HUMAN-CENTRIC VISUAL MANIPULATION UNDERSTANDING VIA LLM

by

LI GUANZHEN
(B.Eng., University of International Business and Economics)

A THESIS SUBMITTED FOR THE DEGREE OF

MASTER OF COMPUTING

in

ARTIFICIAL INTELLIGENCE

in the

GRADUATE DIVISION

of the

NATIONAL UNIVERSITY OF SINGAPORE

2024

Supervisor:
Associate Professor Min-Yen Kan

Examiners:
Associate Professor Min-Yen Kan
Dr. Christian Von Der Weth

Declaration

I hereby declare that this thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

Li Guanzhen

9 November 2023

To everyone who helped me

Acknowledgments

Firstly, I would like to express my sincere gratitude to Prof. Min-Yen Kan for warmly welcoming me as a member of the Web Information Retrieval and Natural Language Processing Group (WING) and patiently instructing me throughout this project. I also want to extend my honest appreciation for the invaluable guidance and assistance provided by my PhD student mentor, Xie Yuxi. Working alongside so many outstanding colleagues in the lab has not only been a privilege for me but has also contributed significantly to my well-rounded skill development.

Besides, I desire to convey my sincere thanks to Dr. Christian Von Der Weth for being my dissertation examiner. His insightful lectures provided me with a wealth of prior knowledge in NLP, and his constructive feedback played a crucial part in further revising my dissertation.

Moreover, I'm grateful to Prof. Lu Dongyuan from UIBE, who is also an alumnus of the WING, for supporting me in joining this fabulous group. Additionally, I also appreciate the powerful computing resources provided by the NSCC Workshop.

Last but not least, I am always very grateful to my parents, Li Pu and Qiao Chang, and my entire family for their unconditional support throughout my studies at NUS. I believe that the lessons learned here will undoubtedly lay a solid foundation for my future endeavours.

Contents

Abstract

Human-centric Visual Manipulation Understanding via LLM

by

Li Guanzhen

Master of Computing in Artificial Intelligence

National University of Singapore

Manipulated human-centric images have become widespread across social media owing to the prevalence of image editing tools. To assist in regulating malicious visual manipulations, we introduce a Large Language Model driven Visual Manipulation Understanding (LLM-VMU) pipeline, detecting manipulations between images based on their textual captions instead of directly relying on visual information. Besides, we propose a Visual Information Checking module to mitigate the hallucination and utilize a tailored taxonomy to guide the workflow by considering edits' indications. Our approach significantly surpasses the baseline on the EMU benchmark by 2.2 in CIDEr and exhibits robustness across various LLM and MLLM backbones. This foretells the pipeline's adaptability and promising effectiveness with potentially more advanced models in the future. Additionally, further analysis reveals that our pipeline holds the potential to alleviate errors besides hallucination, including key component missing and reference misunderstanding, suggesting its broader applicability in detecting various forms of visual manipulations.

Keywords: Visual Manipulation, Human-centric Image, Large Language Model, Multi-modal Large Language Model, Hallucination, Editing Intention

List of Tables

List of Figures

Chapter 1

Introduction

Image forgery has become increasingly ubiquitous and easily accessible due to the recent progress in image-editing techniques [9, 24]. Human-centric images, primarily featuring individuals such as celebrities or politicians [37], constitute a substantial portion of the editing cases, particularly within the realm of social media. Within them, improper manipulations of human-centric images can lead to profound effects on both individuals and society, such as personal reputation sabotage and social division exacerbation [10].

To counteract the detrimental effects of malicious image manipulations, numerous studies in computer vision and media forensics focused on identifying edited images [42, 38, 49, 37]. However, as not all manipulations are inherently harmful (shown in Figure 1.1), it is essential to further understand how an image is edited. For example, adjusting photos' overall brightness is a common practice for photographers, which can absolutely be considered proper. By contrast, replacing the background of a politician from a conference setting with a nightclub scene can damage the personal reputation and raise ethical concerns. Hence, in this work, we go beyond merely identifying forged images, focusing on "what" specific edits have been made as a further step to assist in scrutinizing malicious visual manipulations.

To detect and interpret the manipulations, we introduce a Large Language Model driven Visual Manipulation Understanding (LLM-VMU) pipeline. Given an image pair (*i.e.*, source image and edited image), LLM-VMU leverages the reasoning ability of LLMs to infer manipulations based on the images' textual captions. Different from previous works which directly rely on visual information, we distill and reason about the complex information within human-centric images via captioning [40,

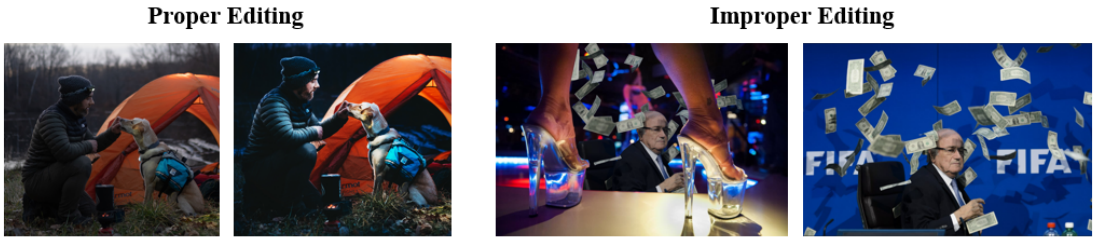

Figure 1.1: A comparison between proper editing and improper editing. It demonstrates that not all manipulations are harmful. It depends on "what" manipulation has been conducted.

12]. Human-centric images closely resemble real social scenes and often contain a multitude of elements (*i.e.*, intricate context). Consequently, to conduct indications on the subjects, manipulations within these images typically alter a substantial portion of the image (*i.e.*, removing and introducing plentiful elements). For example, in Figure 3.1, to indicate the shift in the main subject's role, the entire office setting, along with all the decorations and office supplies, has been replaced with an old wall. This imposes a stringent demand on visual models to discern relationships between the various components across two images. Hence, by representing images as captions, we can directly understand the entire scene's story and avoid confusion from numerous trivial details.

To better deal with the intricate context in the human-centric images, [48] demonstrate the effectiveness of guiding models to focus on specific areas. Nevertheless, the attention mechanisms previously proposed at the pixel level may be not equally effective for human-centric images. For one thing, flexible manipulations can alter the images' overall layout, the common objects' coordinates within two images, and the image size (shown in the case of our study in Figure 1.2). In this way, visual models will be misled to detect pixel-level differences everywhere and fail to find the relationships across images. Therefore, describing which object to focus on via natural language might be more effective in referring to the mutual components within an image pair rather than directly pinpointing a certain region in the image. For another thing, in consideration of the manipulations' rich meanings, "where to focus" largely depends on "what the manipulations try to indicate", rather than being solely related to visual information as in many other domains. So, to specify

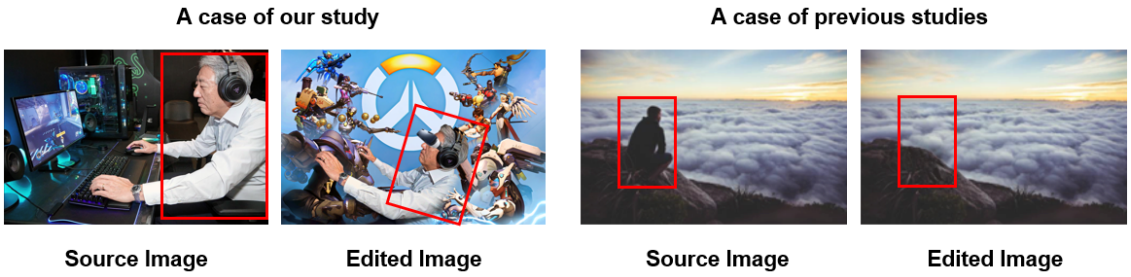

Figure 1.2: Complicated Background and Altered Layout. This is a comparison between a case we focus on and a case for the previous studies.

which component to focus on for our pipeline, we construct a taxonomy that maps the relationships from the edits' indications (defined as perception level features) to the crucial objects that require more attention (defined as focal points).

In summary, the main contributions of this thesis are as follows:

- We research two-image relationships in the human-centric image manipulation domain, which is a meaningful topic for scrutinizing malicious visual editing but remains insufficiently studied.
- We validate the feasibility of detecting manipulations based on images' captions and introduce an LLM-VMU pipeline, which is capable of handling the intricate context and versatile manipulations within human-centric images. Besides, the LLM-VMU shows robustness across different backbones, indicating its adaptability to more powerful Large Language Models (LLMs) and Multi-modal Large Language Models (MLLMs) in the future.
- We construct a Human-centric Image Manipulation Taxonomy, which is a comprehensive collection of editing types and explicitly demonstrates the relationships between the edits' indications and the objects that demand greater attention.

Chapter 2

Related Work

Human-centric image manipulation is a form of social media image forgery, where understanding the indication behind editing plays an important role. Hence, initially, we explore how previous studies connect the indications to the visual information. Subsequently, we delve into research that integrates human factors with the visual-and-linguistic reasoning process, providing a methodological foundation for us to combine audiences' perceptions of the edits' indications with manipulation understanding. However, one difference between these existing studies and our research is the task formulation: we analyze both the source image and its manipulated version, which emphasizes the relationships between images. Therefore, finally, we survey a limited number of existing two-image studies to identify points for reference and potential areas for enhancement in our research.

Social Media Image Forgery. The recent advancements in large-scale image editing techniques have amplified the threat of misinformation, making social media misinformation a prominent research direction in both the fields of computer science and social science.

Digital forgery detection [42, 38] and multi-modal fake news detection [49, 37] are two artificial intelligence topics related to our study. However, owing to the versatile nature of visual editing with various intentions, one common challenge for them is their outcomes' generalization. The existing detection methods are typically customized for specific editing types [52, 62] or only demonstrate effectiveness within a limited scope of images [37]. Therefore, a thorough and systematic consideration of editing intentions and editing types is essential to devise an approach capable of managing diverse categories of manipulations.

