

GETReason: Enhancing Image Context Extraction through Hierarchical Multi-Agent Reasoning

Shikhhar Siingh^{*1}, Abhinav Rawat^{*1}, Chitta Baral¹, Vivek Gupta¹

¹ Arizona State University ^{*}equal contribution





Motivation

Publicly significant images are rich in context

Need to go beyond the surface level visual parts of the image to **reason** this information out of the image

Existing frameworks fail to extract this contextual information

What is this image about?



Who are the people in this image?



Who are the people in this image?



Where could this image be from?



Jill Biden

President
Rene Preval

Michelle
Obama

Where could this image be from?



Jill Biden

President
Rene Preval

Michelle
Obama

Port-au-Prince, Haiti

What is this image about?



Jill Biden

President
Rene Preval

Michelle
Obama

Port-au-Prince, Haiti

What is this image about?



Visit of Michelle Obama and Jill Biden to Haiti
after the 2010 Haiti Earthquakes

What is this image about?



Event Visit of Michelle Obama and Jill Biden to Haiti
after the 2010 Haiti Earthquakes

What is this image about?



Jill Biden

President
Rene Preval

Michelle
Obama

Location

Event

Visit of Michelle Obama and Jill Biden to Haiti
after the 2010 Haiti Earthquakes

What is this image about?



Event Visit of Michelle Obama and Jill Biden to Haiti
Time after the 2010 Haiti Earthquakes



Problem Statement

Given an image, extract:



Location (Geospatial)



Time (Temporal)



Event (Socio Political significance)

