Supplementary Material For Paper: Enhancing Weakly Correlated Multimodal Sentiment Analysis with Meta-knowledge

Anonymous submission

Experimental Setup

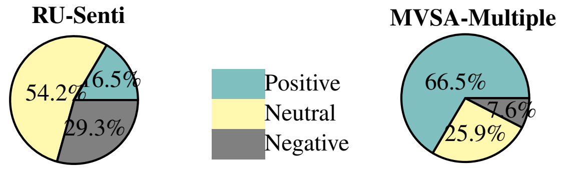
During the training process, we select the Adam optimizer (Kingma and Ba 2014) with a learning rate of 5e-4, which is decreased every 10 epochs with a decay rate of 0.1. We train our model by setting the number of epochs, dropout rate, and batch size to 60, 0.1, and 64 respectively. All our scores are the averages of runs with different random seeds for more than five times. In Table A1, we summarize the hyperparameter settings.

Table A1: Details of the hyper-parameter settings.

Param.	Value
Leaning rate	5e-4
Batch size	64
Epoch size	60
Hidden size	768
Dropout	0.2
Text factor (α)	0.7
CPU	Intel i9
GPU	NVIDIA A100

Datasets Details

We evaluate the effectiveness of our framework on the widely used weakly correlated multimodal sentiment analysis dataset RU-Senti (Liu et al. 2024) and the traditional multimodal sentiment analysis dataset MVSA-Multiple (Niu et al. 2016). We present the specific distributions of sentiment categories on the RU-Senti and MVSA-Multiple datasets in Figure A1. The sentiment categories in both datasets are divided into three types: positive, neutral, and negative. Regarding the RU-Senti dataset, the neutral sentiment category takes up the largest share, making up 54.2%. As for the MVSA-Multiple dataset, the positive sentiment category occupies the highest proportion, reaching 66.5%. To comprehensively evaluate the performance of the model, we randomly partition each dataset into a training set, a validation set, and a test set in an 8:1:1 ratio.



Dataset	Total	Positive	Neutral	Negative	Avg Length
RU-Senti	113,588	18,715	61,553	33,320	19.99
MVSA-Multi	17,024	11,318	4,408	1,298	12.34

Figure A1: The data distribution of the RU-Senti dataset and MVSA-Multiple dataset.

The Effect of Different Prompt Template

We are curious about the impact of different prompt templates for generating knowledge on multimodal sentiment analysis. In Figure A2, we conduct experiments using prompt templates generated by ChatGPT[1], Llama 3, and human designed ones respectively. To ensure the consistency and comparability of the experiments, we provide the same prompt to human annotators and large language models: "Please provide a prompt template for a large vision language model to generate historical knowledge based on the information of images and texts." We observe that in all metrics of the two datasets, the prompt templates generated by large language models are significantly superior to the human-designed ones. Among them, the prompt templates generated by ChatGPT achieve the most outstanding results. Upon in-depth exploration of the reasons, it is highly likely that ChatGPT, with its powerful language understanding and generation capabilities, can generate templates that provide more in-depth analysis. These templates can accurately reflect the important events, meanings, and rich background information behind the images and texts. Such detailed and comprehensive information provides a solid foundation for multimodal sentiment analysis, helping the model capture

[1] https://chat.openai.com

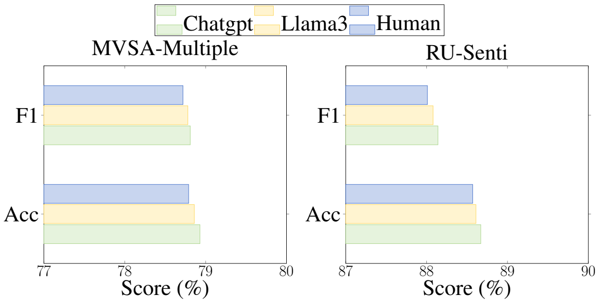


Figure A2: Experimental results of different prompt templates.

Table A2: Experimental results of different LLMs for knowledge generation.

	MVSA-Multiple	RU-Senti
LLaVA	78.81	88.14
Video-LLaMA3	78.79	88.09
Qwen2.5-VL	78.73	88.02

sentiment cues more accurately and thus enhancing the accuracy of sentiment analysis. Human designed prompt templates often seek historical insights directly from the surface information of images and texts, which have certain limitations in the depth and breadth of information. When designing templates, humans may be influenced by factors such as knowledge reserves and mental sets, making it difficult to explore potential historical information as comprehensively and deeply as large language models. This leads to a slight deficiency in the accuracy and comprehensiveness of multimodal sentiment analysis based on human designed templates.

