

Supplementary Materials: MM-Forecast

Anonymous Authors

1 ADDITIONAL EXPERIMENTS

To illustrate the limitations of existing MLLMs in the task of temporal event forecasting, we also conduct experiments using the Gemini-1.0-Pro-Vision model directly with images as sub-events. The prompt is consistent with the unimodal approach, except that the image associated with each sub-event is embedded directly into the prompt. In addition to the ICL-based methods detailed in the paper, we also compare the RAG-based methods. As show in the Table 1, a similar trend emerges as with the ICL-based methods, where the input of multiple images results in a degradation of prediction performance. This finding further illustrates the challenge existing MLLMs face in performing effective multimodal event forecasting.

2 DATASET CONSTRUCTION

In this section, we provide a comprehensive description of the MidEast-TE-mm dataset construction pipeline. This includes details on the data sources utilized to construct the dataset, the dataset construction pipeline, the dataset statistics and the evaluation settings of our datasets.

2.1 Data Source

We build the dataset based on a subset of MidEast-TE. The original MidEast-TE dataset includes structured atomic events and news articles, on this basis we added news images matching news articles as multi-modal information. Therefore, we name our dataset as MidEast-TE-multimodal, short as MidEast-TE-mm. Given the large scale of MidEast-TE, we sample a subset of complex events from it and used relevant news articles as the original documents to construct our dataset. Specifically, the time spans of the complex events in MidEast-TE are of high divergence, ranging from several days to more than three months. To reduce the potential bias introduced by outliers, we just sample 120 complex events whose time spans are within 40-60 days. After the sampling, we input the title of the news articles into Google Image, downloading the relevant images by ranking order with our own crawler script. As some images are of bad quality, we will analyze the size of the image and filter out images with small data size, which are generally of poor resolution.

2.2 Construction Pipeline

Sub-event Extraction. Due to the difference in data representation, we perform structured event extraction and unstructured event extraction by LLMs, separately. First, based on the original dataset approach and the three-layer structure of CAMEO ontology, we conduct a hierarchical sub-event extraction for structured data. Each next level of event extraction is based on the results of the prior level. This avoids the cost and performance degradation that could result from massive event definitions in the prompt. Subsequently, we summarize the news article to generate unstructured sub-events. We design the prompt to ensure the summarized sub-events are accurate, comprehensive and independent in terms of the content selection, descriptive way employed, and interrelationships among the sub-events.

Entity Linking. For structured sub-events, due to the absence of a predefined entity set in our event extraction process, entities are duplicated and formatted diversely. Subsequently, we employ GPT-4 for entity linking. The initial entity set for linking is relatively small, comprising only tens of thousands of entities, making the cost of using GPT-4 negligible. Initially, we apply K-means clustering to group all the original entities into multiple clusters. Then, we batch-process the entities within each cluster and request GPT-4 to perform entity linking. To ensure potential identical entities have a chance to be inputted into GPT-4 within the same batch, we iteratively conduct two rounds of such clustering followed by entity linking.

Image Filtering. Though our images are the most news-related on Google Images, many of them are not actually related to the news articles. For example, the icon of the media publisher of the news article or the poster of a magazine. In order to filter out irrelevant images, we instruct the Gemini-1.0-Pro-Vision model to determine the relevance of images to news articles. We give three options: highlighting, complementary and irrelevant. Highlighting means that the iamge and the content of the news are highly matched, and complementay means that the image has supplementary meaning to the content of the news. Images beyond these two relationships are regarded as irrelevant. We further remove images judged to be irrelevant to the news article.

2.3 Dataset Statistics

Table 2 and Table 3 detail the statistics of our curated dataset MidEast-TE-mm. We split the dataset into train, validation, and test sets in a temporal manner. Specifically, we use the sub-events in the last year for testing, the second-to-last year for validation, and the rest about five years for training. Note that the number of sub-events is not the same as the number of documents, as a single news document may correspond to multiple sub-events.

2.4 Human Evaluation

Due to the specialized format of structured sub-events, their extraction is more difficult than the extraction of unstructured sub-events. In order to evaluate the quality of the structured sub-events, we conduct human evaluation. Specifically, given a news article and the extracted sub-events, we ask the evaluator to tell whether the sub-events are valid or not based on the news article. We randomly sample 20 documents from the document set of the dataset, and conduct human evaluation based on the following criteria:

- Time: the sub-events extracted are events that have already occurred or are currently happening, rather than future events.
- Relation: the extracted sub-events appear in the original text and faithfully reflect the semantics of the original text.
- Entity: the extracted entities are concrete and real-world entities, and they appear in the original news article.