Studies in the field of social science can provide valuable inspiration for inducing

editing indications and how they manifest via visual information. Much previous research focused on "what visual elements (known as objective features) are more influential" in shaping viewers' impressions and engagement on social media. Factors such as settings, emotions, and face attributes are elucidated as objective features contributing significantly to these effects [33, 28]. Nevertheless, there is a dearth of research on how specific visual features influence the audience's perception, where the relationships between the objective features and the perceived features play an important role. Peng et al. [34] conducted a survey on visual misinformation and constructed a theoretical framework for visual features' persuasive mechanism by comprehensively enumerating formats, objective features, and perceived features of the visual media. However, their findings only qualitatively revealed that objective features affect viewers' understandings via the intermediate perceived features while falling short of explicitly and systematically exposing the specific relationships, which are essential to bridging the gap between visual messages' intrinsic properties and the audiences' perceptions [25, 41].

Visual-and-Linguistic Reasoning. Visual-and-linguistic reasoning encompasses a wide spectrum of topics, such as descriptive information extraction [3], physical relation inference [23, 20], science questioning-answering [31, 46], and more. Recently, there has been a growing focus on human-centric reasoning, which can facilitate a more seamless integration of AI into human daily life [32]. Human-centric reasoning takes meaningful human factors into account, including psychology [55], motivation [21], scenario premise [14], personality [63], and theory of mind [50]. In visual manipulation, the audience's perception is a crucial human factor, which drives us to incorporate perceived features into our approach.

A cause of this trend may be the development of Large Language Models (LLMs), as they emerged as potential solutions to the inherent challenges of human-centric reasoning tasks: complex scenes, commonsense knowledge and reasoning ability [54]. Most current research extends the modality by adding vision modules to project visual inputs into either discrete text words [19, 56] or continuous features [2, 15] to harness LLMs' strong capabilities. Besides, LLMs can also serve as AI agents to select proper vision experts for certain tasks [53]. These methods generally show desirable results compared with the other state-of-the-art models and provide a foundation for our methodology. However, hallucination remains a common issue in

reasoning [5, 59], compelling us to particularly discuss it along with other potential errors.

Two-image Task. Various computer vision and multi-modal benchmarks consist of image pairs, such as WSRD for image enhancement [43], HEB for homography estimation [6], Office-Home for style transfer [44]. However, they primarily focus on generating one single image, with limited attention given to the visual relationships between images. Topics pertaining to face recognition [1, 26], and medical image registration [4, 18] concentrate on the images' relationships, yet it's hard to generalize their results to different tasks due to their narrow focus on specific domains.

In recent studies, natural language has played as a more flexible and explicit means of describing relationships between images. The NLVR2 dataset [39] proposes a task to judge the correctness of a statement according to the given image pair. Initiatives like spot-the-diff [22] and relational speaker [40] are designed to detect the differences between a pair of similar images. Nevertheless, limited editing types and minor manipulating scopes (while preserving the image's general layout) are common constraints for these studies.

EMU (Edited Media Understanding) [12] dataset aims to understand human-centric image manipulations, which fit well with our requirements. However, their proposed method (PELICAN) connects the visual information directly to downstream tasks (like inferring the implication), without revealing what edits have been conducted as an intermediate result. This lack of transparency makes it challenging to interpret the results and transfer the model to other tasks.

To address the identified gaps between existing studies and our objectives, we first establish a Human-centric Image Manipulation Taxonomy in Section 3, aiming to delineate the specific relationships between content level editing types and perception level editing types. Additionally, to enhance the integration of the audience's perception of indications within the reasoning process and address potential issues (like hallucination) when applying existing LLM pipelines to our specific task, we introduce a Large Language Model driven Visual Manipulation Understanding (LLM-VMU) pipeline and utilize the constructed taxonomy to guide its workflow (Section 4).

Chapter 3

Human-centric Image Manipulation Taxonomy

In this section, to explicitly demonstrate the mapping relationships from each perception level editing type to multiple possible content level editing types, we establish the Human-centric Image Manipulation Taxonomy in a hierarchical manner and will elaborate on the taxonomy construction in the following discussion.

For general visual misinformation, there is a distinction between the objective and perceived features. Perceived features reflect mediating psychological states (i.e. the intentions for editing in our research), while objective features are intrinsic properties of visual messages and exist independently of audience perceptions [34].

Despite the gap in perceived features and objective features [25, 41], relationships between them are essential for human-centric image manipulation, as "which objects to focus on" (also defined as focal points in our research) largely depends on the manipulations' indications. So, we introduce the Human-centric Image Manipulation Taxonomy to establish these relationships, while few existing taxonomies have addressed this explicitly.

Human-centric Image Manipulation Taxonomy (Table 3) provides a comprehensive compilation of editing types at both the perception and content levels. Perception level editing types are based on "what the manipulations indicate about the main subjects", while the content level editing type categories distinguish themselves by considering "how the manipulations change the objective features". Each content level editing type is designated a specific focal point. Inferring the perception level editing type can assist in identifying the content level editing types by filtering out the irrelevant options (e.g., solely altering a teacher's facial expression cannot

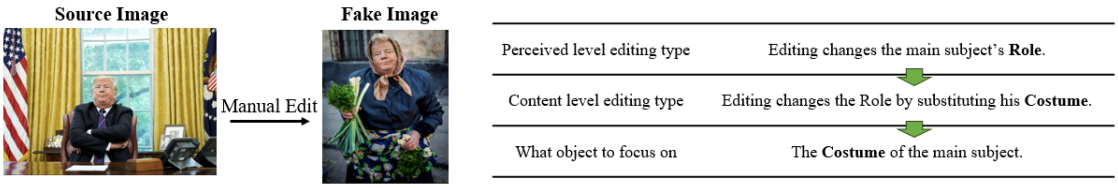

Figure 3.1: An example for demonstrating Human-centric Image Manipulation Taxonomy. The editing intends to change the politician's **Role** (perception level editing type) to dressing as an old lady by substituting the man's **Costume**, which is a content level editing type.

change his/her role), and thus narrowing down the range of possible choices. For example, the manipulations in Figure 3.1 intend to transform the politician into an old lady and change the main subject's Role at the perception level. At the content level, even though the background, the man's costume, the object held by the man, and the overall lightness are all changed, only the costume substitution contributes to the intention to change the main subject's Role. Therefore, the visual focus should be on the main subject's costume consequently.

Hence, referring to existing studies on "important objective features within social media images" and "common digital forgery types", we gather the elements associated with human-centric image manipulations and define them as perception level editing types or content level editing types. We define five meaningful dimensions related to humans (Behaviour, Role, Identity, Emotion, Scenario) as perception level editing types, and outline all possible objective manipulations for each defined dimension.

Behaviour. Behaviour is one contributing factor in shaping personal reputation, which means how individuals are perceived by others [64]. Manipulations on main subjects' behaviours can lead to individual reputation sabotage [10]. Therefore, we define the alterations of human behaviours as a meaningful editing indication when the main subjects are portrayed engaging in different actions within a given context. Manipulating the behaviours of main subjects can be achieved through the following five types of content level edits (B1-B5, shown as Figure 3.2):

- **Environment context substitution** (B1). Replace the overall background of an image, which indicates the main subjects are attending a different activity (e.g., changing the settings from a formal conference to a nightclub).

- **Movement alteration** (B2). Partially change the main subjects' actions, signifying an altered intention (e.g., transitioning a person from holding a meal plate to issuing commands to soldiers).
- **Interaction alteration** (B3). Alter the interaction between main subjects, with the potential to misinterpret their relationships (e.g., a woman is introduced, which indicates that the man is murdering the woman).
- **Object substitution** (B4). Substitute the object held by the main subjects or something the main subjects are using (e.g., a policewoman is changed from holding a dog to holding the Pepsi).
- **Content substitution** (B5). Substitute the words or content in the papers, signs, screens, paintings and so on. In human-centric images, people often interact with these elements, so we regard it as a way to change the main subjects' behaviours (e.g., the words on the board are substituted).

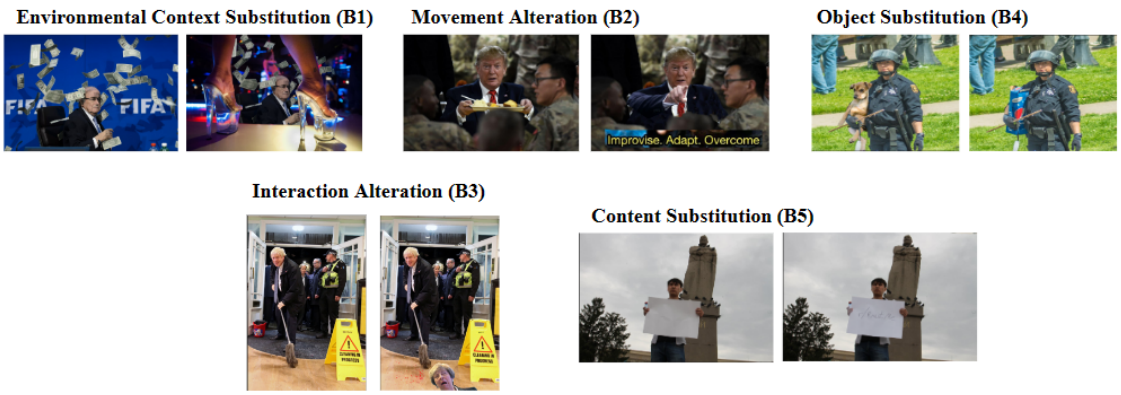

Figure 3.2: Behaviour Editing type cases.

Role. Social roles are tied to social status [36], serving not only as an indicator of people's positions in the social hierarchy [27] but also the socioeconomic resources they possess [13]. Particularly for certain groups (like country governors), their social roles directly determine their credibility [47]. Thus, editing on subjects' social roles (such as occupation, outfit, and group association) can have effects on both individuals and society, making it a meaningful perception level editing type. The

main subjects' roles can be changed by the following three content level edits R1-R3, shown as Figure 3.3):

- **Group association indication** (R1). By introducing a new group of people who did not appear in the source image, manipulations create a new group association for the main subjects. Group association indication alters how viewers perceive the roles of the main subjects via the cheerleader effect [45] (e.g., a group of men are introduced, which indicates that an officer in the source image becomes a member of them).
- **Costume substitution** (R2). Replace the main subjects' clothing to change their overall appearance (e.g., the substitution of a man's clothes makes him seem like a fighter on the battlefield).
- **Virtual character introduction** (R3). Virtual character here means the characters exclusively appear in fictional films or cartoons (like therianthropes). This editing process involves either introducing virtual characters into images or manipulating human subjects in the original images to resemble virtual characters (e.g., a man is manipulated as a cartoon bear).

