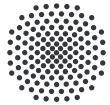


University of Stuttgart
Germany

Multimodal Feature Fusion via Graph Neural Networks for Visual Dialog

Topic Introduction

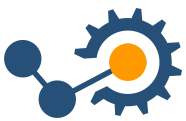
Nishan Chatterjee

April 27, 2022

Perceptual User Interfaces Group, University of Stuttgart

www.perceptualui.org

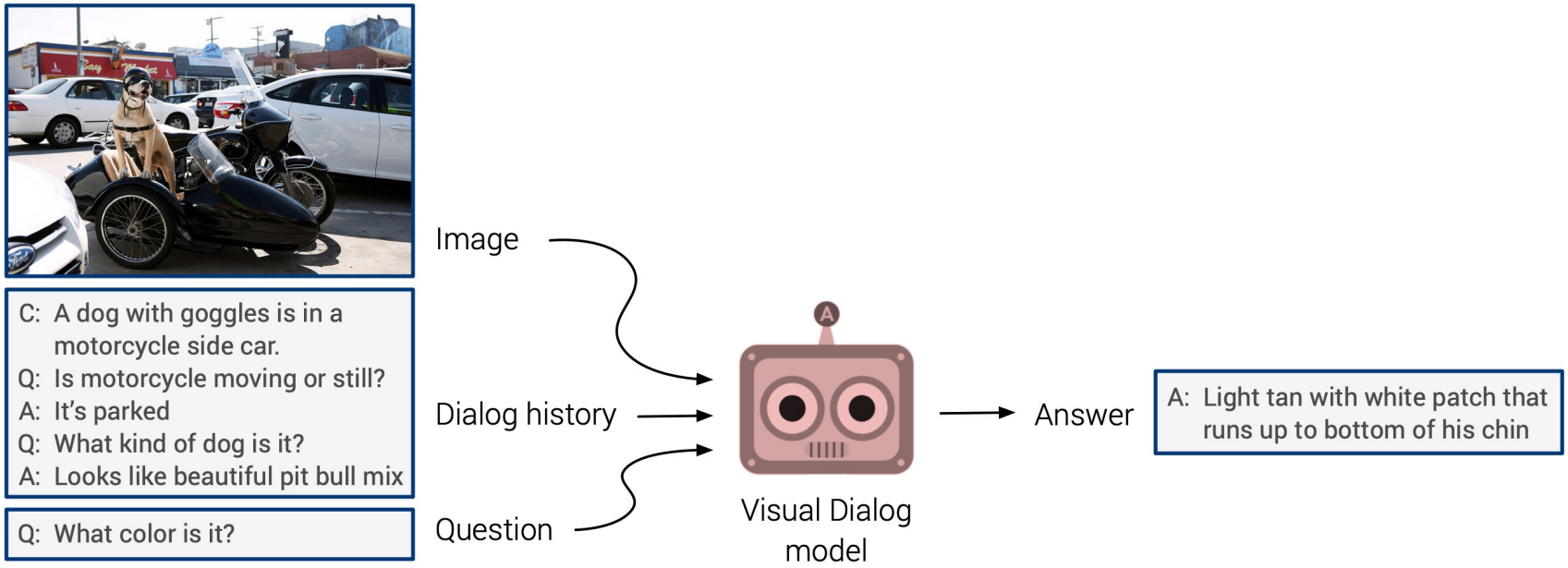
Table of Contents



Visual Dialog

Task Introduction

What is Visual Dialog?



Source: Visual Dialog Task Description [Das et al., 2016]

Given an image, a dialog history, and a question, an AI agent has to ground the question in the image and infer from context history to answer questions accurately.

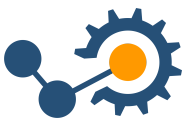
Visual Dialog

Neural Architecture Choices

Why Graph Neural Networks over Deep Neural Networks?

- Deep Neural Networks are good for Euclidean Data (data with an underlying grid-like structure) [Bronstein et al., 2016]
- Complex data having no fixed node ordering or reference point makes it difficult for Deep Learning
- Geometric Deep Learning (Like Graphs) redefine isolated data-points as networks and relations between entities
- The edges and the topology portray the relation between these entities

