HIJACKING CONTEXT IN LARGE MULTI-MODAL MODELS

Joonhyun Jeong[1,2]
NAVER Cloud, ImageVision[1]
Korea Advanced Institute of Science and Technology (KAIST)[2]
{joonhyun.jeong}@navercorp.com

ABSTRACT

Recently, Large Multi-modal Models (LMMs) have demonstrated their ability to understand the visual contents of images given the instructions regarding the images. Built upon the Large Language Models (LLMs), LMMs also inherit their abilities and characteristics such as in-context learning where a coherent sequence of images and texts are given as the input prompt. However, we identify a new limitation of off-the-shelf LMMs where a small fraction of incoherent images or text descriptions mislead LMMs to only generate biased output about the hijacked context, not the originally intended context. To address this, we investigate whether replacing the hijacked visual and textual contexts with the correlated ones via GPT-4V and text-to-image models can help yield coherent responses.

1 INTRODUCTION

Large Language Models (LLMs) pre-trained with huge amount of text corpus have demonstrated remarkable generalization on various language tasks such as human-like dialogues Google (2023); OpenAI (2023a;b), reasoning-tasks Wei et al. (2021); Lewkowycz et al. (2022); Yao et al. (2022), and few-shot in-context learning Min et al. (2021; 2022) where a few demonstration examples are given as a context for answering the test question. By bootstrapping LLMs with several add-on layers, Large Multi-modal Models (LMMs) OpenAI (2023b); Koh et al. (2023); Liu et al. (2023a); Zhu et al. (2023) were enabled to understand the visual contents and answer the corresponding textual instruction. LMMs also inherited the in-context learning capability of LLM, where a coherent sequence of images and textual information are given as a context for the input prompt Koh et al. (2023); OpenAI (2023b).

In this paper, we identify a new limitation of these off-the-shelf LMMs, *hijacking contexts*, where a small fraction of incoherent images or text descriptions mislead them to only generate biased output about the incoherent contexts. In Figure 1, we provide the qualitative example of context hijacking on visual story telling (VIST) dataset Ferraro et al. (2016) with FROMAGe Koh et al. (2023) as LMM. Given all the former sequence of images and their corresponding captions, the LMM is asked to provide a caption or description of the final query image. Normally, when all the visual and textual information is coherent and consistently aligned, LMM successfully provides a coherent response under consideration in the given contexts. However, when only a single image-text pair that contains an irrelevant subject is appended to the context, LMM falls into the hijacked context and loses coherency concerning the bunch of formerly given original contexts. In real-world applications and scenarios of LMMs, there is no guarantee for the inexistence of such noises and irrelevant contexts. For reliable usage of LMM, it should be robust to such distribution shift and stick to the majority of context without confusion on the minority of irrelevant contexts.

To address this, we investigate whether modifying the hijacked image and text contexts with the correlated ones can help yield better coherent responses. Firstly, we instructed GPT-4 to identify any irrelevant sentences and replace them with appropriate alternative sentences that convey coherency given in the contexts. Subsequently, we prompted text-to-image diffusion models [1] to generate a coherent image corresponding to the newly replaced sentence under consideration for the original

[1]We employed DALLE-3 Betker et al. (2023) for its detailed sensitiveness to the input prompts

Figure 1: **Hijacking context confuses LMMs to only generate responses with incoherent contents.** (top): Given a sequence of visual and textual story plot 1∼4, LMM reasonably outputs a coherent caption for the final image. (bottom): When a single pair of incoherent image and caption (highlighted with red) is appended to the context, LMM only tells about the hijacked context ($i.e.$, *football game*), disregarding all the aforementioned context of visual story plots ($i.e.$, *emergency situation*).

sequence of sentences and images. As a result, a new alternative image and caption that are coherent with the original story plot are interleaved into inputs for LMMs. Notably, we observed that LMMs still exhibit hijacked responses, due to two possible reasons: 1) the generated captions are still confusing which seem to be the combination and mixture between the original and hijacked context. 2) the generated images are not realistic and do not have the consistent texture or styles with the original image sequences.

We showcase our problem statement and aforementioned methods and investigations with qualitative examples one by one[2]. In summary, our key contributions are outlined as follows:

- Identifying a new problem of off-the-shelf LMMs, context hijacking, that can deteriorate reliable usage of LMMs.
- Propose a straightforward remedy for suppressing hijacked contexts via the enormous foundation models.
- Investigate an open question of whether replacing hijacked contexts can help to produce coherent responses, promoting a new future research frontier.