The Effect of Knowledge Generated by Different LLMs

To address concerns about dependency on specific LLM capabilities, we conduct comparative experiments using three distinct vision-language LLaVA, Video-LLaMA3, and Qwen2.5-VL as knowledge generators, evaluating their impact on final sentiment prediction performance. Results are shown in Table A2. It demonstrates that across all tested LLMs, our Meta-MSA achieves nearly identical performance on both datasets. This minimal variance indicates that our framework's effectiveness is not tied to the specific capabilities of a single LLM, but rather to the mechanism of leveraging context-relevant background associations—regardless of the LLM used to generate them.

The Effect of Different Disentanglement Methods

We are curious about the advantages of the multiple disentanglement strategy on the model's performance. In Figure A3, we compare no disentanglement (None), dual disentanglement (which disentangles modality features into modality-invariant features and modality-specific features, Dual), and our multiple disentanglement (Ours). We find that on both datasets, the model exhibits the best performance when our multiple disentanglement strategy is adopted. When dual disentanglement is employed, the performance of the model decreases because this approach overlooks the effective processing of information within modality-specific features that are irrelevant to sentiment analysis. For the RU-Senti dataset, there is a 1.18% reduction in the model's Accuracy and a 0.61% drop in the F1-score. On the MVSA-Multiple dataset, the model's accuracy and F1-score decrease by 1.39% and 1.14% respectively. This indicates that in the process of processing weakly correlated multimodal sentiment analysis data, simply distinguishing between modality-invariant features and modality specific features is insufficient. The unprocessed irrelevant information interferes with the model's judgment, causing a decline in the model's performance. When the model operates without disentanglement, its performance is at its worst, as it fails to effectively distinguish and process features of different natures. In contrast, our multiple disentanglement mechanism can accurately screen out the information valuable for multimodal sentiment analysis. Meanwhile, it effectively excludes the information that may interfere with the model's judgment, thereby significantly enhancing the model's performance.

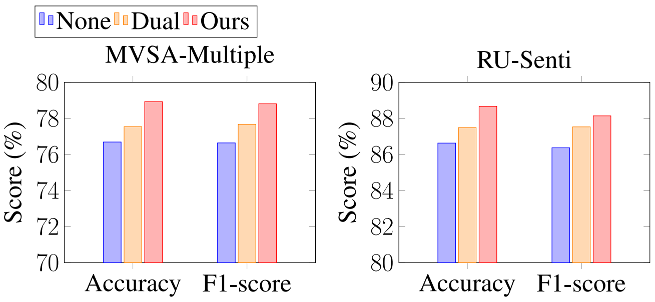


Figure A3: Comparative results of different disentanglement methods on MVSA-Multiple and RU-senti datasets.

The Effect of Disentangled Modality Features

To further validate the necessity of our multiple disentanglement strategy and the distinct roles of each modality-related feature, we conduct experiments by evaluating the model performance when using only one type of disentangled feature (modality-flag, modality-invariant, or modality-redundant features) in isolation. Results are presented in Table A3. It shows that all single-feature settings underperform our full model (Ours) across both datasets, with significant performance drops observed. Specifically, modality-redundant features yield the lowest results , while modality-flag and modality-invariant features perform moderately better. These findings underscore two key insights: (1) each disentangled feature contributes uniquely to the final prediction, and their synergistic integration (as in our full model) is critical for optimal performance; (2) modality-redundant features, which capture noise or irrelevant information, are

Table A3: Performance of individual disentangled modality features on MVSA-Multiple and RU-Senti datasets.

Feature Type	MVSA-Multiple	RU-Senti
Modality-flag	76.26	85.85
Modality-invariant	75.94	85.39
Modality-redundant	70.35	82.37
Ours	78.81	88.14

Table A4: Average inference time per sample on MVSA-Multiple and RU-Senti datasets.

Method	MVSA-Multiple	RU-Senti
TOM	183ms	159ms
Ours	197ms	172ms

particularly detrimental when used alone, highlighting the importance of our strategy to filter out such invalid features.