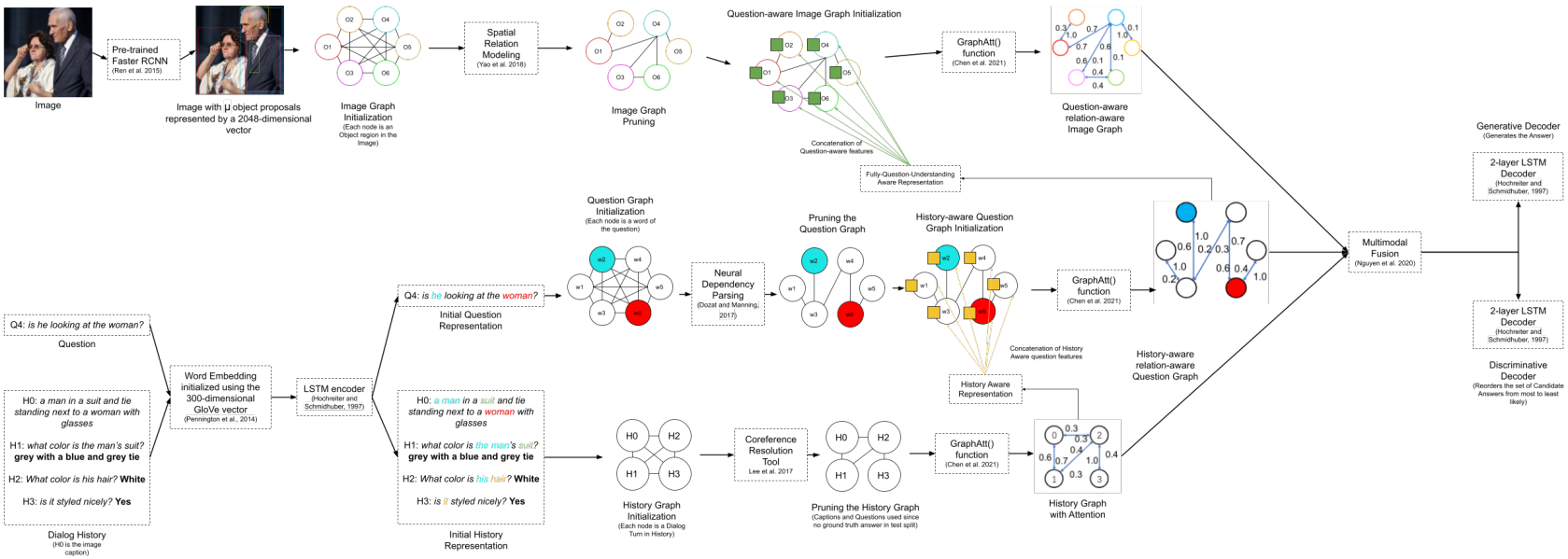
Table of Contents



Implementations

GoG: Relation-aware Graph-over-Graph Network [Chen et al., 2021]

Architecture Overview

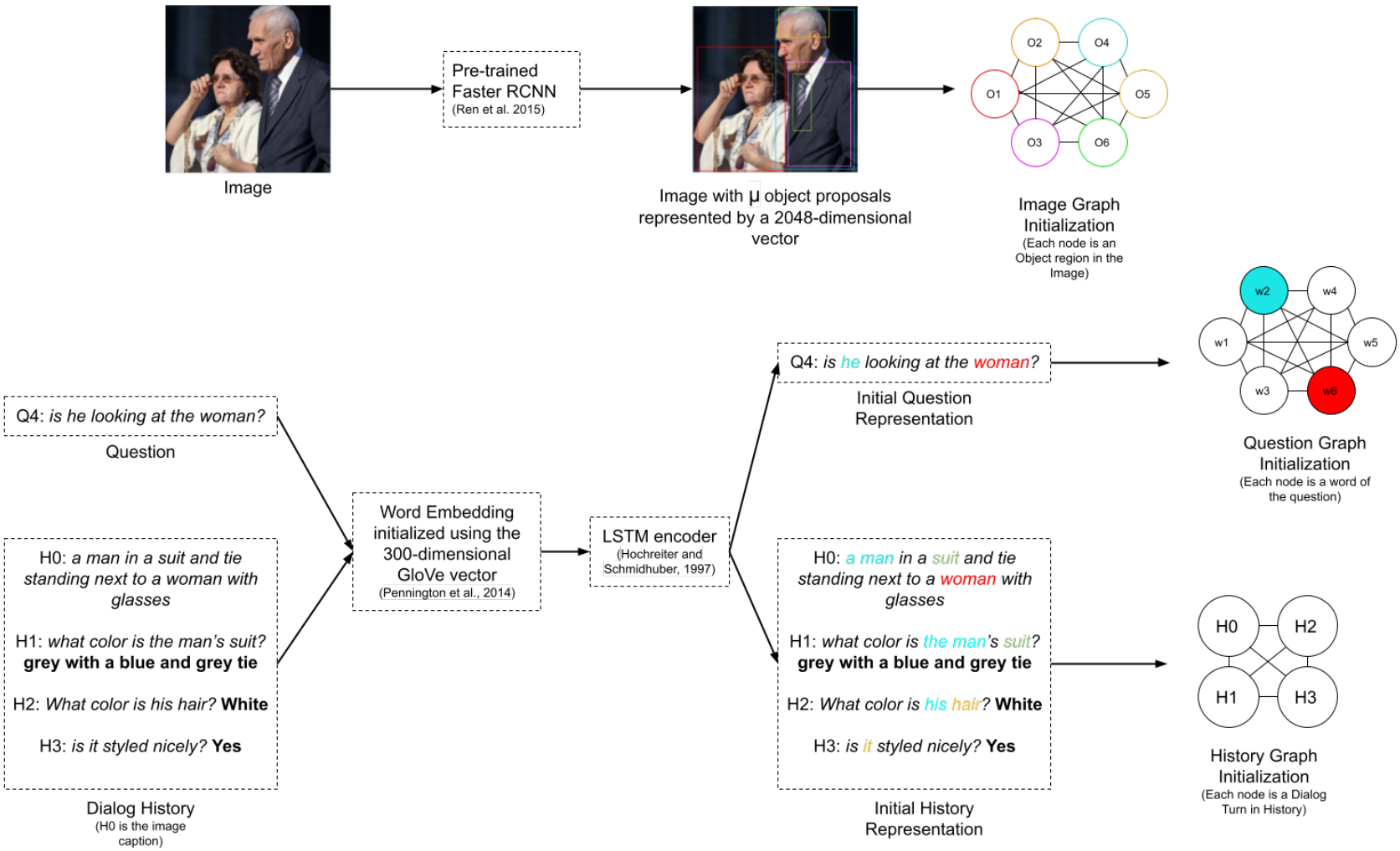


Source: Complete Architecture Overview of [Chen et al., 2021]