2 Related Works

Large language models. Recently, large language models (LLMs) have gained considerable versatility based on enlarging the scale of pre-training datasets Radford et al. (2019); Brown et al. (2020); OpenAI (2023b). Along with the development of human-aligned feedbacks Christiano et al. (2017), LLMs have further been well-suited for conforming to the instruction rather than predicting simple meaningless next tokens and words. These LLMs have also demonstrated versatility on various language domain tasks including in-context learning Brown et al. (2020); Chan et al. (2022) and long-term dialogues Jozefowicz et al. (2016); Dai et al. (2019); Liu et al. (2023b).

Large multi-modal models. Despite the compelling performance of LLMs on language tasks, LLMs could not understand the visual contexts in an image. To bootstrap LLMs with vision modality, several works Liu et al. (2023a); Zhu et al. (2023); Koh et al. (2023) employed an additional pre-trained visual encoder and leveraged a few linear layers to adapt the visual tokens to be understandable by the frozen LLMs. FROMAGe Koh et al. (2023) further demonstrated the in-context

[2]We tested with FROMAGe Koh et al. (2023) as the baseline of LMM.

learning capability of LMMs where consecutive sequences of image-text pairs are given as a context to respond to the query image. In this paper, we focus on the vulnerability of LMMs to incoherent and irrelevant context information.

Hijacking context in language and vision tasks. In the language domain, several works empirically observed that LLMs are easily hijacked by irrelevant context information Shi et al. (2023), small perturbations of characters ($e.g.$, from "film" to "fi1m") Wang et al. (2023), and adversarial addition of suffix words Qiang et al. (2023). In the vision domain, there have been several approaches that investigated the adversarial robustness of LMMs Schlarmann & Hein (2023); Qi et al. (2023) where a small adversarial tweak of pixel values significantly misled LMMs to output violent content. To the best of our knowledge, our work is the first to investigate hijacking contexts in an in-context learning scenario where the consecutive sequence of images and texts along with irrelevant ones are given as the context to provide answers about the query image.

3 Context-Hijacking in Large Multi-modal Models

Preliminary: Large Language Models. An autoregressive LLM p_θ is pre-trained on a large corpus of text tokens with the next-token prediction objective. Formally, given a text of image caption y and its sequence of tokens ($\boldsymbol{s} = s_1, ..., s_{t-1}$) separated by a BPE tokenizer Sennrich et al. (2015), the model is trained to maximize log-likelihood for the next token s_t:

$$\log p_\theta(y) = \sum_{t=1}^{t} \log p_\theta(s_t|\boldsymbol{s})$$

Preliminary: Large Multi-modal Models. To bootstrap LLMs with vision modality, visual embeddings $V_\phi(x)$ from input image x are firstly extracted by a visual encoder model V_ϕ pre-trained on vision tasks. These visual embeddings are mapped into the same dimension of text embedding space by applying several linear layers W. These projected visual embeddings, $V_\phi(x)^T W$, are prepended to the text tokens s and trained to be understandable by the frozen LLMs, letting LLM interpret the image and answer the corresponding captions or instructions. Formally, the image-captioning objective given an image x and a caption y is defined as follows:

$$l(x, y) = \sum_{t=1}^{t} \log p_\theta(s_t|V_\phi(x)^T W, \boldsymbol{s})$$

For in-context learning scenarios where multiple images and corresponding text descriptions are given, FROMAGe Koh et al. (2023) is trained with a random concatenation strategy where multiple image-text pairs are piled for attending and responding more explicitly to each image. Hence, FROMAGe exhibited superior capability for answering the query image x_n given former visual contexts $(x_1, \ldots, x_{n-1})$ and textual contexts $(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_{n-1})$:

$$\sum_{n=1}^{n} \log p_\theta(s_n|V_\phi(x_1)^T W, \boldsymbol{s}_1, \ldots, V_\phi(x_{n-1})^T W, \boldsymbol{s}_{n-1})$$

However, when the contexts in the former images $(x_1, \ldots, x_{n-1})$ or texts $(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_{n-1})$ are composing an incoherent visual story due to the existence of some irrelevant contexts $(x^*, \boldsymbol{s}^*)$, the output response tends to be biased towards the irrelevant context, as shown in Figure 1.

