# Supplementing Missing Visions via Dialog for Scene Graph Generations


Zhenghao Zhao[1,*] Ye Zhu[1,*] Xiaoguang Zhu[2] Yuzhang Shang[1] Yan Yan[1]

[1] Illinois Institute of Technology, USA [2] Shanghai Jiao Tong University, China



## Abstract

**Motivation: Investigation into Incomplete Visual Input.**

- Focus: Scene Graph Generation (SGG) with varying levels of missing visual data.
- Issue: Performance drops due to insufficient visual input.

**Proposed Solution: Supplementary Interactive Dialog (SI-Dial).**

- Model-agnostic framework for natural language dialog interactions.
- Enhances current AI systems with QA capabilities in natural language.

**Experimentation and Results.**

- Task setup with missing visual input to test feasibility.
- Effectiveness of dialog module as a supplementary information source.
- Promising performance improvement over multiple baselines in extensive experiments.


## SI-Dial for Missing Visions

**Input and Output Process.** Update representations with dialog interactions.

$$Input: O' = \{V', E'\} \Rightarrow Output: O = \{V, E\}, \quad (1)$$

**Question Encoder** is used to extract the question embedding for all thee question candidates.

$$x_j = QE(q_{cand.,j}),\ j \in \{1, 2, ..., N_{cand.}\}, \quad (2)$$

**Question Decoder** selects the question that has the highest similarity score with the generated question embedding.

$$q_i = argmin_k\ Sim.(QD(O', x_{his,i-1}), x_j), \quad (3)$$

**History Encoder** is for interactively encoding the QA pairs from the dialog. The output $x_{his,N_R}$ from the history encoder is used as the supplementary information for the missing visual input.

$$x_{his,i} = HE(x_{his,i-1}, x_{qa_i}), \quad (4)$$

**Vision Update Module.** Preliminary objects $O'$ obtained from the incomplete visions is updated by incorporating the dialog information.


## References

[1] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *CVPR*, 2018.

[2] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. Learning to compose dynamic tree structures for visual contexts. In *CVPR*, 2019.

[3] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *CVPR*, 2020.