1

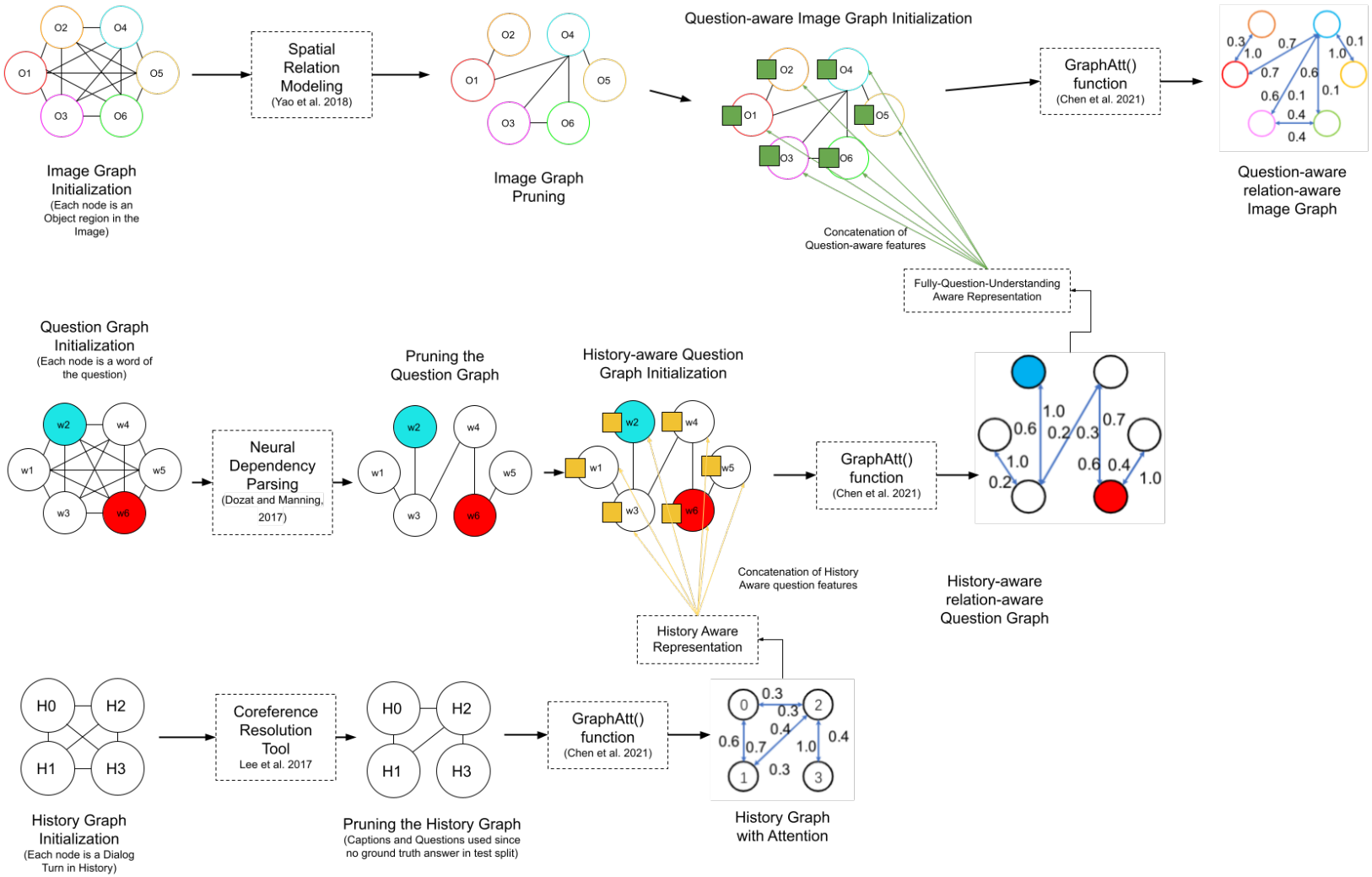
[1] Architecture includes previous methodologies by [Ren et al., 2015], [Yao et al., 2018], [Pennington et al., 2014], [Hochreiter and Schmidhuber, 1997], [Dozat and Manning, 2016], [Lee et al., 2017], [Nguyen et al., 2019]