4 Discussions

Location of hijacked context. We investigate the effect of the sequential location of hijacked context on the coherency of the response from LMMs. In Figure 2, when the hijacking context is interleaved in the former part, the final response still tends to adhere to the original context, "*family*". In contrast, as the hijacking context is interleaved closer to the query image, the response tends to be incoherent to the former contexts and biased towards the hijacked context, "*football*".

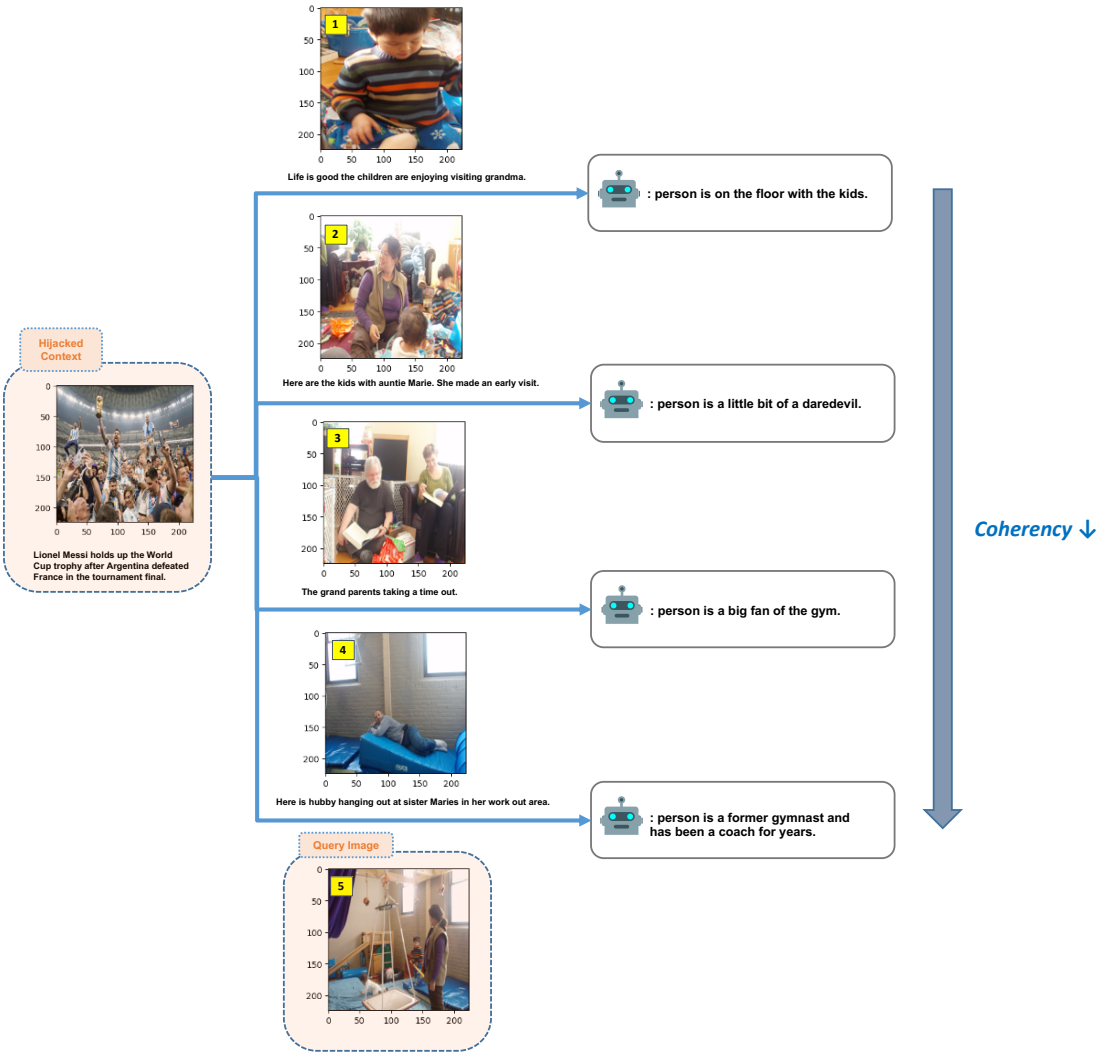

Figure 2: **Effect of location for the hijacking context.** We ablated to insert the hijacking context in between the sequences 1∼5 and visualize the LMM's response with regard to the query image.