Figure 3.3: Role Editing type cases.

Identity. Identity replacement techniques (like deep-fake) have the potential to harm victims due to their utility for reputational sabotage. In many instances, it's

tough to debunk the falsehoods in time, which can lead to irreversible harm. [10]. Noticing the significant consequences of identity manipulations, we define it as a dimension at the perception level and hope it can aid in recognizing detrimental manipulations on identity. Generally, there are three content level editing types (I1-I3, shown as Figure 3.4):

- **Identity theft** (I1). Replace the face or overall body with that of another person, thereby transforming the physical identity of the manipulated subject into a new individual (e.g., an ordinary person is substituted by the president).
- **Physical feature transplantation** (I2). For certain celebrities, their distinctive physical attributes serve as symbols of their identities. This editing transplants these unique features onto the subjects within the original image, for humorously exaggerating the celebrities' characteristics (e.g., the president's symbolic hair is transplanted to another man's head).
- **Celebrity introducing or removing** (I3). Introduce or remove celebrities, which may highlight the significance of specific events or indicate famous people's participation or absence in specific activities (e.g., a president is introduced and talking with another country's president happily).

Figure 3.4: Identity Editing type cases.

Emotion. Emotions are initialized by a gut reaction to a discrepancy of an expected schema and followed by cognitive analyses [17]. Hence, humans' emotions can reflect their attitudes and viewpoints on specific events. For celebrities with social impacts, their emotions can even have profound effects on the stock market [8] and voting outcomes [16]. We embrace emotion as a main indication because it not only shows the manipulated subjects' inner thoughts but also wields profound influence over the real lives. The main subjects' emotions can be manipulated via the following two types of content level edits (E1-E2, shown as Figure 3.5):

- **Facial expression alteration** (E1). Modify the main subject's facial expressions to change their overall mood (e.g., change the people's facial expressions from being angry to being surprised).
- **Emotive gesture substitution** (E2). Substitute a person's gesture to convey his different attitudes (e.g., changing people's gestures from debating to thumbing up).

Figure 3.5: Emotion Editing type cases.

Scenario. Visual Scenes' discrepancy and heterogeneity can significantly influence how people interpret images [57]. The ambience within these scenes has the potential to impact various dimensions (including mood, behaviour, and social interaction) not only among the subjects within the images but also among the viewers [7]. To assist in comprehending manipulations of visual scenes, we define the scenario as one of the editing indications. Three types of content level edits can change the scenario of human-centric images (S1-S3, shown as Figure 3.6):

- **Background manipulation** (S1). Adjust the images' background to emphasize changes in the circumstances of the main subjects, instead of indicating

they are involved in a different activity as **Environment context substitution** does (e.g., change the settings from in the street to a music festival, but the policeman is chasing the man in both images).

- **Style transformation** (s2). Transform the source image's genre or style (e.g., changing from a real photo to a film poster).
- **Lightness adjustment** (s3). Adjust the overall lightness of a given image. Malicious alterations in brightness significantly impact the personalities of the main subjects and the overall atmosphere portrayed in the image (e.g., dimming a person's portrait to evoke a menacing atmosphere and indicate his vicious personality).

Figure 3.6: Scenario Editing type cases.

Each content level editing type is allocated a specific component to focus on, and the taxonomy explicitly reveals the specific relationships between the editing intentions and the focal points. The relationships assist in integrating such indications into our pipeline by instructing the visual models to solely concentrate on the indication-related components. Therefore, to illustrate how the taxonomy guides our pipeline's workflow (Section 4.2), we'll first introduce the architecture of our pipeline in the upcoming section.

Perception level Editing Type	**Content level Editing Type**	**What objects to focus on?**
Behaviour	Environment context substitution	the environment
	Movement alteration	the movement of the main subject
	Interaction alteration	the main subject's behaviour towards other people
	Object substitution	what is the main subject holding or using?
	Content substitution	the content on the paper, sign, painting or screen
Role	Group association indication	the group of main subjects
	Costume substitution	the costume of the main subject
	Virtual character introduction	what the main subject looks like
Identity	Identity theft	the identity of the main subject
	Physical feature transplantation	the physical feature of the main subject
	Celebrities introducing or removing	the celebrities
Emotion	Facial expression alteration	the facial expressions
	Emotive gesture substitution	the gestures
Scenario	Background manipulation	the environment
	Style transformation	the overall setting and style
	Lightness adjustment	the overall lightness

Table 3.1: The Human-centric Image Manipulation Taxonomy is a comprehensive classification system for human-related image editing. It comprises 5 perception level editing types and 16 content level editing types. Each perception level editing type corresponds to multiple possible content level editing types, and each content level editing type has one specific focal point to concentrate on.

Chapter 4

Large Language Model driven Visual Manipulation Understanding

Our method, the Large Language Model driven Visual Manipulation Understanding (LLM-VMU), aims to detect the manipulations between the source image I_S and the edited image I_E based on their respective captions C_S, C_E. To identify the manipulations E, we use large language models (LLMs) to detect all differences based on the information from (C_S, C_E).

One potential challenge in detecting differences solely based on captions is that LLMs may return hallucinated results. This is because when captioning a pair of images with complicated contexts separately, it cannot be guaranteed that the descriptions of unaltered components will remain the same (just like the second inferred manipulation about the blue tie in Figure 4.1). Therefore, in Section 4.1, we introduce Visual Information Checking (VIC) to evaluate if a detected difference is consistent with the visual information and filter out incorrect ones.

Another challenge is detecting manipulations at proper granularity. The distinction between manipulation understanding and simply spotting differences lies in the requirement to comprehend the underlying intentions behind the changes. A detected difference may be accurate, but if it fails to help convey the intended indications, it cannot be regarded as a meaningful manipulation. So, in Section 4.2, we utilize the constructed Human-centric Image Manipulation Taxonomy to specify what objects in the image should be focused on (also known as focal points) according to the edits' indications. The focal point (G) can guide LLM-VMU in only detecting manipulations related to essential objects while ignoring trivial disparities.

Next, we'll illustrate specifically how the VIC and Taxonomy Guideline work to solve the two potential challenges in detail.

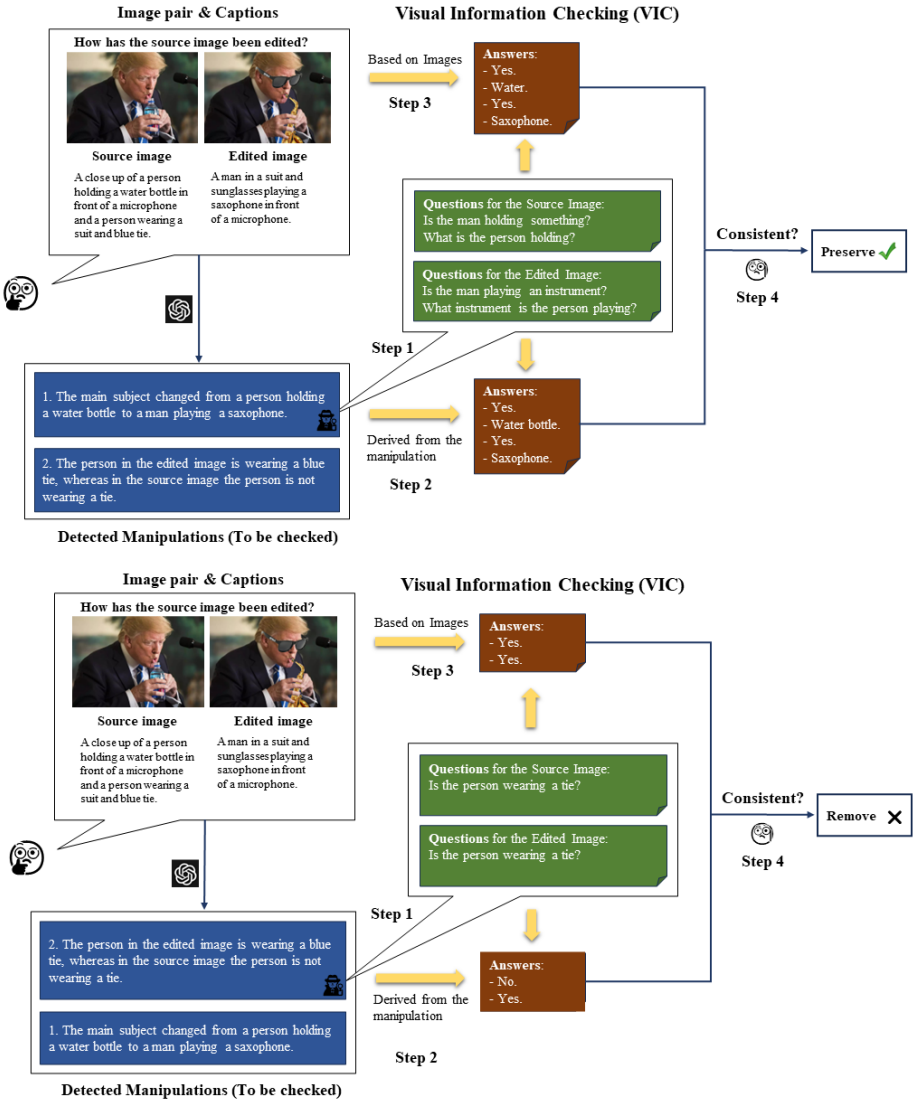

Figure 4.1: The figure illustrates an example where LLM-VMU infers candidate manipulations and ensure the correctness of the first one. The modules using LLM backbones and MLLM backbones are distinguished by colours.