Feature Representation



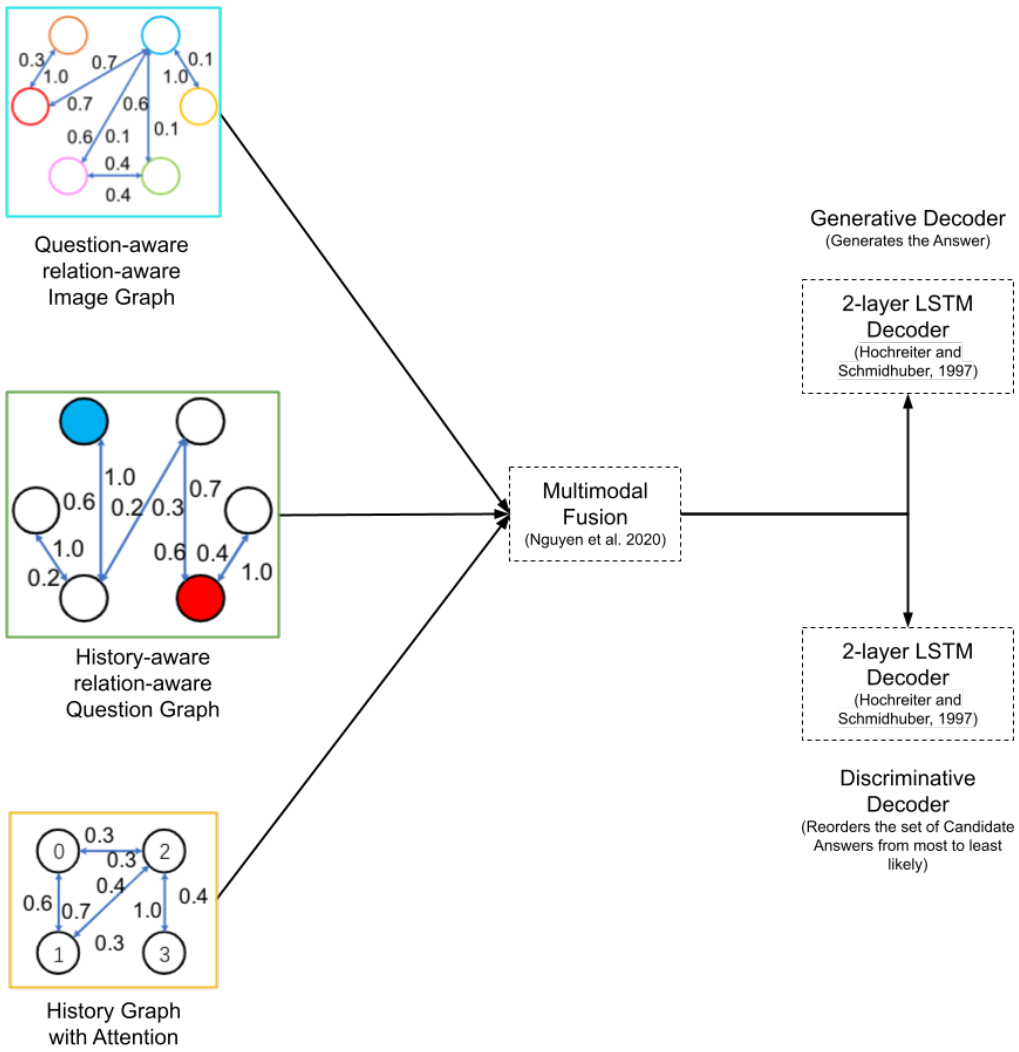
Source: Feature Representation of [Chen et al., 2021]

Pruning, Attention, and Adaptive Relation-awareness



Source: Graph Pruning and Graph Attention Scoring of [Chen et al., 2021]

Evaluation

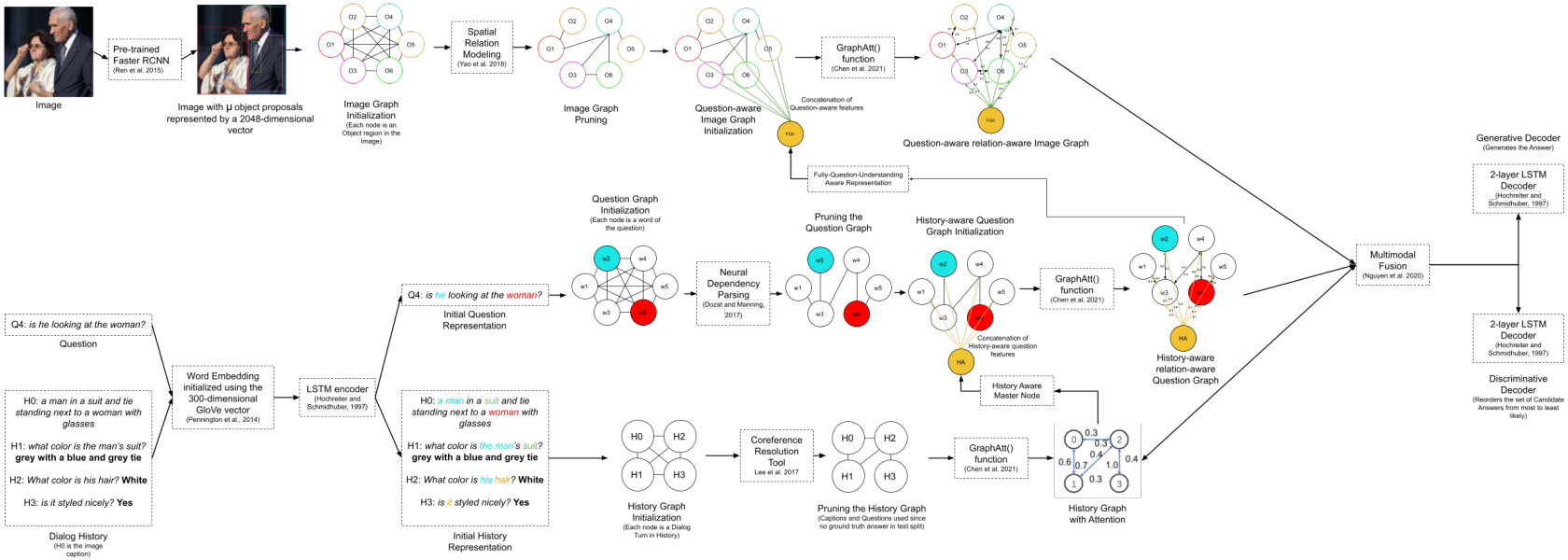


Source: Evaluation of [Chen et al., 2021]