To be noted, for every extracted sub-event, we evaluate it once by going through the three criteria sequentially. For example, if we identify the *Time* is incorrect, we stop the checking of the following

Table 1: Performance (accuracy) comparison between using images directly and our methods in both settings of object entity prediction and relation prediction.

Model Type/Backbone	Forecasting Model	Multimodal Model	Object Entity Prediction		Relation Prediction	
			Text	Graph	Text	Graph
Gemini-1.0-Pro-Vision ³	ICL [1]	MLLM ³	0.3023	0.3319	0.5541	0.6085
	RAG [3]	MLLM ³	0.3305	0.3465	0.5769	0.5848
Gemini-1.0-Pro ³	ICL [1]	Uni-modal	0.3312	0.3657	0.5900	0.6257
		MM-Forecast (ours)	0.3527	0.3837	0.6087	0.6324
	RAG [3]	Uni-modal	0.3340	0.3669	0.6081	0.5866
		MM-Forecast (ours)	0.3425	0.3692	0.6121	0.5991

Table 2: Statistics of our curated dataset MidEast-TE-mm in structured data.

Dataset	#sub-events	#CEs	Image-H	Image-C	#docs
train	8,999	88	2,528	5,059	2,647
val	1,777	19	531	811	473
test	1,766	18	662	911	572
total	12,542	120	3,721	6,781	3,692

Table 3: Statistics of our curated dataset MidEast-TE-mm in unstructured data.

Dataset	#sub-events	Image-H	Image-C	#docs
train	20,194	4,464	3,407	2,647
val	3,715	832	578	473
test	4,231	1,137	488	572
total	28,140	6,433	4,473	3,692

Table 4: Error analysis of the sub-event extraction results in different datasets.

Dataset	#atomic events	Acc.(%)	error type (%)		
			time	relation	entity
GDELT-TE [2]	148	29.73	3.85	31.73	64.42
MidEast-TE [4]	35	55.56	0	92.86	7.14
MidEast-TE-mm(ours)	78	72.60	16.67	27.78	55.56

Table 5: The accuracy of identified relations between image and text.

Data-Type	GPT-4-Vision		Human	
	Text	Graph	Text	Graph
Highlighting	0.68	0.68	0.73	0.83
Complementary	0.88	0.93	0.87	0.86

two criteria. Table 4 shows the accuracy and error types of the sub-event extraction results, including both our dataset MidEast-TE-mm and previous datasets of MidEast-TE and GDELT-TE. To be noted, compared to the previous two datasets, we employ the state-of-the-art large language model for sub-event extraction. Clearly, the dataset constructed by our pipeline is the most accurate one. In our dataset, there are the fewest extraction errors, which makes the following forecasting research more valid and reliable. However,

we admit that an overall accuracy of 72.6% is still unsatisfactory in practice, and further efforts will be devoted to improved event extraction performance. In addition, in order to deeply verify the correctness of our judgement on the relationship between images and news, we also made a manual judgement on it. The results are shown in the Table 5, which are similar to the judgement of GPT4V.

3 PROMPTS: IMAGE FUNCTION

In this section, we show all the prompts that need to be used in the image function identification module. As show in Table 6, the first row is the prompt for image function recognition, which is mainly from the perspective of the subject background and the specific event to judge the function of the image. The last two rows are the prompts of the different functions of the images to achieve their respective functions and transform their information into verbal descriptions. Eventually, the verbal information will be integrated into the LLM-based event forecasting model.

4 CASE STUDY: IMAGE FUNCTION

To further illustrate that our approach does indeed identify truly key events and the required complementary information, we provide additional examples. In the first example of the highlighting function, the image directly depicts Ocasio-Cortez, with the background appearing to be the Congressional sites, thereby emphasizing the relevant key event. Correspondingly, the key event also mentions the relationship between Congress and Ocasio-Cortez. Consequently, an accurate prediction is achieved. In the second example of the highlighting function, the key event highlighted by the image directly mentions the disqualification of Ali Larijani from the election, which perfectly aligns with the results that need to be predicted and the information provided to present those results. For the first example of complementary functions, the image provides information about the signing of a free trade agreement between Turkey and the United Kingdom. While enhanced trade has the potential to lead to employment and economic growth, the image offers complementary information on the role of labor. Therefore, an accurate forecast is achieved. In the second example about the complementary function, the image shows Bernie Sanders who is a democratic progressive socialist like Ocasio-Cortez. They share many commonalities and connections to Congress, which can provide supplementary information to more accurately predict the outcome. Through these examples, the distinct functions of highlighting key events and providing complementary information

Table 6: Prompts of image function identification module.