## Proposed Method

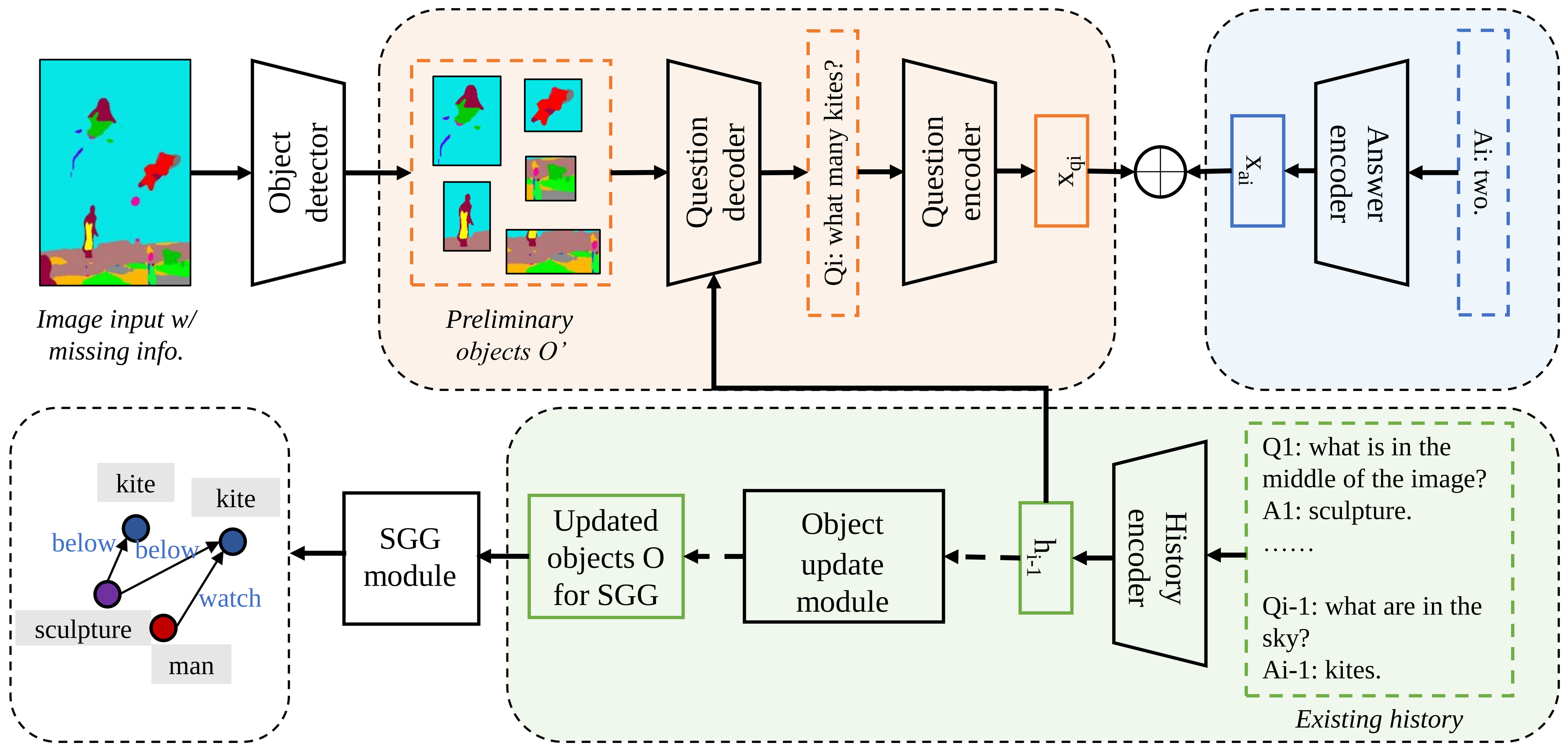


**Figure 1:** The overall architecture of our proposed *SI-Dial* framework. We first obtain the preliminary objects from the object detector based on the incomplete visual input, and propose to conduct an interactive dialog process. Note that the dashed lines denote the operations only after the dialog is completed) for the final scene graph generation.

## Experiments

| Vision Input | Model | Predicate Classification mR@20 | mR@50 | mR@100 | Scene Graph Classification mR@20 | mR@50 | mR@100 | Scene Graph Detection mR@20 | mR@50 | mR@100 |
|---|---|---|---|---|---|---|---|---|---|---|
| Original VG | IMP† | - | 9.8 | 10.5 | - | 5.8 | 6.0 | - | 3.8 | 4.8 |
| | FREQ.† | 8.3 | 13.0 | 16.0 | 5.1 | 7.2 | **8.5** | **4.5** | **6.1** | **7.1** |
| | MOTIF† | 11.5 | 14.6 | 15.8 | **6.5** | **8.0** | **8.5** | 4.1 | 5.5 | 6.8 |
| | VCTREE† | **11.7** | **14.9** | **16.1** | 6.2 | 7.5 | 7.9 | 4.2 | 5.7 | 6.9 |
| Object Blur | MOTIF | 11.23 | 14.36 | 15.71 | 6.20 | 7.51 | 7.90 | 4.13 | 5.48 | 6.82 |
| | VCTREE | 11.48 | 14.61 | 15.88 | 5.97 | 7.46 | 7.85 | 4.24 | 5.66 | 6.93 |
| | MOTIF+Random QA | 11.52 | 14.78 | 16.08 | 5.82 | 7.34 | 7.86 | 4.64 | 6.18 | 7.24 |
| | VCTREE+Random QA | 11.55 | 14.83 | 16.17 | 5.51 | 7.03 | 7.48 | 4.71 | 6.23 | 7.43 |
| | MOTIF+SI-Dial | 13.31 | 16.74 | 18.09 | **6.96** | **8.51** | **9.00** | 5.77 | 7.76 | 9.12 |
| | VCTREE+SI-Dial | **13.40** | **16.88** | **18.26** | 6.67 | 8.12 | 8.59 | **5.88** | **7.92** | **9.28** |
| Image Blur | MOTIF | 11.69 | 14.31 | 15.64 | 6.24 | 7.56 | 7.90 | 3.88 | 5.21 | 6.37 |
| | VCTREE | 11.72 | 14.38 | 15.78 | 6.03 | 7.39 | 7.83 | 3.82 | 5.18 | 6.33 |
| | MOTIF+Random QA | 11.57 | 13.93 | 15.14 | 6.75 | 8.21 | 8.69 | 4.10 | 5.39 | 6.26 |
| | VCTREE+Random QA | 11.70 | 14.26 | 15.53 | 6.68 | 8.03 | 8.55 | 4.07 | 5.34 | 6.25 |
| | MOTIF+SI-Dial | 12.90 | 16.26 | 17.91 | **8.41** | **10.33** | **11.00** | 5.05 | 6.96 | **8.23** |
| | VCTREE+SI-Dial | **13.62** | **17.18** | **18.49** | 7.93 | 10.02 | 10.86 | **5.24** | **7.08** | 8.11 |
| Semantic Masked | MOTIF | 11.61 | 14.28 | 15.57 | 4.45 | 5.41 | 5.68 | 2.80 | 3.89 | 4.76 |
| | VCTREE | 11.68 | 14.32 | 15.59 | 4.40 | 5.38 | 5.69 | 2.80 | 3.87 | 4.68 |
| | MOTIF+Random QA | 12.00 | 15.32 | 16.67 | 5.83 | 7.14 | 7.65 | 2.79 | 3.86 | 4.68 |
| | VCTREE+Random QA | 12.28 | 15.69 | 17.04 | 5.66 | 7.01 | 7.28 | 2.92 | 4.01 | 4.85 |
| | MOTIF+SI-Dial | **12.79** | 16.26 | 17.58 | **6.44** | **7.85** | **8.33** | 3.03 | 4.21 | 4.92 |
| | VCTREE+SI-Dial | 12.73 | **16.35** | **17.63** | 6.21 | 7.68 | 8.05 | **3.15** | **4.28** | **5.00** |

**Table 1:** Quantitative evaluations for the SGG with missing visions. The results are reported on mean Recall.

## Results


**Figure 2:** Qualitative results for the SGG with missing visions. Displayed in sequence from top to bottom are scenarios with Object Blur, Image Blur, and Semantic Masked.

ILLINOIS TECH

ICASSP 2024 KOREA