Computational Efficiency Analysis

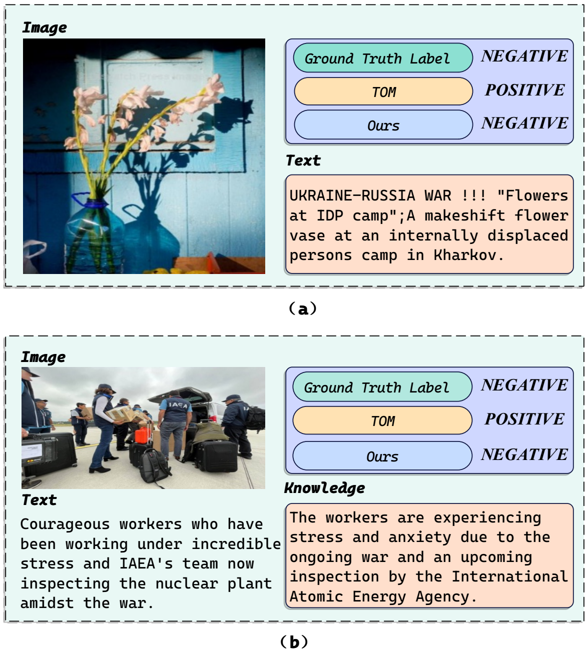
To evaluate the practicality of our method, we conduct additional experiments to compare the computational time overhead between our Meta-MSA and the state-of-the-art model TOM. Specifically, we measure the average inference time per sample on both MVSA-Multiple and RU-Senti datasets, with results summarized in Table A4. Table A4 reveals that our Meta-MSA incurs a moderate increase in inference time compared to TOM. Notably, our framework adopts an offline preprocessing strategy for the LLM-based meta-knowledge generation: meta-knowledge is batch-processed during the pre-training phase (186ms per sample on MVSA-Multiple and 146ms on RU-Senti) and stored in a structured format. This design avoids real-time LLM invocations during inference, thus eliminating excessive latency and ensuring the model remains applicable to practical scenarios.

Human Evaluation of Meta-Knowledge Quality

To quantitatively assess the reliability and effectiveness of the generated meta-knowledge, we conduct a human evaluation on 500 randomly selected samples. Three NLP experts with expertise in multimodal sentiment analysis independently rate the meta-knowledge across three key dimensions using a 5-point scale (1 = poor, 5 = excellent): 1)Fluency: The degree to which the generated knowledge is grammatically coherent and naturally expressed. 2)Relevance: The extent to which the knowledge is contextually aligned with the original multimodal content (text-image pairs). 3)Reliability: The accuracy and trustworthiness of the semantic information conveyed by the meta-knowledge. Table A5 shows that the meta-knowledge achieves high average scores across all dimensions. These consistent high ratings confirm that the generated meta-knowledge exhibits strong semantic complementation ability—effectively providing coherent, context-relevant, and reliable background information to enhance the understanding of weakly correlated multimodal data.

Table A5: Human evaluation results of meta-knowledge quality(average scores over 500 samples, 5-point scale).

	Average Score
Fluency	4.36
Relevance	4.15
Reliability	4.24



UKRAINE-RUSSIA WAR !!! "Flowers at IDP camp";A makeshift flower vase at an internally displaced persons camp in Kharkov.

(a)

Courageous workers who have been working under incredible stress and IAEA's team now inspecting the nuclear plant amidst the war.

The workers are experiencing stress and anxiety due to the ongoing war and an upcoming inspection by the International Atomic Energy Agency.

(b)

Figure A4: Case study of weakly correlated multimodal sentiment analysis.