→ Move beyond object recognition to real-world
reasoning

GETReason

Geospatial Event and Temporal Reasoning

Hierarchical **Multi-Agent** Framework

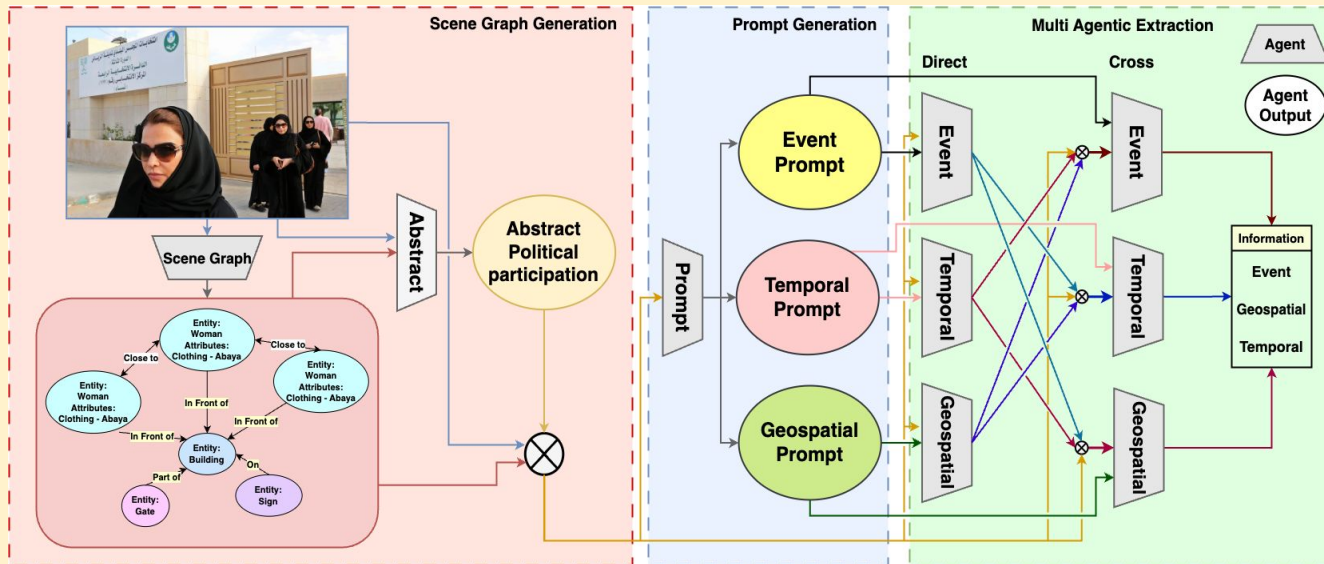
Structured Outputs

Prompt Engineering

GETReason

Geospatial Event and Temporal Reasoning

Architecture



Scene Graph
Generation

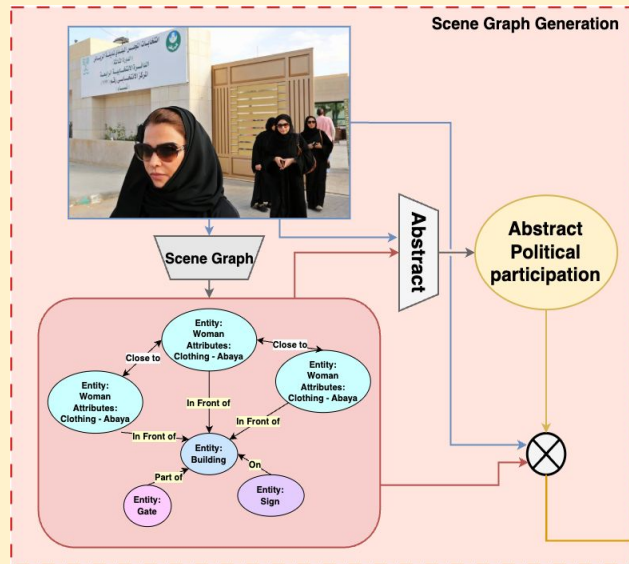
Prompt Generation

Multi Agentic
Extraction

GETReason

Geospatial Event and Temporal Reasoning

Scene Graph Generation



```

{
  "entities": [
    {
      "attributes": [
        {
          "attribute": "",
          "value": ""
        }
      ],
      "entity": ""
    }
  ],
  "relationships": [
    {
      "reasoning": "",
      "relationship": ""
    }
  ]
}
  
```

Scene graph Agent

```

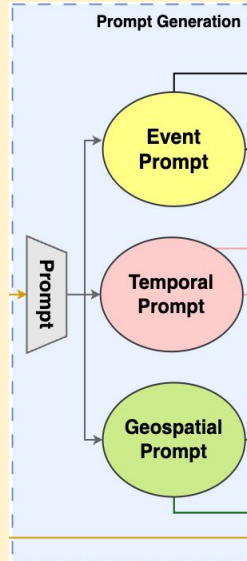
{
  "idea": "",
  "reasoning": ""
}
  
```

Abstract Agent

GETReason

Geospatial Event and Temporal Reasoning

Prompt Generation



Prompt Generator

```
{
  "global_event_specialist": {
    "prompt": "",
    "reasoning": ""
  },
  "spatial_specialist": {
    "prompt": "",
    "reasoning": ""
  },
  "temporal_specialist": {
    "prompt": "",
    "reasoning": ""
  }
}
```

Event Prompt

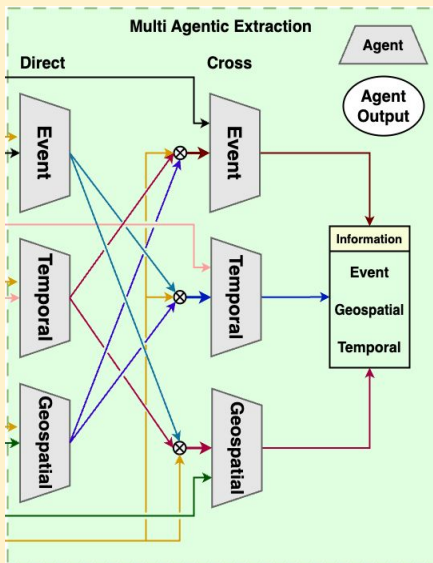
Temporal Prompt

Geospatial Prompt

GETReason

Geospatial Event and Temporal Reasoning

Multi Agentic Extraction



```
{
  "event": {
    "value": "",
    "reasoning": ""
  },
  "background": {
    "value": "",
    "reasoning": ""
  }
}
```

Event Agent

```
{
  "century": "",
  "day": "",
  "decade": "",
  "month": "",
  "year": ""
}
```

Temporal Agent

```
{
  "city": "",
  "country": "",
  "state_or_province": ""
}
```

Geospatial Agent



Dataset

TAR**A**: 11,241 images

WikiTi**Lo**: 6,296 images

→ **JSON-based** structure: location, time, event & reasoning

Event

```
{
  "event": {
    "value": "",
    "reasoning": ""
  },
  "background": {
    "value": "",
    "reasoning": ""
  }
}
```

Geospatial

```
{
  "city": "",
  "country": "",
  "state_or_province": ""
}
```

Temporal

```
{
  "century": "",
  "day": "",
  "decade": "",
  "month": "",
  "year": ""
}
```



Restructuring & Augmentation

TARA*

```
{
  "id": "",
  "event": {
    "value": "",
    "reasoning": ""
  },
  "background": {
    "value": "",
    "reasoning": ""
  },
  "geospatial_information": {
    "city": "",
    "country": "",
    "state_or_province": ""
  },
  "temporal_information": {
    "century": "",
    "day": "",
    "decade": "",
    "month": "",
    "year": ""
  }
}
```