Identification	You are a professional news writer.
	Please judge the relationship between images and news based on the following rules:
	1. Final judgment please choose between [highlighting, complementary, irrelevant].
	2. The relationship between an image and a news article is highlighting if the image’s subject matter and depicted event are highly related to the news and the specific event shown in the image is already mentioned in detail in the article’s description.
	3. The relationship between an image and a news article is complementary if the image’s overall theme and background information are highly related to the news, but the specific event depicted in the image is not mentioned in detail in the article, and the visual information in the image can complement the news story as a whole.
	4. Except in cases where the relationship is highlighting or complementary, in other cases, the relationship between the image and the text is irrelevant.
Highlighting	You are a professional news writer.
	Please determine which sub-event in the news the image is most relevant to based on the following rules:
	1. For the final judgement, please answer with the serial number of the sub-event. For example: [The number of the sub-event most relevant to the image is 1.]
	2. Identify the main subjects or objects prominently featured in the image. Sub-events that provide details, background information or context directly about these central visual elements are highly relevant.
	3. If people are depicted, identify who those individuals are. Sub-events involving those particular people should take priority.
	4. Analyze the overall activities, actions, emotions or mood being portrayed in the image. Relevant sub-events likely delve into similar situations, occurrences or sentiments illustrated.
	5. Take note of the specific location, setting or environment depicted in the image. Prioritize sub-events that discuss that geographic area, type of place, or related events.
	6. Look for any text, logos, labeled items or signs visible in the image content. Sub-events elaborating on the organizations, companies, products or public figures represented by those texts are applicable.
Complementary	You are a professional news writer.
	Please extract the image information according to the following rules based on the content of the provided news:
	1. Extract the image information as a sub-event. Instead of multiple sub-events.
	2. The phrases: [In the image], [The image shows], [In the picture], [The image is], [In the photo], etc, should never appear in the summarised sub-event.
	3. Identify the primary focus or subject of the image that represents the core piece of information being conveyed. This main subject should serve as the central point around which the image information is extracted.
	4. Directly relate the extracted image information to the associated news event covered in the article. The image summary should complement and enhance the understanding of the news content, not introduce unrelated information.
	5. Prioritize and emphasize the most newsworthy and significant details visible in the image. These could include specific actions, emotions, or identifying characteristics of the main subject.
	6. Ensure that all information included in the image summary originates directly from the provided image and news article. Avoid introducing fabricated content, speculative details.
	7. Aim for a succinct summary, using clear and straightforward language. Avoid excessive detail or subjective commentary.
	8. Maintain an objective and impartial tone when describing the image. Avoid inserting personal opinions or interpretations.

are elucidated, substantiating the effectiveness of our approach in leveraging multimodal information for accurate temporal event forecasting.

REFERENCES

[1] Dong-Ho Lee, Kian Ahrabian, Woojeong Jin, Fred Morstatter, and Jay Pujara. 2023. Temporal Knowledge Graph Forecasting Without Knowledge Using In-Context Learning. In *EMNLP*. Association for Computational Linguistics, 544–557.

[2] Kalev Leetaru and Philip A Schrodt. 2013. Gdelt: Global data on events, location, and tone, 1979–2012. In *ISA annual convention*, Vol. 2. Citeseer, 1–49.

[3] Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *NeurIPS*.

[4] Yunshan Ma, Chenchen Ye, Zijian Wu, Xiang Wang, Yixin Cao, Liang Pang, and Tat-Seng Chua. 2023. Structured, Complex and Time-complete Temporal Event Forecasting. *CoRR* abs/2312.01052 (2023).

Figure 1: The case study of *highlighting* function of image.

User:

Before Highlighting

[Query]: (The Guardian Council, Veto, 2289)

[Related events]:

[Date]1828:

- * Iranians voted on Friday to elect a new parliament.
- * Many Iranians chose to abstain from voting due to disappointment with the government and its hollow promises.
- * A survey by the Institute for Social Studies at Tehran University in early February indicated that less than one out of four Iranians in Tehran would vote, in contrast to a 62% turnout in 2016.
- * Women activists in Iran called for an election boycott to protest the regime's brutal handling of demonstrators.
- * Iranians have faced economic hardship due to US sanctions and the government's "maximum resistance" policy.
- * Iranian security forces brutally cracked down on protests against the economic crisis, with parliamentarians remaining silent.
- * The Iranian regime needs legitimacy, and a high turnout is crucial for Khamenei to interpret as a sign of trust in the political system.
- * The Guardian Council barred almost 9,000 candidates from participating in the polls, including 92 incumbent members of parliament, mostly reformist politicians.