Implementations

Multi-modal Feature Fusion Model (Main)

Architecture Overview

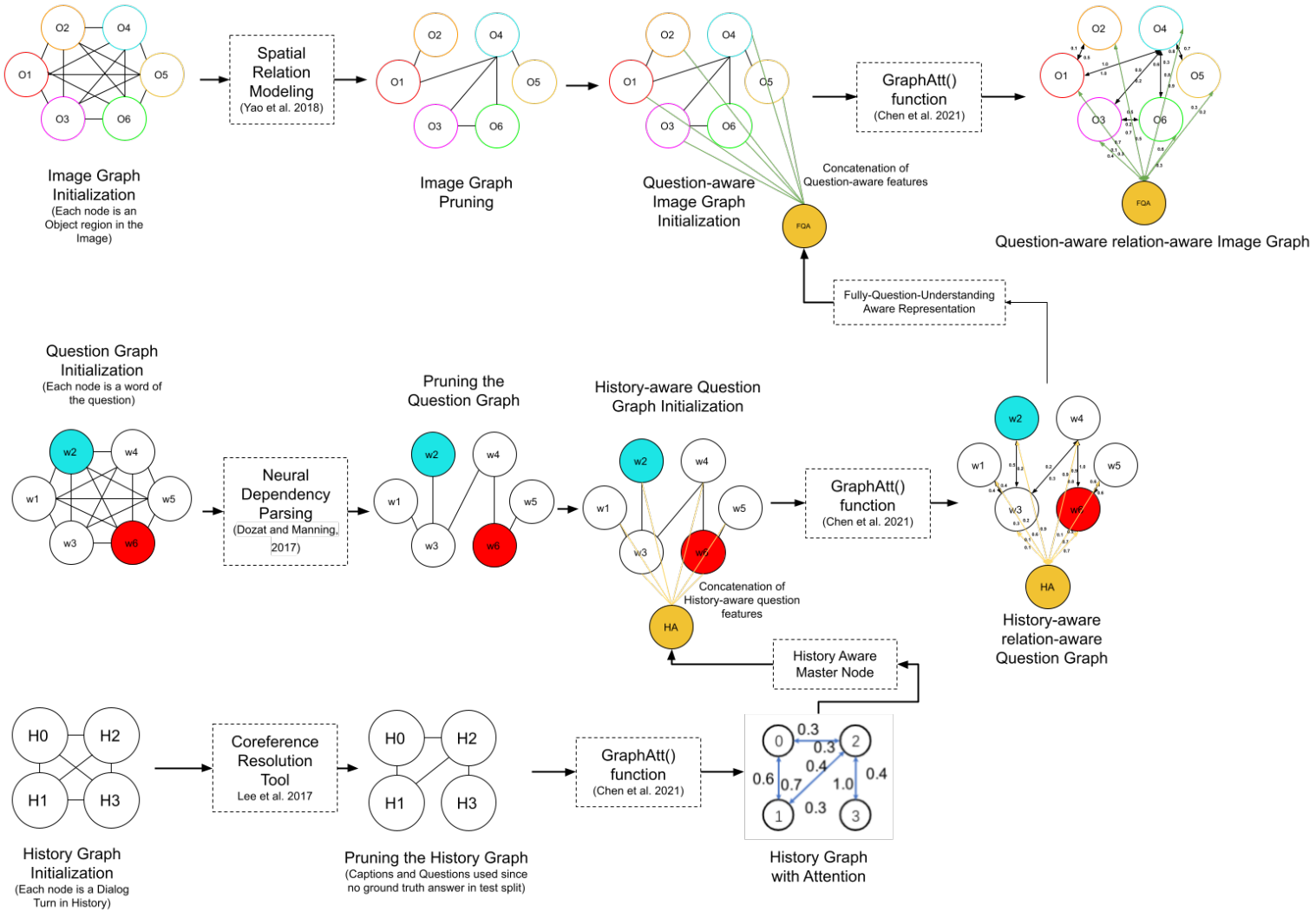


Source: Complete Architecture Overview of the Main novel Implementation

2

[2]Architecture includes previous methodologies by [Ren et al., 2015], [Yao et al., 2018], [Pennington et al., 2014], [Hochreiter and Schmidhuber, 1997], [Dozat and Manning, 2016], [Lee et al., 2017], [Nguyen et al., 2019]