Case Study

We conduct a case study by comparing our method with TOM to more intuitively demonstrate the advantages of our method in weakly correlated multimodal sentiment analysis. As presented in Figure A4, our method accurately identifies the sentiment polarity, while the TOM model makes incorrect predictions in both cases. In Figure A4 (a), the image shows blooming flowers, which, from a visual perspective alone, easily evoke "positive" sentiment associations. The TOM model, influenced by this, misjudges the sentiment as "positive". However, the textual content clearly conveys that it is during a war, with people displaced, cherishing life extremely, and full of "negative" sentiment such as hatred for the war between the lines. If the image and text information are simply fused directly, there would likely be a misjudgment of the overall sentiment as "positive". In contrast, our method employs unimodal auxiliary tasks to carefully capture and analyze the sentiment cues in the textual modality. It successfully captures the "negative" sentiment embedded therein, thereby enabling it to make correct predictions. In Figure A4 (b), the TOM model solely relies on the surface information of bravery in the image and text, as well as the topic information generated from the text, and mistakenly predicts the sentiment of this multimodal information as “positive”. However, the actual situation is more complex. Our method skillfully integrates the background knowledge of the Russia-Ukraine conflict. This knowledge indicates that currently, in the context of war, civilians are living under great threat. Through the understanding of this background knowledge, our method accurately predicts the “negative” sentiment of anxiety and restlessness among the civilians. This strongly demonstrates the crucial role of incorporating the implicit associative knowledge behind image-text pairs into sentiment analysis. In summary, these cases demonstrate the effectiveness of our model. By integrating meta-knowledge into the context and leveraging unimodal auxiliary tasks, our model can more thoroughly and comprehensively extract the sentiment cues from multimodal data. This enables it to effectively avoid misjudgments caused by surface level information or information from a single modality.

Error Analysis

In Figure A5 and A6, we conduct an error analysis of our proposed Meta-MSA framework, providing valuable insights for future improvements. We summarize two types of errors. The first error type mainly stems from inaccuracies in the knowledge generated by the model, which in turn misleads the sentiment judgment. The second error type predominantly occurs when dealing with text-image pairs where sentiment features are not obvious, and the model tends to mispredict the sentiment as “neutral”.

In Figure A5 (a), the image shows a scene where a little girl is receiving wound dressing, and the textual information describes that the injured girl sings the Ukrainian national anthem during the treatment. The knowledge generated by the model focuses on the information that “the war conflict has caused casualties among children”. Affected by this, the model determines the sentiment of this text-image pair as “negative”. However, the girl’s act of singing the national anthem and her optimistic expression convey strong patriotic sentiment. In fact, the sentiment of this text-image pair should be more towards the “positive” side. In this case, the knowledge generated by the model one sidedly emphasizes the fact that the girl is injured, ignoring the underlying “positive” sentiment factors, thus leading to an error in sentiment judgment.

Figure A5 (b) presents an image of Ukrainian troops on the march. The knowledge generated by the model is “On a sunny day, the Ukrainian troops set out towards victory”. Based on this, the model wrongly judges the sentiment of this text-image pair as “positive”. But in reality, these soldiers are in a war environment, far away from their hometowns and relatives, and constantly facing the dangers and uncertainties brought by the war. Their true sentiment tone should be “negative”. This case shows that the knowledge generated by the model fails to comprehensively consider the war situation that the soldiers are in and their complex inner feelings, thus resulting in a deviation in sentiment judgment.

Figure A6 (c) shows image-text information related to Russia’s electricity supply to Ukraine. The model, based solely on surface information, deems it as a simple statement and thus predicts the sentiment as “neutral”. However, this scene is set in a war environment. The electricity supply is affected by the war, and people will feel uneasy and worried about the future uncertainties. The true sentiment should be “negative”. This indicates that when dealing with such situations, the model fails to dig deep into the potential sentiment factors and only makes judgments based on surface information, leading to incorrect sentiment predictions.