[Date]2273:

- * Iran opened the door for candidacies for the upcoming presidential vote.
- * President Hassan Rouhani rejected the decision of the Guardian Council to study the candidates' applications.
- * The Guardian Council specified that "all nominees must be between 40 and 70 years of age, hold at least a master's degree or its equivalent, have experience of at least four years in managerial posts... and have no criminal record".
- * The new terms come in implementation of a 2016 directive from Iran's supreme leader Ali Khamenei for the council to clarify and "determine" the requirements.
- * The updated requirements are likely to exclude some well-known figures, such as Rouhani's telecom minister, Mohammad Javad Azari Jahromi, who is 39, or the elite Revolutionary Guards' Saeed Mohammad, whose rank is below major general.
- * The increasing number of Revolutionary Guard generals who have announced their candidacy or are likely to be intending to compete in the presidential race, have raised fears of the "militarization" of the political circle within the Iranian regime.

[Date]2278:

- * Iran's election in June will determine the next president and possibly the successor of Supreme Leader Ali Khamenei.
- * Khamenei, who has been in power since 1989, is 81 years old and has several health problems.
- * The position of supreme leader is critical for the survival of the Islamic Republic and the continuation of Ayatollah Khomeini's revolutionary concept of Velayat-e Faqih.
- * The supreme leader has the final say on Iran's domestic and foreign policy and significant control over the legislative, executive and judicial systems.
- * The supreme leader also controls the IRGC, intelligence service, bonyad charitable trusts, and Setad organization.
- * The position of president can be a crucial stepping stone to succeeding the supreme leader.

[Options]:

A: Vienna Talks B: Khamenei C: Ali Larijani D: Hemmati E: Mahmoud Ahmadinejad F: Mohsen Mehralizadeh

Response:

B: Khamenei



After Highlighting

User:

[Query]: (The Guardian Council, Veto, 2289)

[Key events]:

[Date]2287:

- * Moderate conservative Ali Larijani was disqualified, potentially boosting ultraconservative judiciary chief Ebrahim Raisi's chances of winning

[Related events]:

[Date]1828:

- * Iranians voted on Friday to elect a new parliament.
- * Many Iranians chose to abstain from voting due to disappointment with the government and its hollow promises.
- * A survey by the Institute for Social Studies at Tehran University in early February indicated that less than one out of four Iranians in Tehran would vote, in contrast to a 62% turnout in 2016.
- * Women activists in Iran called for an election boycott to protest the regime's brutal handling of demonstrators.
- * Iranians have faced economic hardship due to US sanctions and the government's "maximum resistance" policy.
- * Iranian security forces brutally cracked down on protests against the economic crisis, with parliamentarians remaining silent.
- * The Iranian regime needs legitimacy, and a high turnout is crucial for Khamenei to interpret as a sign of trust in the political system.
- * The Guardian Council barred almost 9,000 candidates from participating in the polls, including 92 incumbent members of parliament, mostly reformist politicians.

[Date]2273:

- * Iran opened the door for candidacies for the upcoming presidential vote.
- * President Hassan Rouhani rejected the decision of the Guardian Council to study the candidates' applications.
- * The Guardian Council specified that "all nominees must be between 40 and 70 years of age, hold at least a master's degree or its equivalent, have experience of at least four years in managerial posts... and have no criminal record".
- * The new terms come in implementation of a 2016 directive from Iran's supreme leader Ali Khamenei for the council to clarify and "determine" the requirements.
- * The updated requirements are likely to exclude some well-known figures, such as Rouhani's telecom minister, Mohammad Javad Azari Jahromi, who is 39, or the elite Revolutionary Guards' Saeed Mohammad, whose rank is below major general.
- * The increasing number of Revolutionary Guard generals who have announced their candidacy or are likely to be intending to compete in the presidential race, have raised fears of the "militarization" of the political circle within the Iranian regime.

[Date]2278:

- * Iran's election in June will determine the next president and possibly the successor of Supreme Leader Ali Khamenei.
- * Khamenei, who has been in power since 1989, is 81 years old and has several health problems.
- * The position of supreme leader is critical for the survival of the Islamic Republic and the continuation of Ayatollah Khomeini's revolutionary concept of Velayat-e Faqih.
- * The supreme leader has the final say on Iran's domestic and foreign policy and significant control over the legislative, executive and judicial systems.
- * The supreme leader also controls the IRGC, intelligence service, bonyad charitable trusts, and Setad organization.
- * The position of president can be a crucial stepping stone to succeeding the supreme leader.