Relation-awareness using Master Node

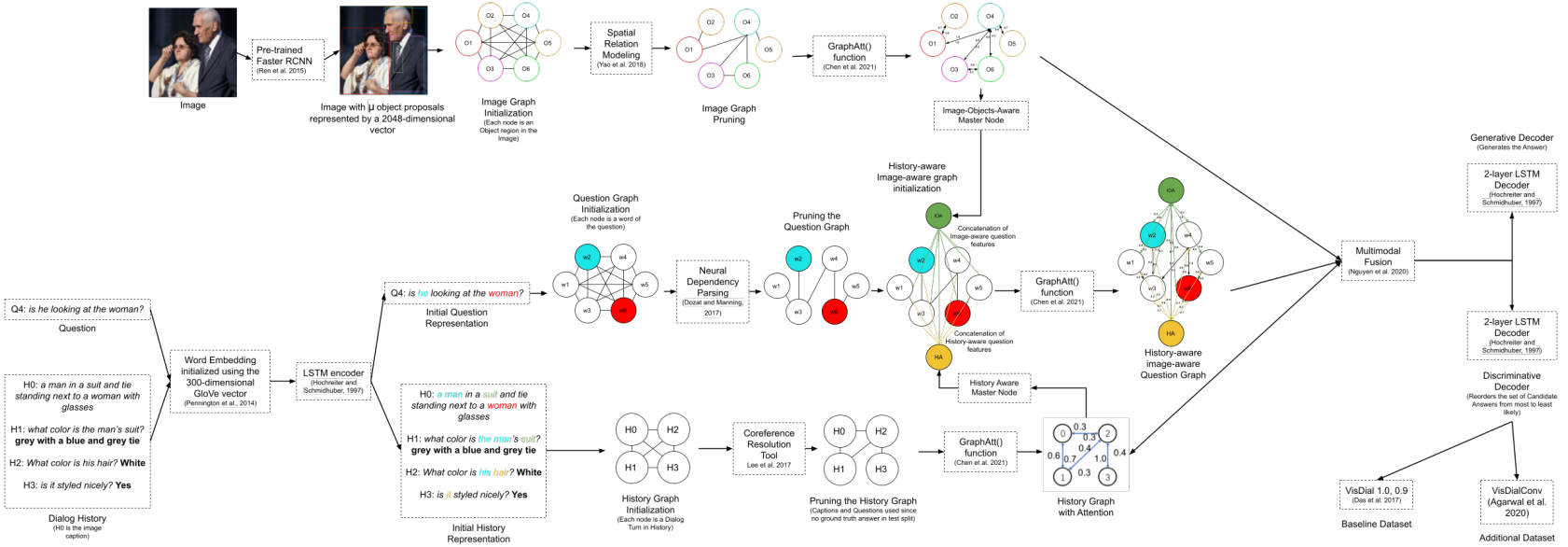


Source: Implementation strategy of the Main novel Implementation

Implementations

Multi-modal Feature Fusion Model (Optional)