4.1 Visual Information Checking (VIC)

Visual Information Checking (VIC) evaluates the correctness of each detected manipulation e_i returned from LLMs and decides to preserve, remove, or partially adjust it. To align the textual information within detected manipulations with the

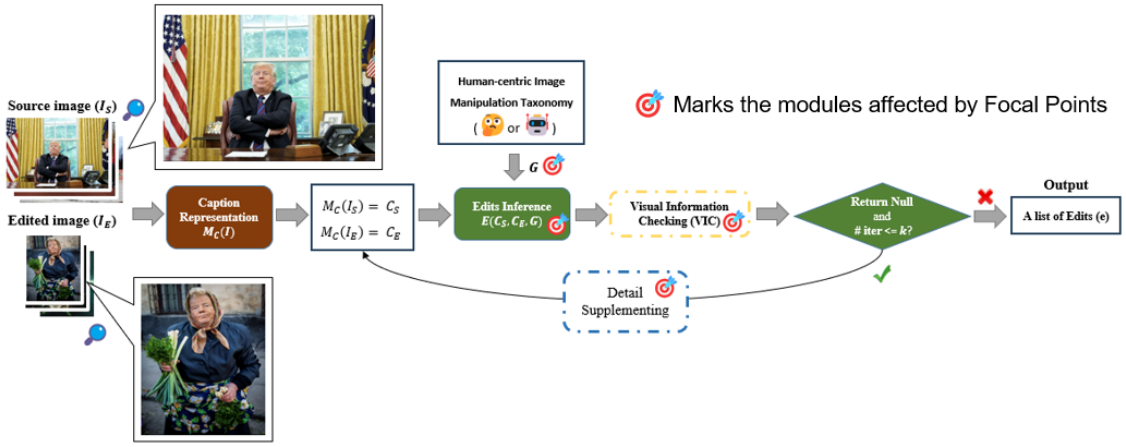

Figure 4.2: The figure demonstrates the architecture of LLM-VMU pipeline. The modules using LLM backbones and MLLM backbones are distinguished by colours. The symbol "target" indicates the modules affected by the Taxonomy Guideline. The input focal point (G) is defined based on our proposed Human-centric Image Manipulation Taxonomy, which can be gained either by manually assigning an editing type to an image pair, or automatically detecting the editing type of the image pair by Focal Point Detection module.

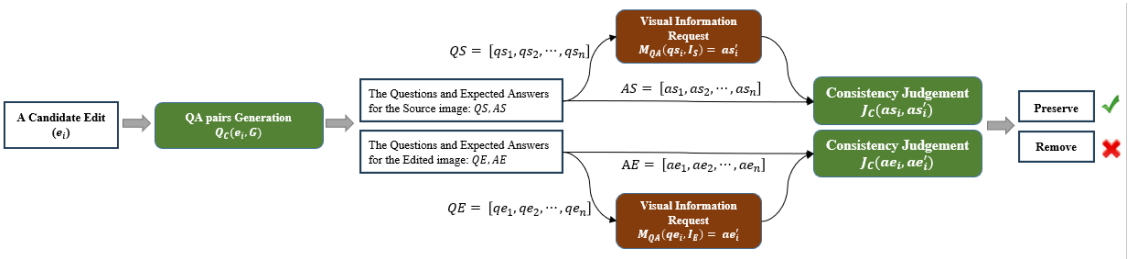

Figure 4.3: The figure demonstrates the process for VIC to check the correctness of one candidate manipulation (e_i). LLM backbones and MLLM backbones are distinguished by different colours.

visual information in image pairs, LLM-VMU generates specific questions (by QA Pairs Generation) and assesses the consistency (via Consistency Judgement) between the answers based on the manipulations and the answers derived from images (as shown in Figure 4.3). The final determination of the action on each manipulation depends on their consistency.

4.1.1 QA Pair Generation

QA Pair Generation aims to generate a list of effective questions and the corresponding answers for the source image ($I_S : \{QS = [qs_1, ..., qs_n], AS = [as_1, ..., as_n]\}$) and the edited image ($I_E : \{QE = [qe_1, ..., qe_n], AE = [ae_1, ..., ae_n]\}$) separately. To achieve this, we prompt LLMs to consider "to verify the correctness of all the points within e_i, what questions should be posed (QS, QE), and what are the expected answers based on e_i (AS, AE)".

4.1.2 Consistency Judgement

Besides the expected answers (AS, AE) derived from manipulations, we employ Visual Question and Answering (VQA) backends to generate answers to the same set of questions based on the images ($AS' = [as_1^{'}, ..., as_n^{'}]$, $AE' = [ae_1^{'}, ..., ae_n^{'}]$). For assessing if the two versions of answers are consistent, we provide the triples $(qs_i, as_i, as_i^{'})$ or $(qe_i, ae_i, ae_i^{'})$ and prompt LLMs to judge whether as_i and $as_i^{'}$ (or ae_i and $ae_i^{'}$) are the same. Besides, we require LLMs to attempt generating Yes/No questions, to ensure that the consistency assessment is objective and convincing.

4.1.3 Preserve or Remove?

For each detected manipulation to be checked (e_i), multiple QA pairs are generated and the consistency judgement contains a list of results ($J = [J_1, ..., J_n]$) consequently. We introduce two variations, Automatic Correction and Question Selection to decide how to process e_i based on J.

Automatic Correction. Automatic Correction applies all the generated questions in QS and QE for verification and automatically revises partially correct manipulations by removing incorrect elements. This process contains three distinctive situations:

1. If e_i passes the verification of all the questions in QS and QE, we directly accept it as a correct manipulation (like the operations on the first candidate manipulation in Figure 4.1)

2. If e_i fails to pass the checking of all the questions in either QS or QE, we assert its description of at least one image is entirely incorrect, rendering this

statement regarding the relationships between the images meaningless. Then, we reject and remove e_i.

3. If e_i passes the checking of partial questions in both QS and QE, we regard e_i as partially correct and prompt LLMs to eliminate incorrect elements.

The motivation of automatic correction is to preserve as much correct information as possible, while it carries the risk of generating incoherent sentences or mistakenly eliminating precise points owing to an inaccurate removal of incorrect elements.

Question selection. Question selection extracts one question from QS and QE separately, and verifies e_i based on the two selected questions. Here are two possible situations:

1. If e_i passes both questions' verification, we preserve it.

2. If e_i fails to pass the checking of either question, we reject and remove it.

The standard of selecting questions is the relevance with key objects (G) provided by the taxonomy guideline. Referring to Verbalized Confidence [51], we instruct LLMs to directly score the relevance and select the question with the highest score. When multiple questions have the same highest score, we randomly sample one from them.

With question selection, we only inspect the information related to the core component (G) and permit the existence of other errors. However, selecting an improper question may lead to discarding the correct elements or retaining incorrect edits.

4.1.4 Detail Supplementing

The detail supplementing module addresses an error when the pipeline doesn't return any manipulations. A null return occurs when the initialized captions (C_S, C_E) are not specific enough to cover the manipulated objects. Inspired by an iterative method for requesting additional information from images [61], we instruct LLMs to generate questions related to unclear entities within the initialized captions (C_S, C_E, or the focal object G from the guideline if available) and retrieve the corresponding answers using VQA models. With each request for extra information, LLM-VMU

updates the captions of both images (C_S, C_E) by incorporating the details in the question and the answer. The following VIC steps will be conducted on these updated captions, and the process continues in a loop until at least one manipulation is returned or the maximum iteration count (hyperparameter k) is reached.

4.2 Taxonomy Guideline

Within the Human-centric Image Manipulation Taxonomy, each content level editing type correlates with one specific focal point, while a manipulation's content level editing largely depends on the inferred (perception) editing type. Hence, we illustrate two questions here: 1) how the focal points affect our pipeline (Section 4.2.1), and 2) how to identify each image pair's focal point (or content level editing type) except via manual annotation (Section 4.2.2).

4.2.1 Guided modules

Generally, with a focal point G, we use it to guide the workflow of LLM-VMU by simply inserting G into the prompt for certain modules.

Before inferring manipulations, we instruct LLMs to extract information solely related to G for filtering out irrelevant noise, following a process similar to Context Selection [11]. Besides, G assists in inferring the manipulations by providing an emphasis, while guiding the VIC and Detail supplementing by controlling the topics for question generation. Generally, the taxonomy guideline directs LLM-VMU to only concentrate on the core components while filtering out trivial noise (as shown in Figure 4.2).

4.2.2 Focal point detection

Besides manual annotations, we introduce an automatic focal point detection module for specifying which objects to focus on (shown in Figure 4.4). To take the edits' indications into account, this module identifies the perception level editing type first and then explores all possible content level editing types and the corresponding focal points (G_i) according to the mapping relationships in the Human-centric Image Manipulation Taxonomy. In consideration that multiple editing types may occur simultaneously, for each possible category, we compare whether two images

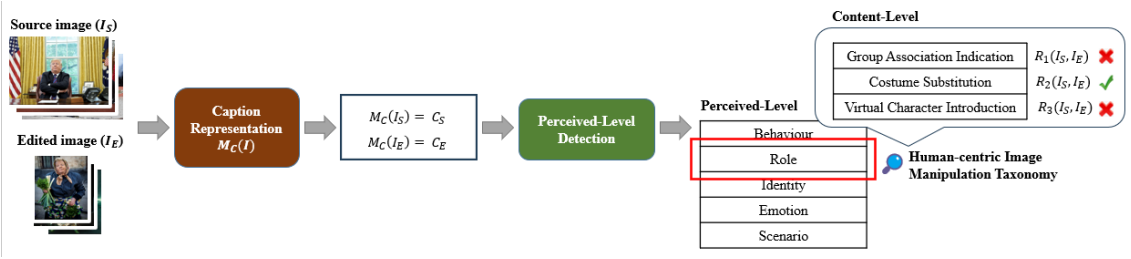

Figure 4.4: The figure demonstrates how Focal Point Detection works. The module detects the editing type and the corresponding focal points of an image pair following the definition of Human-centric Image Manipulation Taxonomy. LLM backbones and MLLM backbones are distinguished by different colours.

exhibit disparities in G_i to verify the presence of specific content level editing types. The aggregation of all detected G_i forms G, which guides the functioning of LLM-VMU. For example, in Figure 4.4, after inferring the intention is manipulating the gentleman's role, we use three questions (Is there a group of people? What is the person wearing? Is there any virtual character?) to detect the existence of three potential content level editing types via MLLMs. Upon comparison, disparities in the two images' responses are observed only for the second query, and thus we need to focus on the costume of the main subject for Costume Substitution (R_2).