[Options]:

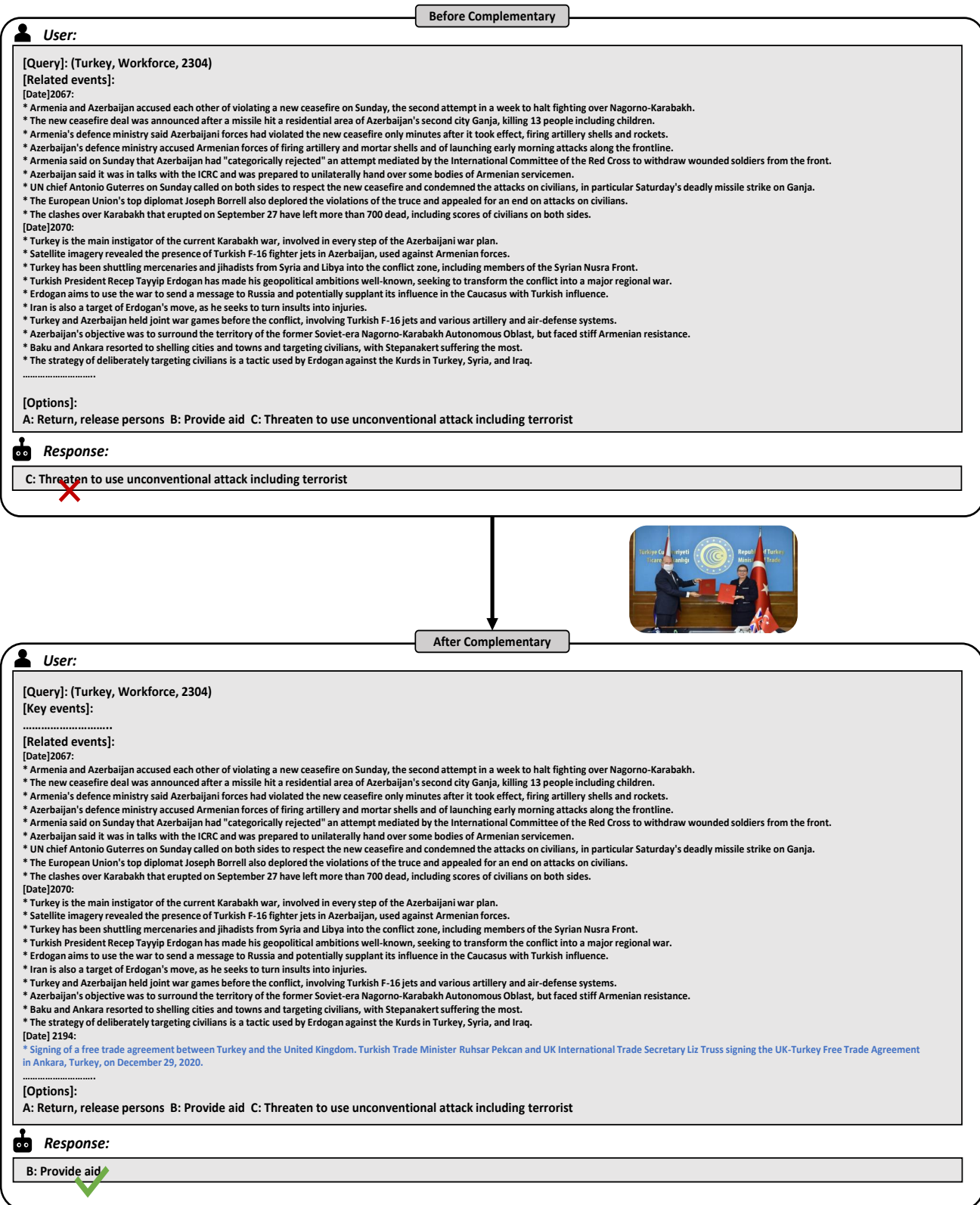
A: Vienna Talks B: Khamenei C: Ali Larijani D: Hemmati E: Mahmoud Ahmadinejad F: Mohsen Mehralizadeh

Response:

C: Ali Larijani



Figure 2: The case study of *highlighting* function of image.

Figure 3: The case study of *complementary* function of image.