Architecture Overview

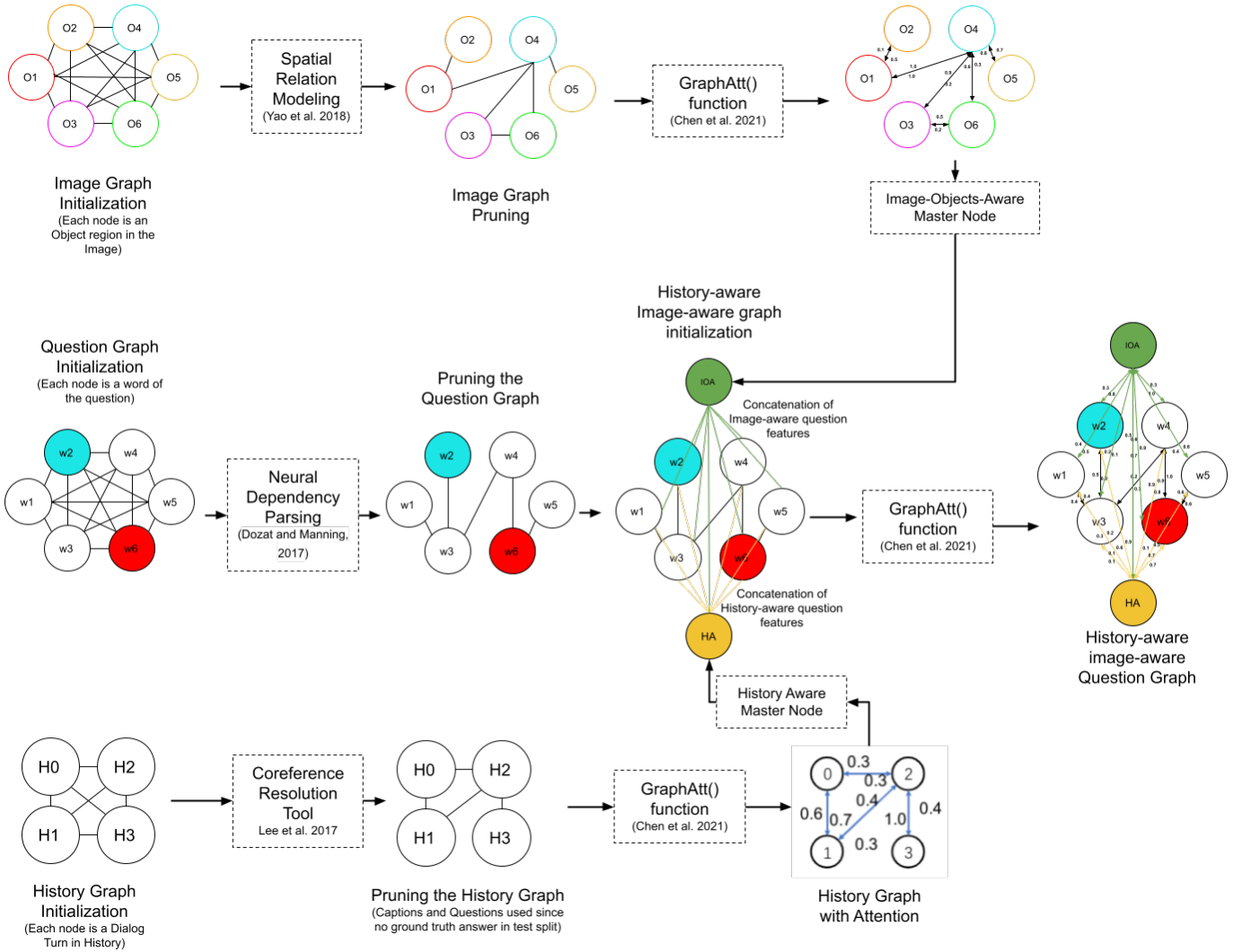


Source: Complete Architecture Overview of the Optional novel Implementation

3

[3] Architecture includes previous methodologies by [Ren et al., 2015], [Yao et al., 2018], [Pennington et al., 2014], [Hochreiter and Schmidhuber, 1997], [Dozat and Manning, 2016], [Lee et al., 2017], [Nguyen et al., 2019]