Chapter 5

Experiments

In this section, we commence with experiments to validate the effectiveness of our proposed LLM-VMU by comparing it against the baselines and the previous state-of-the-art models. Subsequently, we conduct an ablation study to investigate each module's function and determine which modules contribute to the overall improvement. Besides, we test the robustness of the LLM-VMU and investigate whether it still outperforms baselines after substituting the internal backbones. The results can also offer insights into the values and applications of our pipeline when more capable LLMs and MLLMs emerge in the future. Finally, we examine specific cases to assess if the conclusions drawn from the experimental results align with our intuitive judgements.

5.1 Set up

Dataset. We conduct our experiments on the Edited Media Understanding (EMU) dataset [12] to validate our pipeline. The EMU dataset includes both the source images and their edited versions, along with ground truths regarding the manipulations' corresponding implications, intents, effects and impacts on the main subjects' mental states. The edits within the EMU dataset are meaningful, flexible, significant and human-centric (such as the cases in Figure 5.1, Figure 5.4), which aligns well with the objectives of understanding human-centric image manipulations. Hence, we apply the test set of EMU, containing 720 image pairs to evaluate our pipeline.

Evaluation. We employ both automatic evaluation and human evaluation to ensure the credibility of the evaluation results for this open-ended task. For

automatic evaluation, we provide the detected manipulations and prompt LLMs to generate the answers to four downstream tasks (implications, intents, effects, and main subjects' mental states) whose ground truths are accessible in the EMU dataset. In this way, we indirectly evaluate the quality of the detected manipulations by evaluating the pipeline's performance on these four downstream tasks. As the ground truths are provided in a template format, we eliminate the constant components and calculate the CIDEr score to reduce the weight of tokens with high frequency. Besides, we also conduct a human evaluation of the predicted manipulations on a subset containing 100 cases. The human evaluation focuses on three dimensions:

- **Correctness.** Whether a description is consistent or in conflict with the provided image pair.
- **Relevance.** Whether a manipulation is related to the main indication behind the editing.
- **Completeness.** Whether the predicted manipulations cover all the valid points which contribute to conveying the indications of the given image pair.

Backbones and Hyperparameter. Our main experiments are based on the Instruct-GPT LLM and BLIP MLLM. In the LLM-VMU pipeline, all modules are set as zero-temperature and zero-shot, to exclude the influences of LLM variance and ensure the replicability of the results. In our automatic evaluation, to map the detected manipulations to the four downstream tasks, we design four-shot prompts to guide LLM in generating responses in certain formats.

5.2 Performance Comparison

We initially compare our proposed baseline with the previous state-of-the-art models (Table 5.1). The baseline here refers to a pipeline only inferring manipulations based on captions without the attendance of either VIC or Taxonomy Guideline. For this comparison, we follow the evaluation metrics (ROUGE-L and METEOR) used in the prior study [12], and our proposed baseline significantly outperformed the previous models by 10.2 on ROUGE-L and 5.8 on METEOR. The results affirm

Models	ROUGE-L	METEOR
GPT-2 [35]	10.3	6.5
Cross-Modality GPT-2	12.0	7.9
Dynamic RA [40]	13.2	8.9
VIP [60]	19.5	10.8
PELICAN [12]	22.1	11.6
Baseline (LLM-VMU w/o VIC, w/o TG)	**32.3**	**17.8**

Table 5.1: The performance comparison between previous SOTA models and our proposed baseline. The baseline refers to the LLM-VMU without VIC and Taxonomy Guideline. TG denotes the Taxonomy Guideline. Previous research uses ROUGE-L and METEOR to evaluate their results, so we follow their metrics for making comparisons.

the feasibility and advantages of representing visual information as captions for detecting manipulations between images with complicated context.

Besides, we also compare the LLM-VMU with the baseline to validate the effectiveness of the VIC and Taxonomy Guideline (Table 5.2). Notably, LLM-VMUs with different settings overall outperform the baseline on all four downstream tasks. However, the pipelines guided by automatically detected focal points don't perform as well as the pipelines with manually annotated focal points. Based on our analysis, one main reason is that the focal point detection module cannot specify the key objects accurately enough, with 26.8% of cases failing to identify any focal point, and only 35.3% of cases aligning with humans' judgements. So, for tasks like Mental state, which are not directly related to the event or need further reasoning, noise introduced by the incorrect focal points can lead to an even worse performance than the baseline in implicit long reasoning chains. Nevertheless, the results still pose a promising direction for future research and underscore the potential of focal point detection, as evidenced by the significant effects of the Taxonomy Guideline with manually annotated focal points.

	Automatic Evaluation ↑					**Human Evaluation ↓**		
	Overall	Implication	Intent	Mental State	Effect	Correctness	Relevance	Completeness
Baseline	8.5	5.2	6.4	10.7	8.6	3.05	2.98	**1.47**
LLM-VMU(Automatic Correction + Detected TG)	9.6	7.1	8.6	10.3	9.3	2.28	2.05	2.06
LLM-VMU(Question Selection + Detected TG)	9.4	6.7	8.2	10.4	9.2	2.16	2.05	1.66
LLM-VMU(Automatic Correction + Human Labelled TG)	**10.7**	**8.1**	8.5	**11.9**	**11.0**	1.64	**1.82**	1.98
LLM-VMU(Question Selection + Human Labelled TG)	10.4	7.8	**8.8**	11.1	10.9	**1.62**	1.86	2.23

Table 5.2: Comparisons of automatic and human evaluation between the baseline and the LLM-VMU. Detected TG refers to the pipeline with an Automatically Detected focal point for the Taxonomy Guideline, whereas Human labelled TG means the pipeline with a manually annotated focal point. We calculate the CIDEr score for automatic evaluation and utilize the average ranks within the five pipelines for human evaluation. Average rank denotes the average of each pipeline's ranks for all the cases within the subset.

According to the human evaluation, LLM-VMU considerably improves the Correctness and Relevance of the manipulations compared with the baseline, whereas with the cost of sacrificing the Completeness. This trade-off meets our expectations, as both the VIC and the Taxonomy Guideline aim to filter out improper (incorrect or trivial) detected manipulations from the baseline's results. Consequently, the baseline's results tend to be more complete, given that LLM-VMU's removal process may not be entirely accurate.

Generally, the automatic evaluation metric aligns with the human evaluation rankings when comprehensively considering Correctness, Relevance and Completeness. Additionally, although the pipeline with Automatic Correction VIC and Question Selection VIC exhibit similar performances, Automatic Correction is more suitable for inferring downstream tasks as it preserves more correct information. Therefore, our subsequent discussions will place greater emphasis on the automatic evaluation and the automatic correction VIC.

5.3 Ablation Study

To test the VIC's and the Taxonomy Guideline's effects, we ablate the two modules separately (Table 5.3). The significant drops indicate that they both play an important role in the LLM-VMU, but the variations in the distribution of these drops indicate that they contribute to different specific tasks. Specifically, the Taxonomy Guideline contributes more to inferring the Implication and Mental state, as these tasks are closely related to the intentions behind the editing and only require the identification of dominant manipulations. However, when considering the effects of the editing, it focuses more on the details of the manipulations and poses a higher requirement for correctness. VIC aims to ensure the correctness of all details in the detected manipulations and thus contributes to this category of tasks more.

5.4 Robustness Analysis

When the LLM and MLLM backends are altered, LLM-VMU still significantly outperforms the baseline on almost all the tasks (except inferring the Intent when using the Shikra MLLM backend). However, an unintuitive finding is that when

Models	Overall	Implication	Intent	Mental State	Effect
LLM-VMU(Automatic Correction + Human Labelled TG)	**10.7**	**8.1**	**8.5**	**11.9**	**11.0**
w/o VIC	9.5	7.1	7.7	11.2	8.3
w/o Taxonomy Guideline	9.1	6.4	7.8	10.4	8.9

Table 5.3: The automatic evaluation results of ablation study based on LLM-VMU with automatic correction VIC and Taxonomy Guideline with human labelled focal points. The table shows the CIDEr Score. Ablating the VIC and Taxonomy Guideline separately both exhibit a significant drop in the results.

Backbone	Pipeline	Overall	Implication	Intent	Mental State	Effect
Instruct-GPT + BLIP	LLM-VMU	**10.7**	**8.1**	**8.5**	**11.9**	**11.0**
	Baseline	8.5	5.2	6.4	10.7	8.6
Instruct-GPT + Shikra	LLM-VMU	**8.6**	**7.2**	6.2	**9.8**	**7.3**
	Baseline	7.3	4.6	**6.8**	9.1	5.7
LLaMA2 + BLIP	LLM-VMU	**8.9**	**5.4**	**7.2**	**10.9**	**9.0**
	Baseline	7.8	4.4	5.7	10.1	7.5

Table 5.4: The automatic evaluation results when the LLM backbones and the MLLM backbones are changed. We utilize the CIDEr Score as the evaluation metric. The results show that LLM-VMU can still generally outperform the baseline with different backends.

the MLLM backbone is changed to a more capable one (Shikra) [30], the results show a significant decrease. Through our observations of the cases, we find two contributing reasons. For one thing, in the VIC module, the generated questions are relatively straightforward, imposing minimal requirements on MLLMs' reasoning and understanding capabilities, i.e. even though using a stronger MLLM backbone, the VIC's performance cannot be boosted by solely answering the generated questions more accurately. For another thing, although stronger MLLMs have the ability to generate captions with greater linguistic complexity and richness, this can also introduce noise, which may confuse the LLM when it infers the differences based on textual image captions.