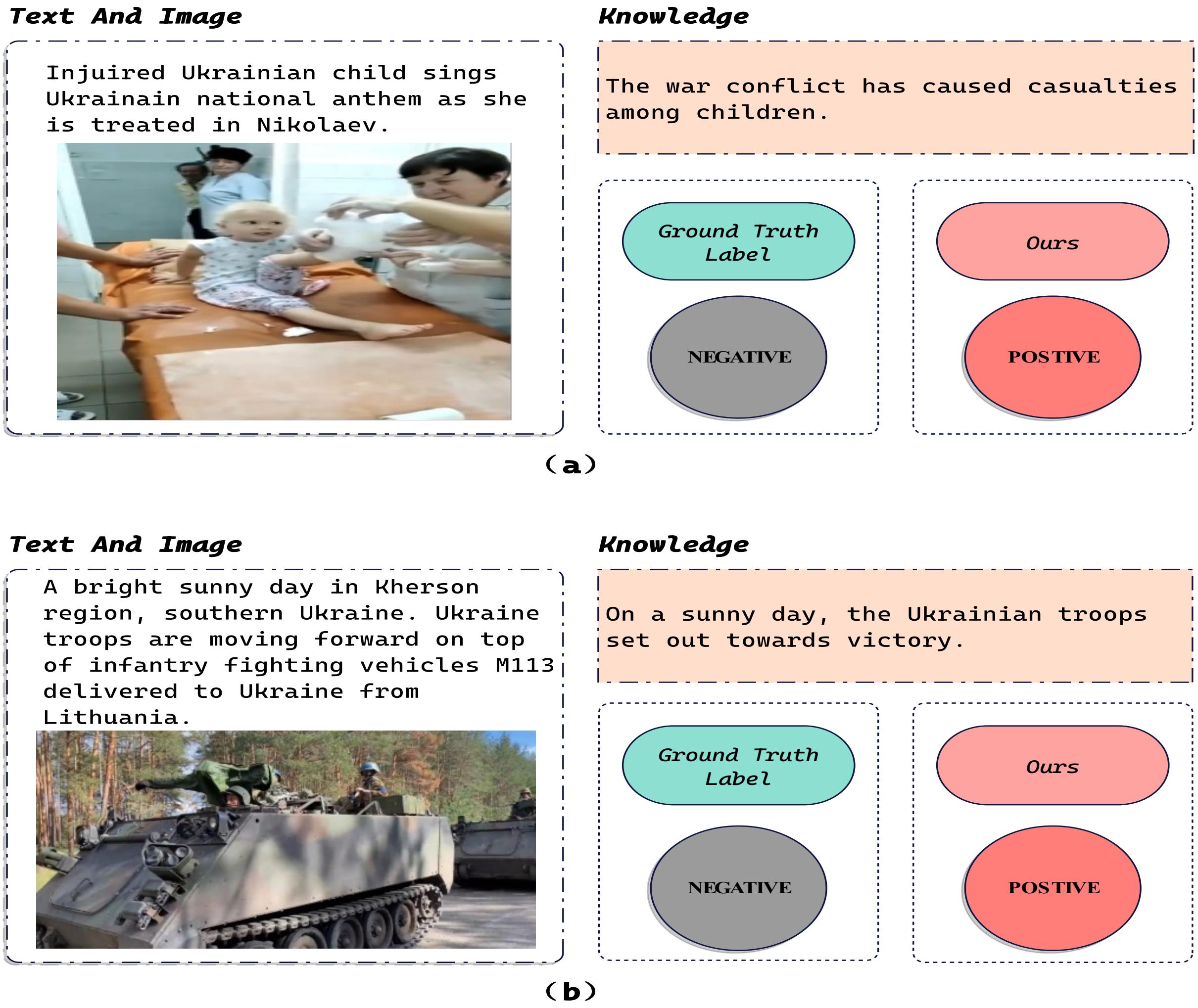
Figure A6 (d) presents an image of a group of citizens queuing up to receive winter supplies. Due to the relatively weak sentiment cues in the multimodal information, the model fails to effectively identify the sentiment tendency and wrongly determines the sentiment as “neutral”. It should be noted that this scene takes place during the Russia-Ukraine war. Behind the citizens’ act of queuing up for supplies actually lies their fear and helplessness towards the war, as well as the “negative” sentiment resulting from the shortage of living supplies and the turbulent life caused by the war. This reflects that the model lacks sufficient sensitivity and in-depth mining ability when dealing with text-image pairs with weak sentiment cues. In conclusion, in the future, it is necessary to further optimize the knowledge generation mechanism to make it more accurate and comprehensive. Meanwhile, the model’s ability to capture and analyze sentiment cues in complex situations should be enhanced to improve the accuracy of sentiment judgment.

References

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980.*

Liu, W.; Li, W.; Ruan, Y.-P.; Shu, Y.; Chen, J.; Li, Y.; Yu, C.; Zhang, Y.; Guan, J.; and Zhou, S. 2024. Weakly correlated multimodal sentiment analysis: New dataset and topic-oriented model. *IEEE Transactions on Affective Computing*.

Niu, T.; Zhu, S.; Pang, L.; and El Saddik, A. 2016. Sentiment analysis on multi-view social data. In *MultiMedia Modeling: 22nd International Conference, MMM 2016, Miami, FL, USA, January 4-6, 2016, Proceedings, Part II 22*, 15–27. Springer.