Event
Augmentation

Spatio-Temporal
Augmentation

Deduction
Augmentation


WikiTiLo*

```
{
  "id": "",
  "temporal_information": {
    "century": "",
    "decade": "",
    "year": "",
    "month": "",
    "day": ""
  },
  "geospatial_information": {
    "country": "",
    "state_or_province": "",
    "city": ""
  }
}
```

Evaluation

GREAT (**G**eospatial **R**easoning **E**vent **A**ccuracy with
Temporal alignment)

 Event: Semantic cosine similarity

 Geospatial: Haversine distance + hierarchy

 Temporal: Weighted unit-wise scoring

GREAT

(Geospatial Reasoning Event Accuracy with Temporal alignment)

Event Evaluation

Cosine Similarity of Sentence
Embeddings (Event +
Background)

$$CS_{\text{shifted}} = \frac{CS + 1}{2}$$

$$ES_i = \frac{CS((e_i + b_i), (E_i + B_i)) + 1}{2}$$

GREAT

(Geospatial Reasoning Event Accuracy with Temporal alignment)

Geo-spatial Evaluation

**Haversine Distance-Based
Similarity**

$$d = 2R \arcsin \left(\sqrt{\sin^2 \left(\frac{\Delta\phi}{2} \right) + \cos(\phi_1) \cos(\phi_2) \sin^2 \left(\frac{\Delta\lambda}{2} \right)} \right)$$

$$S_{\text{geo}} = \max \left(0, 1 - \frac{d}{D_{\text{max}}} \right)$$

GREAT

(Geospatial Reasoning Event Accuracy with Temporal alignment)

Temporal Evaluation

**Granularity-Weighted
Temporal Accuracy**

$$S_u = \begin{cases} 1, & \text{if } gt_u = pred_u \text{ (exact match, century level)} \\ \max \left(0, 1 - \frac{|gt_u - pred_u|}{T_u} \right), & \text{otherwise} \end{cases}$$

$$TS_i = \frac{\sum_u w_u S_u}{\sum_u w_u}$$



Results Summary

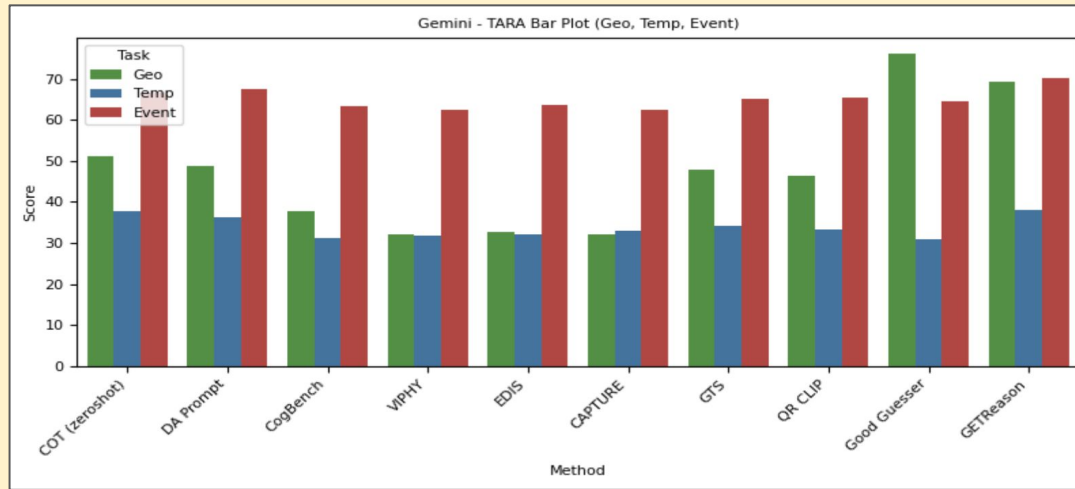
GETReason achieves highest scores in **Event**, **Geo**,
Temporal inference