When we degrade the LLM backbones, the pipeline with LLaMA2 (13B parameters) does not perform as effectively as the pipeline with Instruct-GPT (175B parameters). This suggests that LLMs with enhanced reasoning abilities can signifi-

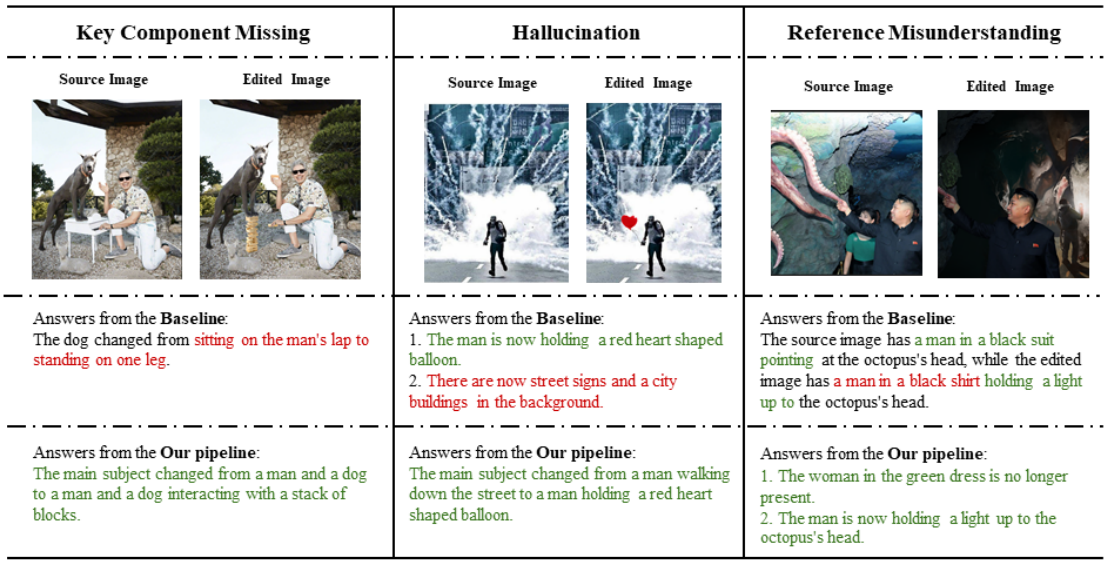

Figure 5.1: Three types of errors occur frequently for the baseline: Key component missing, Hallucination, and Reference misunderstanding. Comparing the LLM-VMU's answers with the baseline's answers, our pipeline can mitigate the potential errors.

cantly boost the performance of our pipeline, as they are more adept at identifying distinctions between two image captions and generating suitable questions under control. This also indicates that strong LLMs should be given higher priority over MLLMs when only limited space costs and time costs are allowed.

5.5 Case Study

Through our observations of cases, we have identified several types of potential errors when detecting manipulations solely based on the images' textual captions (the baseline). Some of these errors can be alleviated by our proposed LLM-VMU pipeline, while others still remain as challenges.

The LLM-VMU can mitigate three types of errors: Key Component Missing, Hallucination, and Reference Misunderstanding.

- Key Component Missing (E1). The returned manipulations fail to mention the necessary objects (the stack of blocks in Figure 5.1) for describing the editing.

There are two underlying causes: 1) the MLLM incorrectly understands the object as something else (the man's leg in Figure 5.1), and 2) the initialized captions are not specific enough to cover the crucial objects.

- Hallucination (E2). Hallucinated manipulations are not consistent with the visual information within images. In the initialized pair of captions, the MLLM cannot guarantee that the descriptions of unmanipulated components (the street signs and city buildings in the background in Figure 5.1) remain entirely the same. This inconsistency leads the LLM to identify incorrect relationships between captions and consequently detect hallucinated differences.

- Reference Misunderstanding (E3). Reference Misunderstanding occurs when there is a mismatch between the subjects and the behaviours. When representing images as natural language, different subjects may use the same preposition for reference (both the president in black clothes and the person holding a flashlight can be denoted as "the man" in Figure 5.1), which can perplex the LLM in determining which subject is associated with specific actions.

For the Key component missing (E1) error, the Detail Supplementing module and Taxonomy Guideline can instruct the pipeline to add more details about the key components into the original captions to rectify their lack of specificity. In response to the Hallucination (E2) error, both the VIC module and the Context Selection can contribute by eliminating incorrect hallucinated manipulations or removing the descriptions of trivial components from the initialized captions. LLM-VMU doesn't intentionally address the Reference misunderstanding (E3) error, but the Automatic Correction VIC can indirectly mitigate this issue by partially removing the incorrect elements from the detected manipulations.

In assessing whether our pipeline can generally alleviate the observed errors, we compare the error type distribution (how many pieces of the detected manipulations are correct, or incorrect with certain types of errors) and the error occurrence frequency (how many image pairs contain incorrect manipulations with each type of errors) of the baseline and the LLM-VMU on a subset containing 100 cases. In both visualizations, our pipeline demonstrates a general reduction in the frequency of all error types (in Figure 5.2) and notably improves the proportion of accurately detected

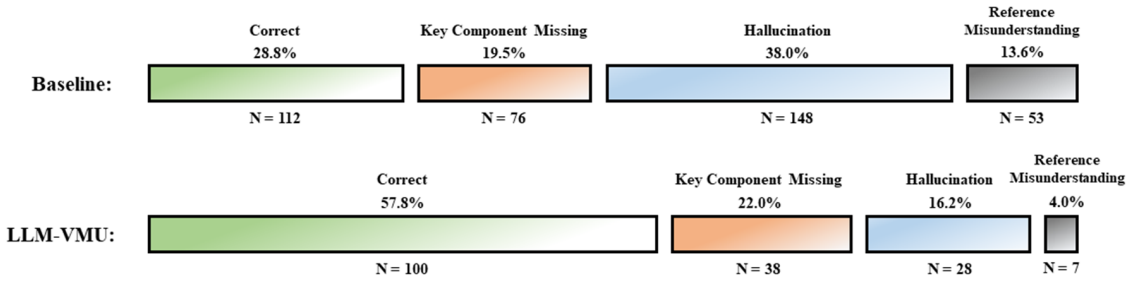

Figure 5.2: The bar chart shows the distribution of error type among all detected manipulations. The statistics for the baseline and the LLM-VMU are conducted on a subset including 100 image pairs.

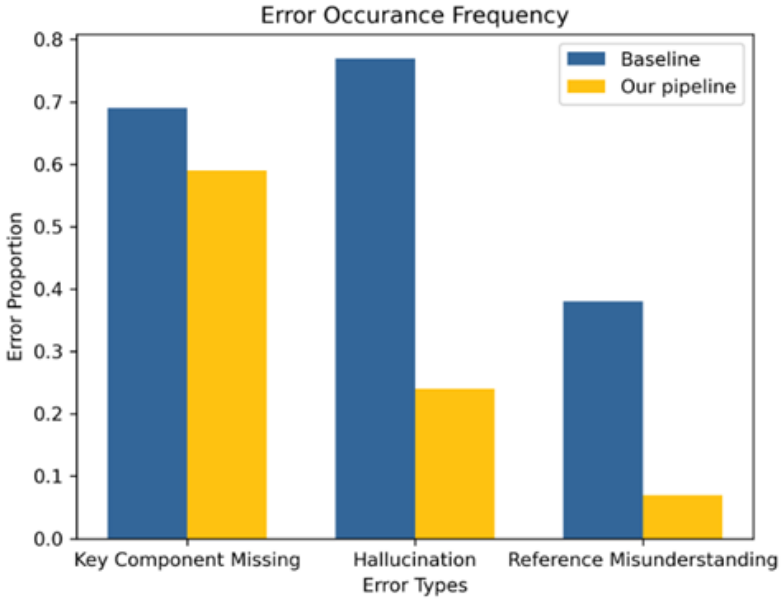

Figure 5.3: The bar chart illustrates the frequency of each error type occurring within the subset cases. The statistics for the baseline and the LLM-VMU are conducted on a subset including 100 image pairs.

manipulations (in Figure 5.3). However, the Key component missing error remains a primary challenge for LLM-VMU (according to the labelled proportion in Figure 5.2), due to the limitations in the MLLM's capability to accurately comprehend specific components.

Moreover, although LLM-VMU generally improves the baseline, we observe some instances where LLM-VMU operates erroneously (shown in Figure 5.4). We envision that future research will enhance our pipeline by generating precise questions

Source Image	Edited Image	Errors
		**Improper Generated Question** Wrongly Removed Edit: The overall lightness changed from not specified to black and white. QG: Source Q: What color are the main subjects wearing? EA: Unspecified. RA: Black. ✗ Edited Q: Is the overall lightness black and white? EA: Yes. RA: Yes ✓
		**Incorrect Requested Visual Information** Wrongly Removed Edit: The edited image also shows the man in the back of the helicopter holding a child in the air with a gun and a helmet on the back of the air. QG: Source Q: Is there a gun in the image? EA: No. RA: Yes. ✗ Edited Q: Is there a gun in the image? EA: Yes. RA: Yes. ✓
		**Inaccurate Context Selection** Guidance: *the object held by the main subject* Original Source Caption: A group of men in suit and white shirts sitting around in a room and talking with each other. Original Edited Caption: A group of people standing around the table **with a box**. Selected Source Context: A group of men discussing in a room. Selected Edited Context: A group of people discussing around a table.

Figure 5.4: Cases illustrating three common error types for LLM-VMU.

for verifying each detected manipulation (to mitigate the Improper Generated Question error), distilling accurate information from images via stronger MLLMs (for handling Incorrect Request Visual Information), and extracting comprehensive information related to focal points from original captions (for solving Inaccurate Context Selection). Two potential factors may lead to the incorrect cases. For one thing, the error cases are caused by the limited capabilities of LLM (or MLLM) backbones, which is a common bottleneck for related studies owing to LLMs' weakness in common or specialized knowledge, spatial relationship understanding, and generating fully correct descriptions. Taking the Incorrect Request Visual Information as an example, the cause of this error is that existing MLLM backbones face the challenges of precisely comprehending specific visual components. For another thing, different from the pixel-level comparisons, our LLM-VMU suffers from the absence of visual information in most steps (except Visual Information Request in the VIC module mentioned in 4.1.2). This leads LLMs (or MLLMs) to generate erroneous instances when summarizing original captions (*i.e.*, Inaccurate Context Selection) and acting as Questioners (*i.e.*, Improper Generated Question), although previous studies have demonstrated LLMs' extraordinary capabilities in

these two tasks. The analysis also encourages us to discuss the pixel-level methods' and our LLM-based pipeline's separate functions.

We compare two datasets with distinct features to delineate the boundaries of our LLM-VMU pipeline. As Table 5.5 shows, the LLM-VMU performs better in terms of the more flexible (*i.e.* the source image's overall layout is changed, and the number of conducted manipulations is uncertain) and complex (*i.e.* the background is complicated and the average lengths of descriptions are longer) manipulations. By representing visual information as textual captions, they reflect the stories conveyed by the whole scene by ignoring trivial details and extracting only the important information related to the events depicted within the images. Hence, our LLM-VMU will not be confused by the countless particular manipulations as it only cares about the main stories. By contrast, for cases in the Image Editing Request dataset, the manipulations are tiny and are mainly related to one specific component. The results indicate that comparing this kind of image pairs pixel by pixel may be more efficient than our pipeline, as there is a likelihood of overlooking crucial details when representing images as textual captions.