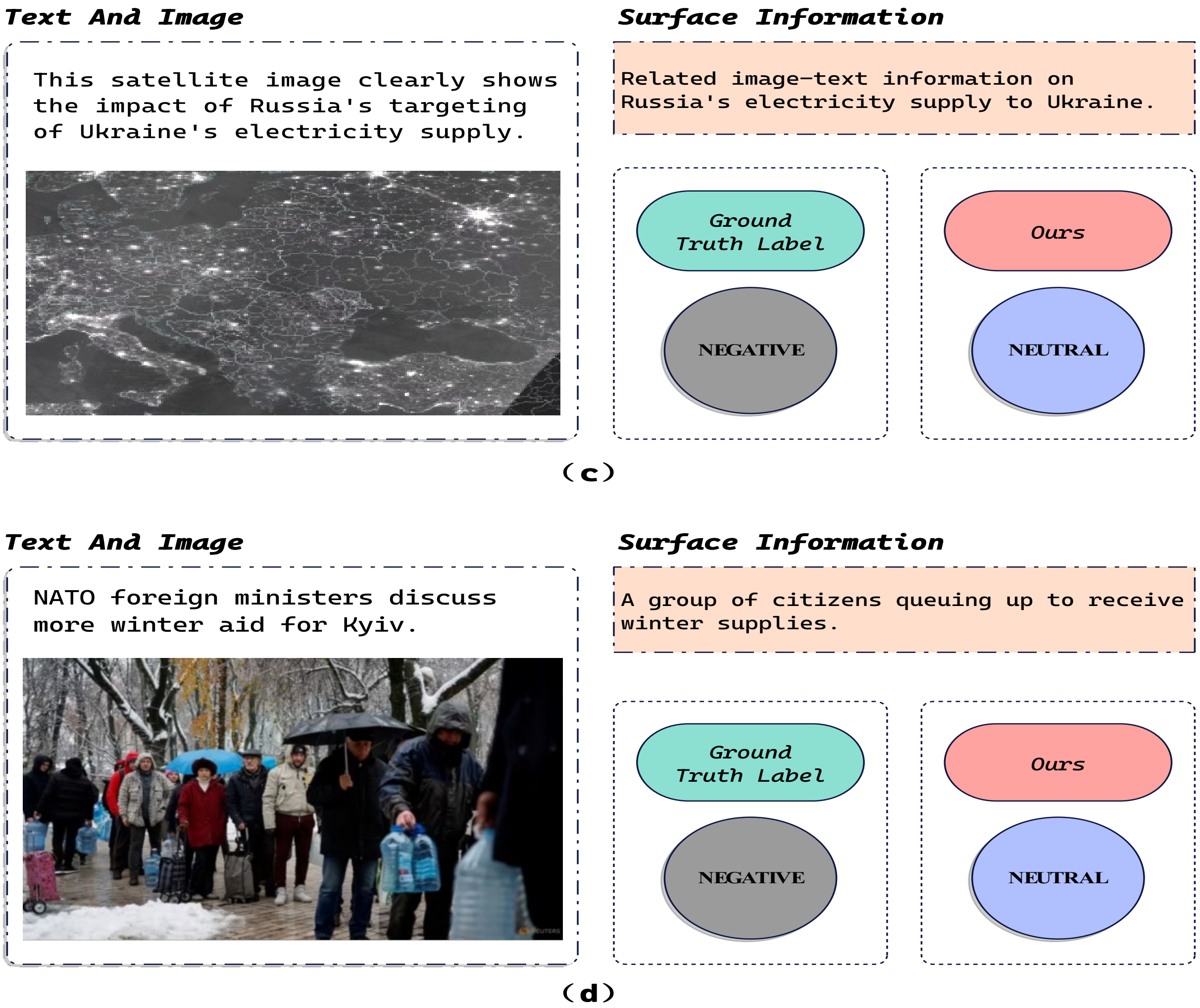
Injuired Ukrainian child sings Ukrainain national anthem as she is treated in Nikolaev.

The war conflict has caused casualties among children.

A bright sunny day in Kherson region, southern Ukraine. Ukraine troops are moving forward on top of infantry fighting vehicles M113 delivered to Ukraine from Lithuania.

On a sunny day, the Ukrainian troops set out towards victory.

Figure A5: Error analysis of our method.



This satellite image clearly shows the impact of Russia's targeting of Ukraine's electricity supply.

Related image-text information on Russia's electricity supply to Ukraine.

NATO foreign ministers discuss more winter aid for Kyiv.

A group of citizens queuing up to receive winter supplies.

Figure A6: Error analysis of our method.