th, 2006. Accessed: Nov, 2023. 4

[17] Akiomi Kamakura. https://github.com/akiomik/pilgram. 4

[18] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023. 4

[19] Liunian Harold Li*, Pengchuan Zhang*, Haotian Zhang*, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *CVPR*, 2022. 4

[20] Li Lin, Yixin Cao, Lifu Huang, Shu'ang Li, Xuming Hu, Lijie Wen, and Jianmin Wang. What makes the story forward? inferring commonsense explanations as prompts for future event generation, 2022. 2

[21] Shilong Liu, Hao Cheng, Haotian Liu, Hao Zhang, Feng Li, Tianhe Ren, Xueyan Zou, Jianwei Yang, Hang Su, Jun Zhu, Lei Zhang, Jianfeng Gao, and Chunyuan Li. Llava-plus: Learning to use tools for creating multimodal agents, 2023. 2

[22] Zhaoyang Liu, Zeqiang Lai, Zhangwei Gao, Erfei Cui, Ziheng Li, Xizhou Zhu, Lewei Lu, Qifeng Chen, Yu Qiao, Jifeng Dai, and Wenhai Wang. Controlllm: Augment language models with tools by searching on graphs, 2023. 2

[23] Jieyi Long. Large language model guided tree-of-thought, 2023. 2

[24] Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. Chameleon: Plug-and-play compositional reasoning with large language models, 2023. 2

[25] Ali Maleky, Shayan Kousha, M. S. Brown, and Marcus A. Brubaker. Noise2noiseflow: Realistic camera noise modeling without clean images. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17611–17620, 2022. 1

[26] Anderson Mancini. https://github.com/ektogamat/R3F-Ultimate-Lens-Flare. 4

[27] Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work?, 2022. 3

[28] OpenAI. Gpt-4 technical report, 2023. 4

[29] Yoonsik Park. https://github.com/yoonsikp/kromo. 4

[30] Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning, 2023. 2

[31] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. 4

[32] Liang Wang, Nan Yang, and Furu Wei. Learning to retrieve in-context examples for large language models, 2024. 4

[33] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. 2

[34] John WHITE and Colin BARRÉ-BRISEBOIS. More performance! five rendering ideas from battlefield 3 and need for speed: The run, advances in real-time rendering in games. 4

[35] Xiaowo Xu, Xiaoling Zhang, Tianwen Zhang, Zhenyu Yang, Jun Shi, and Xu Zhan. Shadow-background-noise 3d spatial decomposition using sparse low-rank gaussian properties for video-sar moving target shadow enhancement. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022. 4

[36] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data, 2024. 4

[37] Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Mm-react: Prompting chatgpt for multimodal reasoning and action, 2023. 2

[38] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models, 2023. 2

[39] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models, 2023. 1, 2, 3, 4

[40] Yao Yao, Zuchao Li, and Hai Zhao. Beyond chain-of-thought, effective graph-of-thought reasoning in language models, 2024. 2

[41] Chenshuang Zhang, Chaoning Zhang, Mengchun Zhang, and In So Kweon. Text-to-image diffusion models in generative ai: A survey, 2023. 3

[42] Haotian* Zhang, Pengchuan* Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Harold Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. Glipv2: Unifying localization and vision-language understanding. *arXiv preprint arXiv:2206.05836*, 2022. 4

[43] Ning Zhang, Lin Zhang, and Zaixi Cheng. Towards simulating foggy and hazy images and evaluating their authenticity. In *International Conference on Neural Information Processing*, pages 405–415. Springer, 2017. 4

[44] Yucheng Zhou, Xiubo Geng, Tao Shen, Chongyang Tao, Guodong Long, Jian-Guang Lou, and Jianbing Shen. Thread of thought unraveling chaotic contexts, 2023. 2

[45] Yuchen Zhuang, Yue Yu, Kuan Wang, Haotian Sun, and Chao Zhang. Toolqa: A dataset for llm question answering with external tools. *Advances in Neural Information Processing Systems*, 36, 2024. 4