Through this comparison, we observe a trade-off between the capabilities of textual caption-based methods and pixel-level methods. This helps explain why there are still instances where our pipeline operates erroneously. It also suggests a promising direction to combine both methods, allowing each of them to deal with the cases where they excel, similar to the model selection studies for reasoning tasks [29, 58].

Dataset		Edited Media Understanding		Image Editing Request	
Examples					
The Number of Manipulations		Uncertain		Constantly one	
Are the images' layouts changed?		Yes		No	
Is the background complicated?		Yes		No	
Average Length of Ground Truths		24.4		7.9	
Performance	Metrics	METEOR	ROUGE-L	METEOR	ROUGE-L
	Previous SOTA (Pixel-level)	11.6	22.1	**12.8**	**37.3**
	LLM-VMU (Ours)	**17.8**	**32.3**	11.3	25.8

Table 5.5: The dataset Edited Media Understanding aligns with our research, while the dataset Editing Request is applied by the previous studies. This table compares the two datasets and shows that the manipulations within Edited Media Understanding are more complicated and flexible. Besides, the results show that our pipeline can outperform the previous SOTA models on the Edited Media Understanding dataset while cannot outperform on the Image Editing Request dataset, which indicates that our LLM-VMU is good at handling the flexible and complicated cases, but with the risk of ignoring specific details.

Chapter 6

Conclusions

We have investigated the feasibility of understanding visual manipulations based on images' textual captions and introduced a Large Language Model driven Visual Manipulation Understanding (LLM-VMU) pipeline effectively mitigating the key component missing, hallucination and reference misunderstanding errors. We observed that detecting based on images' textual captions can handle flexible and intricate manipulations by extracting only important information related to the event, while precise comparison at the pixel level is more efficient for cases where specific details are more important. Inspired by model selection research in reasoning tasks, we believe that combining the two methods may be a promising direction for future studies. Additionally, we construct a taxonomy to bridge the gap between perception level features and content level features for the human-centric image manipulation understanding task and validate its effects for guiding the workflow of LLM-VMU.

However, there are still some limitations in our work. This work has established a Human-centric image manipulation taxonomy, which is a collection of editing types. We anticipate subsequent studies to augment this taxonomy to achieve greater comprehensiveness. We also expect that our study can inspire future research to construct new taxonomies tailored to different motivation-driven tasks. Besides, we propose an Automatic Focal Point Detection module, whose accuracy requires further improvements. Nevertheless, we present it as a convincing future direction and encourage the community to explore more flexible solutions to harness LLM's reasoning abilities in assisting visual tasks.

In our future work, we aim to delve deeper into applying LLMs' strong reasoning

abilities at the perception level to instruct the downstream visual tasks at the content level, especially focusing on the topics closely linked to intricate intentions or rich indications.

Bibliography

[1] I. Adjabi, A. Ouahabi, A. Benzaoui, and A. Taleb-Ahmed, "Past, present, and future of face recognition: A review", *Electronics*, vol. 9, no. 8, p. 1188, 2020.

[2] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, *et al.*, "Flamingo: A visual language model for few-shot learning", *Advances in Neural Information Processing Systems*, vol. 35, pp. 23 716–23 736, 2022.

[3] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "Vqa: Visual question answering", in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2425–2433.

[4] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca, "Voxelmorph: A learning framework for deformable medical image registration", *IEEE transactions on medical imaging*, vol. 38, no. 8, pp. 1788–1800, 2019.

[5] Y. Bang, S. Cahyawijaya, N. Lee, W. Dai, D. Su, B. Wilie, H. Lovenia, Z. Ji, T. Yu, W. Chung, *et al.*, "A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity", *arXiv preprint arXiv:2302.04023*, 2023.

[6] D. Barath, D. Mishkin, M. Polic, W. Förstner, and J. Matas, "A large-scale homography benchmark", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2023, pp. 21 360–21 370.

[7] Y. Benkhedda, D. Santani, and D. Gatica-Perez, "Venues in social media: Examining ambiance perception through scene semantics", in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 1416–1424.

[8] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market", *Journal of computational science*, vol. 2, no. 1, pp. 1–8, 2011.

[9] T. Brooks, A. Holynski, and A. A. Efros, "Instructpix2pix: Learning to follow image editing instructions", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 392–18 402.

[10] B. Chesney and D. Citron, "Deep fakes: A looming challenge for privacy, democracy, and national security", *Calif. L. Rev.*, vol. 107, p. 1753, 2019.

[11] A. Creswell and M. Shanahan, "Faithful reasoning using large language models", *arXiv preprint arXiv:2208.14271*, 2022.

[12] J. Da, M. Forbes, R. Zellers, A. Zheng, J. D. Hwang, A. Bosselut, and Y. Choi, "Edited media understanding frames: Reasoning about the intent and implications of visual misinformation", in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 2026–2039.

[13] P. Demakakos, J. Nazroo, E. Breeze, and M. Marmot, "Socioeconomic status and health: The role of subjective social status", *Social science & medicine*, vol. 67, no. 2, pp. 330–340, 2008.

[14] Q. Dong, Z. Qin, H. Xia, T. Feng, S. Tong, H. Meng, L. Xu, Z. Wei, W. Zhan, B. Chang, S. Li, T. Liu, and Z. Sui, "Premise-based multimodal reasoning: Conditional inference on joint textual and visual clues", in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 932–946. [Online]. Available: `https://aclanthology.org/2022.acl-long.66`.

[15] D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, *et al.*, "Palm-e: An embodied multimodal language model", *arXiv preprint arXiv:2303.03378*, 2023.

[16] J. Garry, "Emotions and voting in eu referendums", *European Union Politics*, vol. 15, no. 2, pp. 235–254, 2014.

[17] M. S. Hannula, "Attitude towards mathematics: Emotions, expectations and values", *Educational studies in Mathematics*, vol. 49, no. 1, pp. 25–46, 2002.

[18] G. Haskins, U. Kruger, and P. Yan, "Deep learning in medical image registration: A survey", *Machine Vision and Applications*, vol. 31, pp. 1–18, 2020.

[19] Y. Hu, H. Hua, Z. Yang, W. Shi, N. A. Smith, and J. Luo, "Promptcap: Prompt-guided task-aware image captioning", *arXiv preprint arXiv:2211.09699*, 2022.

[20] D. A. Hudson and C. D. Manning, "Gqa: A new dataset for real-world visual reasoning and compositional question answering", in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 6700–6709.

[21] O. Ignat, S. Castro, H. Miao, W. Li, and R. Mihalcea, "WhyAct: Identifying action reasons in lifestyle vlogs", in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 4770–4785. [Online]. Available: `https://aclanthology.org/2021.emnlp-main.392`.

[22] H. Jhamtani and T. Berg-Kirkpatrick, "Learning to describe differences between pairs of similar images", *arXiv preprint arXiv:1808.10584*, 2018.

[23] J. Johnson, B. Hariharan, L. Van Der Maaten, L. Fei-Fei, C. Lawrence Zitnick, and R. Girshick, "Clevr: A diagnostic dataset for compositional language and elementary visual reasoning", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2901–2910.

[24] B. Kawar, S. Zada, O. Lang, O. Tov, H. Chang, T. Dekel, I. Mosseri, and M. Irani, "Imagic: Text-based real image editing with diffusion models", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6007–6017.

[25] H. S. Kim, "Attracting views and going viral: How message features and news-sharing channels affect health news diffusion", *Journal of Communication*, vol. 65, no. 3, pp. 512–534, 2015.

[26] Y. Kortli, M. Jridi, A. Al Falou, and M. Atri, "Face recognition systems: A survey", *Sensors*, vol. 20, no. 2, p. 342, 2020.

[27] M. W. Kraus and B. Callaghan, "Noblesse oblige? social status and economic inequality maintenance among politicians", *PloS one*, vol. 9, no. 1, e85293, 2014.

[28] Y. Li and Y. Xie, "Is a picture worth a thousand words? an empirical study of image content and social media engagement", *Journal of Marketing Research*, vol. 57, no. 1, pp. 1–19, 2020.

[29] X. Liu, R. Li, W. Ji, and T. Lin, "Towards robust multi-modal reasoning via model selection", *arXiv preprint arXiv:2310.08446*, 2023.

[30] Y. Liu, H. Duan, Y. Zhang, B. Li, S. Zhang, W. Zhao, Y. Yuan, J. Wang, C. He, Z. Liu, *et al.*, "Mmbench: Is your multi-modal model an all-around player?" *arXiv preprint arXiv:2307.06281*, 2023.

[31] P. Lu, S. Mishra, T. Xia, L. Qiu, K.-W. Chang, S.-C. Zhu, O. Tafjord, P. Clark, and A. Kalyan, "Learn to explain: Multimodal reasoning via thought chains for science question answering", *Advances in Neural Information Processing Systems*, vol. 35, pp. 2507–2521, 2022.

[32] S. R. Moghaddam and C. J. Honey, "Boosting theory-of-mind performance in large language models via prompting", *arXiv preprint arXiv:2304.11490*, 2023.

[33] Y. Peng, "What makes politicians' instagram posts popular? analyzing social media strategies of candidates and office holders with computer vision", *The International Journal of Press/Politics*, vol. 26, no. 1, pp. 143–166, 2021.

[34] Y. Peng, Y. Lu, and C. Shen, "An agenda for studying credibility perceptions of visual misinformation", *Political Communication*, vol. 40, no. 2, pp. 225–237, 2023.

[35] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *et al.*, "Language models are unsupervised multitask learners", *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[36] C. Salmivalli, K. Lagerspetz, K. Björkqvist, K. Österman, and A. Kaukiainen, "Bullying as a group process: Participant roles and their relations to social status within the group", *Aggressive Behavior: Official Journal of the International Society for Research on Aggression*, vol. 22, no. 1, pp. 1–15, 1996.

[37] R. Shao, T. Wu, and Z. Liu, "Detecting and grounding multi-modal media manipulation", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6904–6913.