Superior performance on **TARA** and **WikiTiLo** datasets

Ablation confirms value of **cross-agent** iteration and
structured outputs

Results Summary

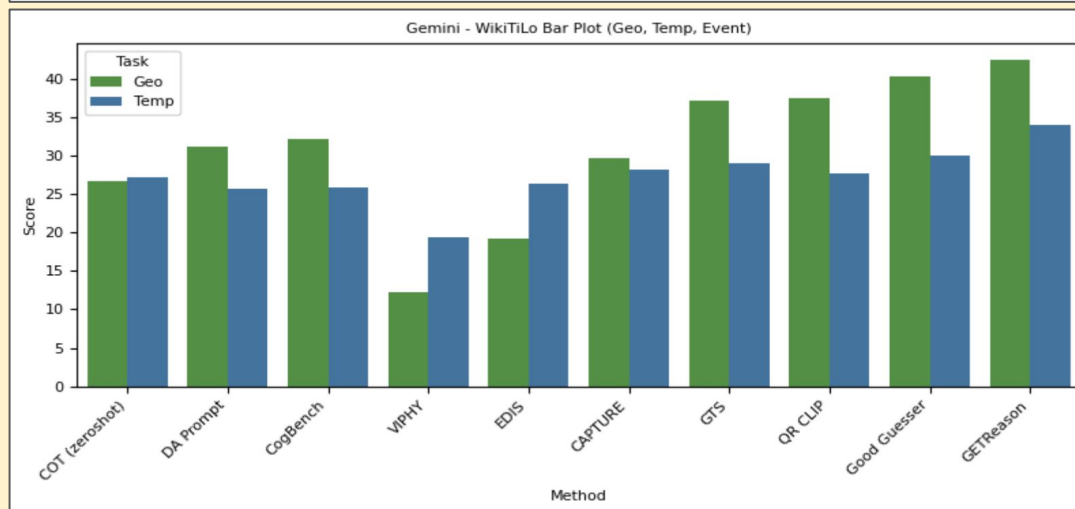
TARA*



Best performing model's performance:

Gemini 1.5 pro
(among the 3 tried models: Gemini 1.5 pro, GPT-4o-mini, Qwen 2.5-VL)

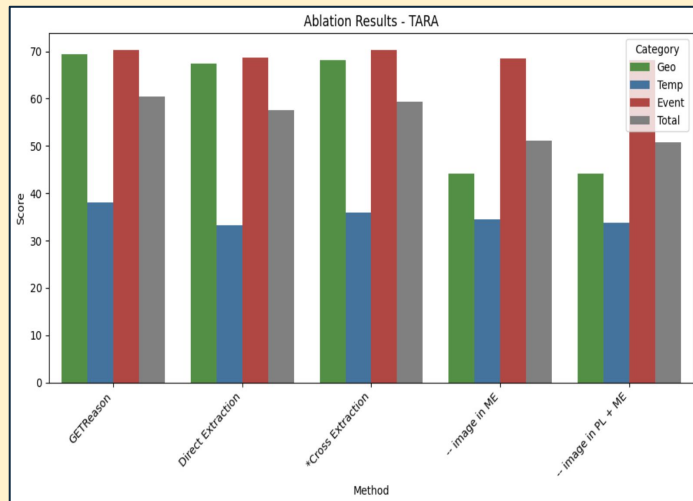
WikiTiLo*



on the two used datasets.

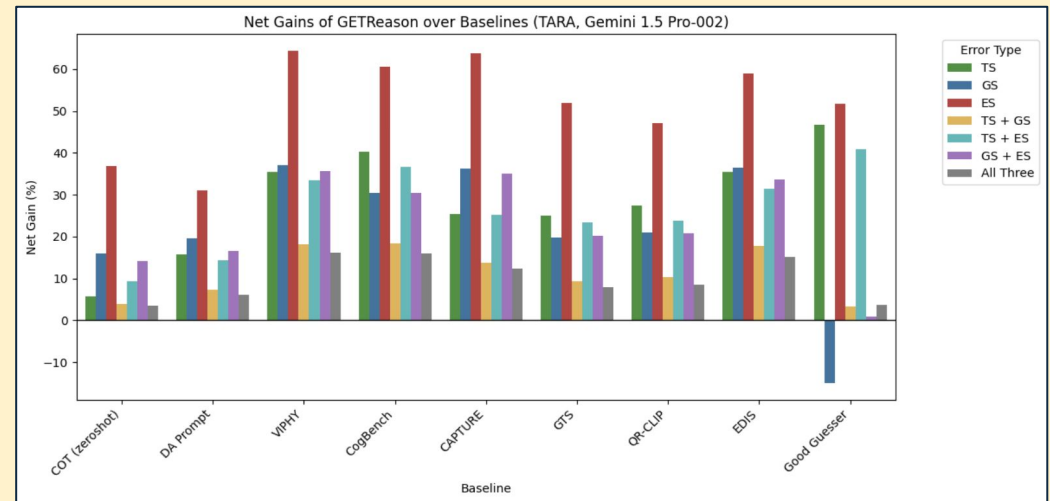
Results Summary

Ablation



Ablation: Impact of cross extraction and images in prompt and multi-extraction layer

Error Analysis



Relative error: Net improvement observed in GETReason for different tasks against baselines



Conclusion & Takeaways

Hierarchical multi-agent design improves contextual reasoning

Structured responses help in controlling the output of an LLM

GREAT metric evaluates reasoning, not just overlap