Robustness of GPT-4V. We further tested GPT-4V whether it is robust to such context-hijacking issues. In Figure 2, we inserted hijacked context between sequences 4 and 5 and instructed GPT-4V to answer the caption corresponding to the query image. Notably, GPT-4V's response was not biased towards the hijacked context, rather sticking to the original context:

Playtime continues with Auntie Marie, as she supervises the little adventurers on their indoor swing set, making memories to cherish.

This result is indicative of the robustness of GPT-4V toward the distribution shift, shedding light on a future direction of explicitly distilling the robust knowledge from GPT-4V.

Reforming contexts via large foundation models. Inspired by robustenss of GPT-4V, we investigate whether such large foundation models can identify and reform the irrelevant visual and textual contexts into coherent ones, for better coherent response. As in Figure 4, we first instructed GPT-4 to provide appropriate sentences that convey coherency given in the contexts. Subsequently, we instructed DALLE-3 to generate a synthetic image that corresponds to the newly reformed text, with the consideration of underlying visual and textual context information.

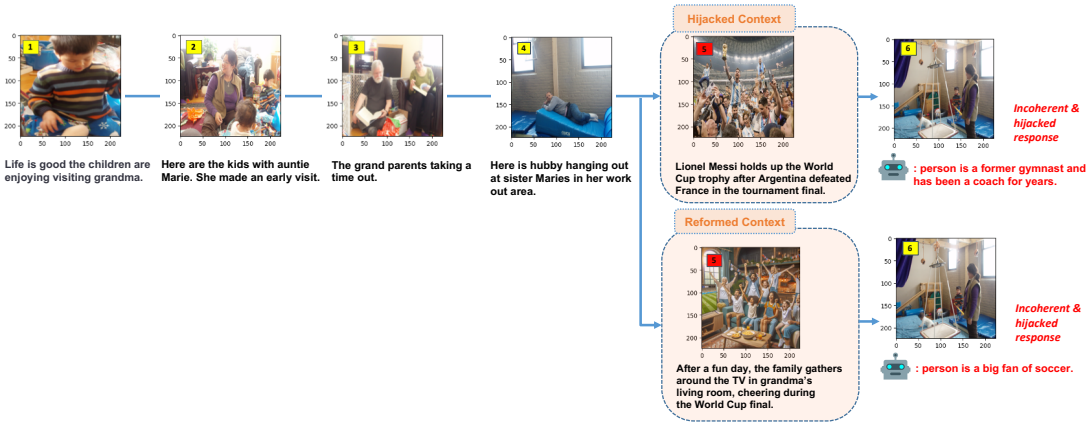

Figure 3: **Effect of reformed context on the LMM response.** We replaced the hijacked context with the one reformed by GPT-4 and DALLE-3.

Limited impact of reforming hijacked contexts. In Figure 3, we observed that our reforming strategy still does not ensure LMMs give coherent responses. We conjecture that the reformed captions still contain the hijacked context ("*football*") while also adhering to the original context ("*family*"). The generated image also correspondingly contains both visual contexts of "family" and "football", confusing LMMs to produce biased responses. We conjecture that the style and texture of the reformed image differ from the original images, which might also induce LMMs to be confused.

5 Limitations

The limitations of this work are that we provide only qualitative example results on the VIST dataset since quantitative evaluation is too labor-demanding; the generated texts from LMMs require a human to check if they are aligned with the image and coherent with the story plot. Therefore, we leave thorough and rigorous quantitative evaluation using various LMMs as future work. Also, our proposed scheme is a straightforward solution yet fragile when the irrelevant contexts become the majority within the entire context where even GPT-4 would also be confused. Therefore, developing a more secure way of filtering or replacing a bunch of the hijacked contexts is a promising future research direction.

6 Conclusion

In conclusion, we newly identified the context-hijacking problem of existing LMM where the irrelevant visual and textual contexts induce the LMM to generate a biased and incoherent response, disregarding the original context. To mitigate this issue, we introduced a simple method to preemptively remove such hijacked contexts and further reform the hijacked contexts into coherent ones, which promotes future research direction to generate more coherent visual and textual information.

References

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2023.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Stephanie CY Chan, Adam Santoro, Andrew K Lampinen, Jane X Wang, Aaditya Singh, Pierre H Richemond, Jay McClelland, and Felix Hill. Data distributional properties drive emergent few-shot learning in transformers. *arXiv preprint arXiv:2205.05055*, 2022.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*, 2019.

Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C Lawrence Zitnick, et al. Visual storytelling. *arXiv preprint arXiv:1604.03968*, 2016.

Google. Bard - chat based ai tool from google, 2023. URL `https://bard.google.com`.

Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410*, 2016.

Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. Grounding language models to images for multimodal generation. *arXiv preprint arXiv:2301.13823*, 2023.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. Solving quantitative reasoning problems with language models, 2022. *URL https://arxiv. org/abs/2206.14858*, 2022.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023a.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*, 2023b.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. Metaicl: Learning to learn in context. *arXiv preprint arXiv:2110.15943*, 2021.

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? *arXiv preprint arXiv:2202.12837*, 2022.

OpenAI. Chatgpt can now see, hear, and speak, March 15 2023a. URL `https://openai.com/blog/chatgpt-can-now-see-hear-and-speak`.

OpenAI. Gpt-4 technical report. Technical report, February 2023b.

Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Mengdi Wang, and Prateek Mittal. Visual adversarial examples jailbreak aligned large language models. In *The Second Workshop on New Frontiers in Adversarial Machine Learning*, 2023.

Yao Qiang, Xiangyu Zhou, and Dongxiao Zhu. Hijacking large language models via adversarial in-context learning. *arXiv preprint arXiv:2311.09948*, 2023.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Christian Schlarmann and Matthias Hein. On the adversarial robustness of multi-modal foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3677–3685, 2023.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pp. 31210–31227. PMLR, 2023.

Jiongxiao Wang, Zichen Liu, Keun Hee Park, Muhao Chen, and Chaowei Xiao. Adversarial demonstration attacks on large language models. *arXiv preprint arXiv:2305.14950*, 2023.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

A APPENDIX

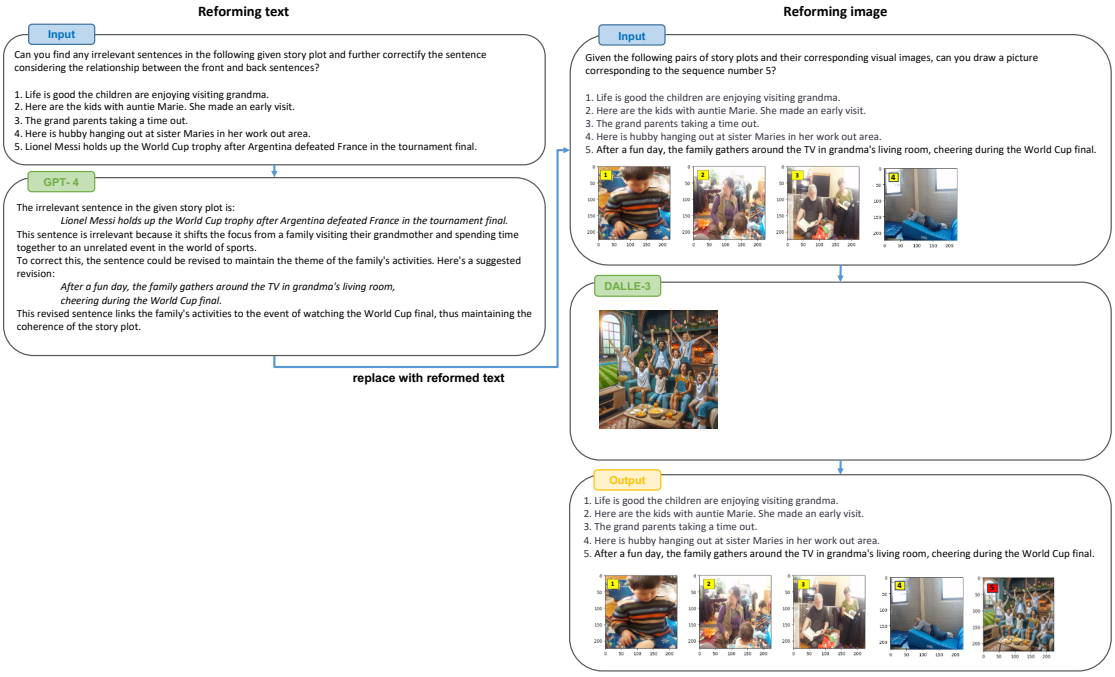

Figure 4: **Reforming hijacked text and images via GPT-4 and DALLE-3.** Given multiple image-text pairs of a visual story plot from VIST dataset Ferraro et al. (2016), GPT-4 is instructed to replace any irrelevant images or text descriptions with coherent ones. Subsequently, DALLE-3 is instructed to generate an image corresponding to the newly reformed text, under consideration of the underlying context.