History-aware Image-aware Question-Graph

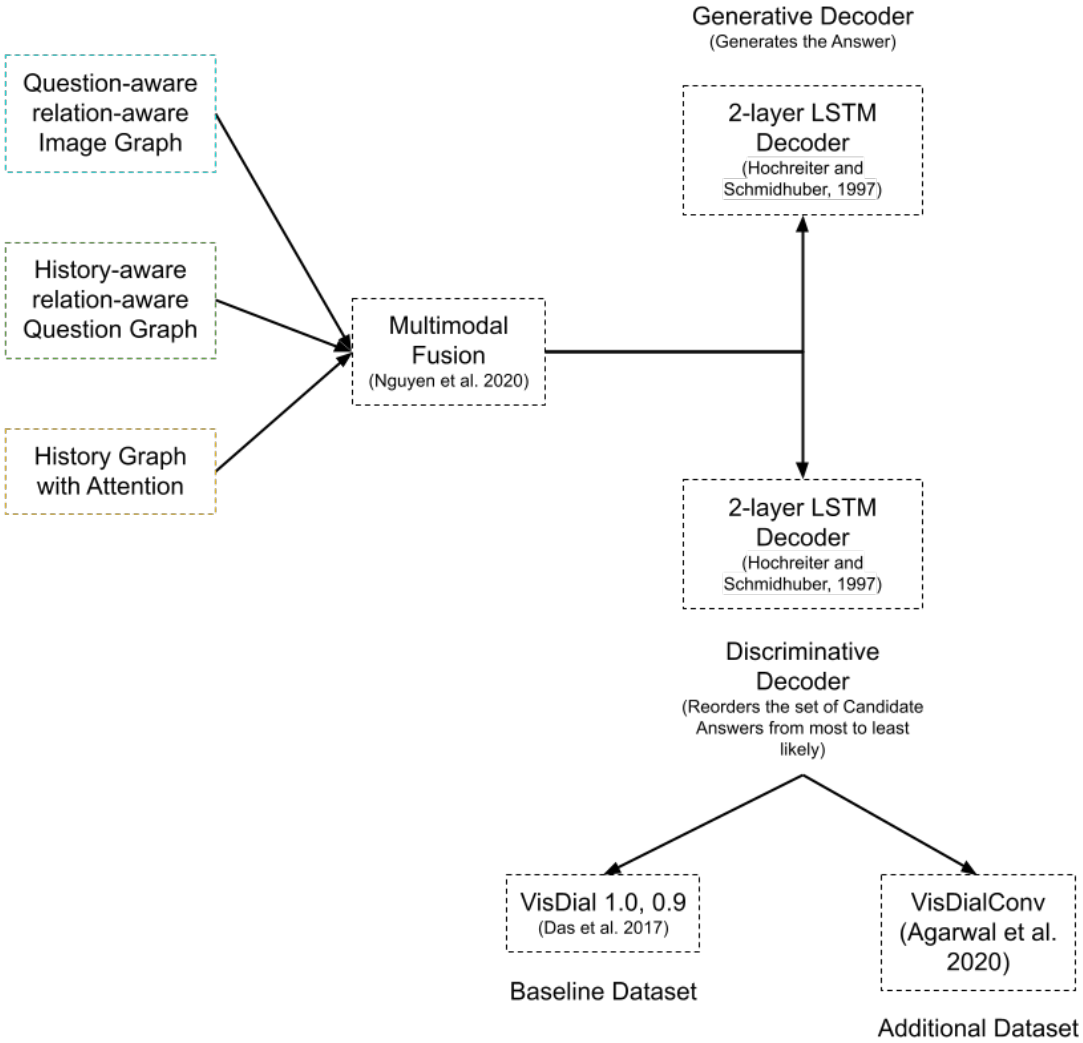


Source: Implementation strategy of the Optional novel Implementation

Implementations

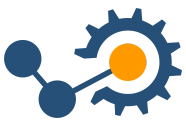
Optional Evaluation Methods

Other Optional Evaluation Methods

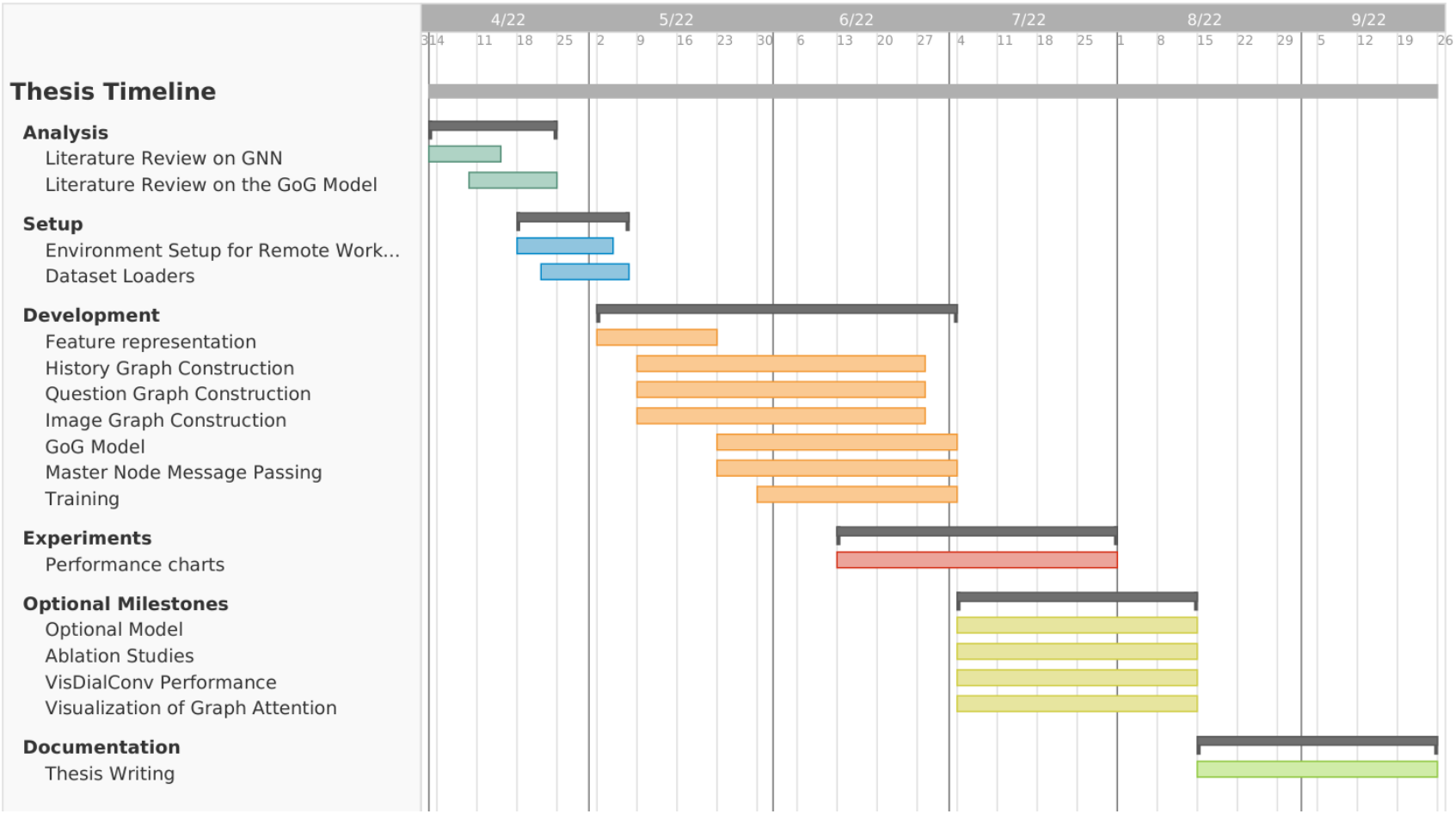


Source: Optional Evaluation Implementation with data from [Agarwal et al., 2020]

Table of Contents



Timeline Overview



Source: Timeline overview

References i

S. Agarwal, T. Bui, J.-Y. Lee, I. Konstas, and V. Rieser. History for visual dialog: Do we really need it? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8182–8197, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.728. URL https://aclanthology.org/2020.acl-main.728.

M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst. Geometric deep learning: going beyond euclidean data. *CoRR*, abs/1611.08097, 2016. URL http://arxiv.org/abs/1611.08097.

F. Chen, X. Chen, F. Meng, P. Li, and J. Zhou. Gog: Relation-aware graph-over-graph network for visual dialog. *CoRR*, abs/2109.08475, 2021. URL https://arxiv.org/abs/2109.08475.

A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. F. Moura, D. Parikh, and D. Batra. Visual dialog. *CoRR*, abs/1611.08669, 2016. URL http://arxiv.org/abs/1611.08669.

T. Dozat and C. D. Manning. Deep biaffine attention for neural dependency parsing. *CoRR*, abs/1611.01734, 2016. URL http://arxiv.org/abs/1611.01734.

References ii

S. Hochreiter and J. Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9 (8):1735–1780, 11 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735. URL https://doi.org/10.1162/neco.1997.9.8.1735.

K. Lee, L. He, M. Lewis, and L. Zettlemoyer. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1018. URL https://aclanthology.org/D17-1018.

V. Nguyen, M. Suganuma, and T. Okatani. Efficient attention mechanism for handling all the interactions between many inputs with application to visual dialog. *CoRR*, abs/1911.11390, 2019. URL http://arxiv.org/abs/1911.11390.

J. Pennington, R. Socher, and C. Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, Oct. 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL https://aclanthology.org/D14-1162.

References iii

S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL `https://proceedings.neurips.cc/paper/2015/file/14bfa6bb14875e45bba028a21ed38046-Paper.pdf`.

T. Yao, Y. Pan, Y. Li, and T. Mei. Exploring visual relationship for image captioning. *CoRR*, abs/1809.07041, 2018. URL `http://arxiv.org/abs/1809.07041`.