[38] N. A. Shelke and S. S. Kasana, "A comprehensive survey on passive techniques for digital video forgery detection", *Multimedia Tools and Applications*, vol. 80, pp. 6247–6310, 2021.

[39] A. Suhr, S. Zhou, A. Zhang, I. Zhang, H. Bai, and Y. Artzi, "A corpus for reasoning about natural language grounded in photographs", *arXiv preprint arXiv:1811.00491*, 2018.

[40] H. Tan, F. Dernoncourt, Z. Lin, T. Bui, and M. Bansal, "Expressing visual relationships via language", *arXiv preprint arXiv:1906.07689*, 2019.

[41] C.-C. Tao and E. P. Bucy, "Conceptualizing media stimuli in experimental research: Psychological versus attribute-based definitions", *Human Communication Research*, vol. 33, no. 4, pp. 397–426, 2007.

[42] R. Thakur and R. Rohilla, "Recent advances in digital image manipulation detection techniques: A brief review", *Forensic science international*, vol. 312, p. 110 311, 2020.

[43] F.-A. Vasluianu, T. Seizinger, and R. Timofte, "Wsrd: A novel benchmark for high resolution image shadow removal", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, Jun. 2023, pp. 1825–1834.

[44] H. Venkateswara, J. Eusebio, S. Chakraborty, and S. Panchanathan, "Deep hashing network for unsupervised domain adaptation", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017.

[45] D. Walker and E. Vul, "Hierarchical encoding makes individuals in a group seem more attractive", *Psychological Science*, vol. 25, no. 1, pp. 230–235, 2014.

[46] L. Wang, Y. Hu, J. He, X. Xu, N. Liu, H. Liu, and H. T. Shen, "T-sciq: Teaching multimodal chain-of-thought reasoning via large language model signals for science question answering", *arXiv preprint arXiv:2305.03453*, 2023.

[47] W. Wang, X.-L. Chen, and L.-F. Zhong, "Social contagions with heterogeneous credibility", *Physica A: Statistical Mechanics and Its Applications*, vol. 503, pp. 604–610, 2018.

[48] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module", in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[49] L. Wu, P. Liu, and Y. Zhang, "See how you read? multi-reading habits fusion reasoning for multi-modal fake news detection", in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, 2023, pp. 13 736–13 744.

[50] Y. Xie, G. Li, and M.-Y. Kan, "Echo: Event causality inference via human-centric reasoning", *arXiv preprint arXiv:2305.14740*, 2023.

[51] M. Xiong, Z. Hu, X. Lu, Y. Li, J. Fu, J. He, and B. Hooi, "Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms", *arXiv preprint arXiv:2306.13063*, 2023.

[52] Y. Yan, W. Ren, and X. Cao, "Recolored image detection via a deep discriminative model", *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 1, pp. 5–17, 2018.

[53] Z. Yang, L. Li, J. Wang, K. Lin, E. Azarnasab, F. Ahmed, Z. Liu, C. Liu, M. Zeng, and L. Wang, "Mm-react: Prompting chatgpt for multimodal reasoning and action", *arXiv preprint arXiv:2303.11381*, 2023.

[54] H. You, R. Sun, Z. Wang, K.-W. Chang, and S.-F. Chang, "Find someone who: Visual commonsense understanding in human-centric grounding", *arXiv preprint arXiv:2212.06971*, 2022.

[55] A. Zadeh, M. Chan, P. P. Liang, E. Tong, and L.-P. Morency, "Social-iq: A question answering benchmark for artificial social intelligence", in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 8799–8809.

[56] A. Zeng, M. Attarian, B. Ichter, K. Choromanski, A. Wong, S. Welker, F. Tombari, A. Purohit, M. Ryoo, V. Sindhwani, *et al.*, "Socratic models: Composing zero-shot multimodal reasoning with language", *arXiv preprint arXiv:2204.00598*, 2022.

[57] F. Zhang, B. Zhou, C. Ratti, and Y. Liu, "Discovering place-informative scenes and objects using social media photos", *Royal Society open science*, vol. 6, no. 3, p. 181 375, 2019.

[58] X. Zhao, Y. Xie, K. Kawaguchi, J. He, and Q. Xie, "Automatic model selection with large language models for reasoning", *arXiv preprint arXiv:2305.14333*, 2023.

[59] W. Zhong, R. Cui, Y. Guo, Y. Liang, S. Lu, Y. Wang, A. Saied, W. Chen, and N. Duan, "Agieval: A human-centric benchmark for evaluating foundation models", *arXiv preprint arXiv:2304.06364*, 2023.

[60] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. Corso, and J. Gao, "Unified vision-language pre-training for image captioning and vqa", in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, 2020, pp. 13 041–13 049.

[61] D. Zhu, J. Chen, K. Haydarov, X. Shen, W. Zhang, and M. Elhoseiny, "Chatgpt asks, blip-2 answers: Automatic questioning towards enriched visual descriptions", *arXiv preprint arXiv:2303.06594*, 2023.

[62] X. Zhu, Y. Qian, X. Zhao, B. Sun, and Y. Sun, "A deep learning approach to patch-based image inpainting forensics", *Signal Processing: Image Communication*, vol. 67, pp. 90–99, 2018.

[63] Y. Zhu, X. Shen, and R. Xia, "Personality-aware human-centric multimodal reasoning: A new task", *arXiv preprint arXiv:2304.02313*, 2023.

[64] R. Zinko, G. R. Ferris, S. E. Humphrey, C. J. Meyer, and F. Aime, "Personal reputation in organizations: Two-study constructive replication and extension of antecedents and consequences", *Journal of Occupational and Organizational Psychology*, vol. 85, no. 1, pp. 156–180, 2012.

Appendix A

Examples of Prompts

We display the prompts of each module within our pipeline with Instruct-GPT LLM and BLIP MLLM. For the pipeline with alternative backbones, the prompts are slightly adjusted but mainly based on the following set of prompts. The bold italicized words represent the elements that can be substituted for different cases.

The prompt for inferring downstream tasks based on the detected manipulations for evaluation (We use four shots in practice. Here is a one-shot case.)
Help me to infer what the edits are trying to imply. The Edits are applied to the Source Image. If a famous person is the subject of the edit, then the focus should be on the implications related to that person, such as making fun of them. Besides, pay attention to any object exchanges within the image, like face or clothes swaps. The source image description is written after 'Source Image:', and the edits are written after 'Edits:'. Write your answer after 'Implications:'. Edits: The man in the source image is replaced with a man wearing sunglasses and riding an ostrich in the target image. The man in the source image is looking down with his head in the other hand, while in the target image, the man is smiling with his eyes closed. Source Image: The image depicts a man hanging from a high building with his feet on a rope holding a cat in one hand and looking down with his head in the other. Implications: This edit could potentially be used to make the man seem crazy, brave, and very talented, it looks like he's riding an ostrich which is dangerous. Edits: ***The source image shows a man wearing a suit, while the edited image shows a woman wearing a scarf around her neck.*** Source Image: ***The man is wearing a suit.*** Implications:

Table A.1: An example prompt for inferring downstream tasks based on detected manipulations for evaluation.

Context Selection (Zero-shot)
Extract the information related to ***the costume of the main subject*** in the picture from the given caption. The caption is written after 'Caption:'. Give your answer after 'Answer: '. Caption: ***a man in a suit sitting at a desk in front of a window with American flags and a phone in front of him on the table.*** Answer:

Table A.2: An example prompt for the Context Selection module.

Question Generation module in Visual Information Checking (VIC, Zero-shot)
I have a statement about the difference between the source image and the edited image. You can ask questions about ***the costume of the main subject*** in the source image and the edited image separately. Try to ask Yes/No questions, but you can also ask other questions if necessary. The information should be related to the two images' different parts in the statement. Tell me how to validate the difference. And if the statement is true, what are the expected answers to the questions? The statement is written after 'Statement'. Write the questions and expected answers for the source image after '##Source Image:##', and write the questions and expected answers for the edited image after'##Edited Image:##'. Write each question after 'Q:', and write the expected answers after 'EA:'. The questions and corresponding answers should be proposed one by one. Statement: ***The source image shows a man wearing a suit, while the edited image shows a woman wearing a scarf around her neck.***

Table A.3: An example prompt for the Question Generation module in VIC.

Edits Inference (Zero-shot)
Tell me the difference in ***the costume of the main subject*** in the source image and the edited image, according to the given captions of the source image and the edited image. If you can detect some differences, express them as how the source image is changed to the edited image. If you can't detect any difference, just return [[NO]], and don't imagine. The caption of the source image is written after 'Source:' and the caption of the edited image is written after 'Edited'. Label the edits with numbers like '1.', '2.'. Source: ***The man is wearing a suit.*** Edited: ***The woman is wearing a scarf around her neck.***

Table A.4: An example prompt for the Edits Inference module.

Details Supplementing (Zero-shot)

I have captions of two images, but the information related to ***the costume of the main subject*** is not sufficient enough. Help me to generate a question, whose answer supplements more details and and information about ***the costume of the main subject***. Don't return Yes or No questions. Don't return questions that have already been asked. The asked questions are written after 'Log:'. The caption of the source image is written after 'Source Image:', and the caption of the edited image is written after 'Edited Image: '. Return your generated question after 'Question: '.
Source Image: ***The man is wearing a suit.***
Edited Image: ***The woman is wearing a scarf on her head, and a scarf around her neck.***
Log: ***What kind of clothes is the main subject wearing?***
Question:

Table A.5: An example prompt for the Detail Supplementing module.

Consistency Judgement in Visual Information Checking (Zero-shot)

I have a question, a reference answer and a candidate answer. Help me to judge whether the candidate answer matches the reference answer or not. If the reference answer is correct, return [[YES]]. If the reference answer is not correct, return [[NO]]. For Yes/No questions, just focus on yes or no. Focus on the important part of the answers, and you can ignore some tiny differences in details. The question is written after 'Question:', the reference answer is written after 'Reference Answer:', and the candidate answer is written after 'Candidate Answer:'. Write your answer after 'Answer:'.
Question: ***Is the main subject wearing a suit?***
Reference Answer: ***Yes.***
Candidate Answer: ***Yes.***
Answer:

Table A.6: An example prompt for the Consistency Judgement module in VIC.