A. Appendix for "Meta Module Network for Compositional Visual Reasoning"

A.1. Visual Encoder and Multi-head Attention

The multi-head attention network is illustrated in Figure 7.

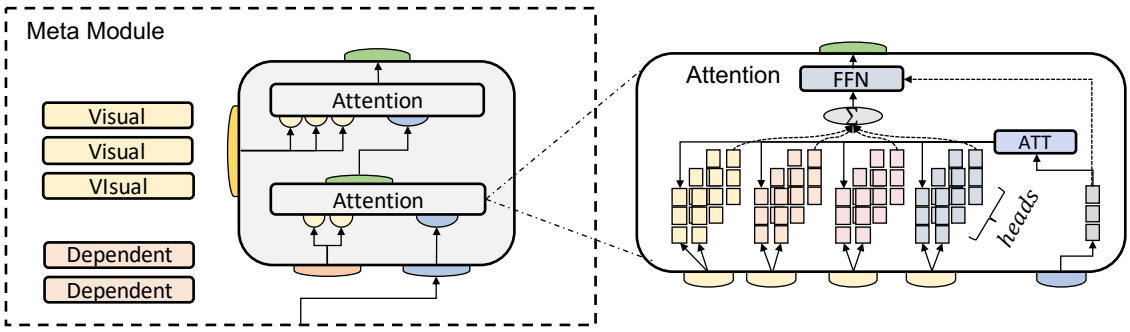

Figure 7. Illustration of the multi-head attention network used in the Meta Module.

A.2. Recipe Embedding

The recipe embedder is illustrated in Figure 8.

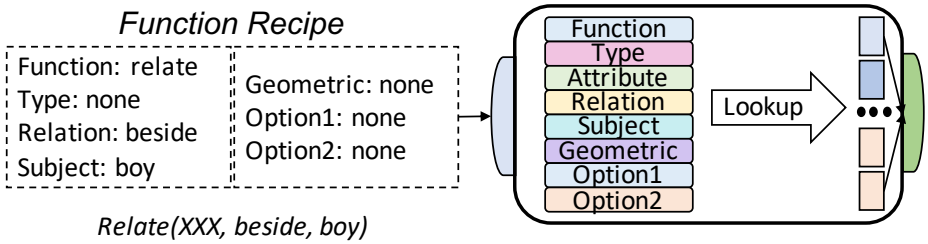

Figure 8. Illustration of the recipe embedder.

A.3. Implementation

The implementation of the proposed model is demonstrated in Figure 9. Our model can be efficiently implemented by adding masks on top of the Transformer model, guided by an additional supervision signal.

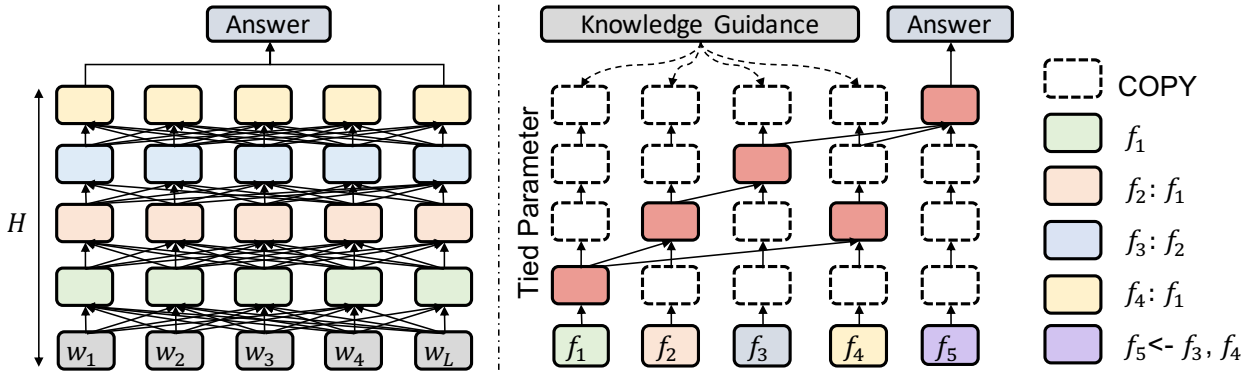

Figure 9. Illustration of the implementation of both the Transformer model (left) and our model (right).

A.4. Function Statistics

The function statistics is listed in Table 5.

Type	Relate	Select	Filter	Choose	Verify	Query	Common	Differ	Bool	Exist	All
Funcs	5	1	8	12	5	6	2	6	2	1	48

Table 5. The statistics of different functions.

Binary						Objects			String	
verify	choose	compare	exist	and	or	filter	select	relate	query[object]	query[scene]
0.74	0.79	0.88	0.88	0.97	0.95	0.67	0.61	0.44	0.61	0.65

Table 6. Error analysis on different types of functions. "Objects" functions only appear in the intermediate step, "String" function only appears in the final step, "Binary" functions can occur in both intermediate and final step

A.5. Inferential Chains

More inferential chains are visualized in Figure 10 and 11.

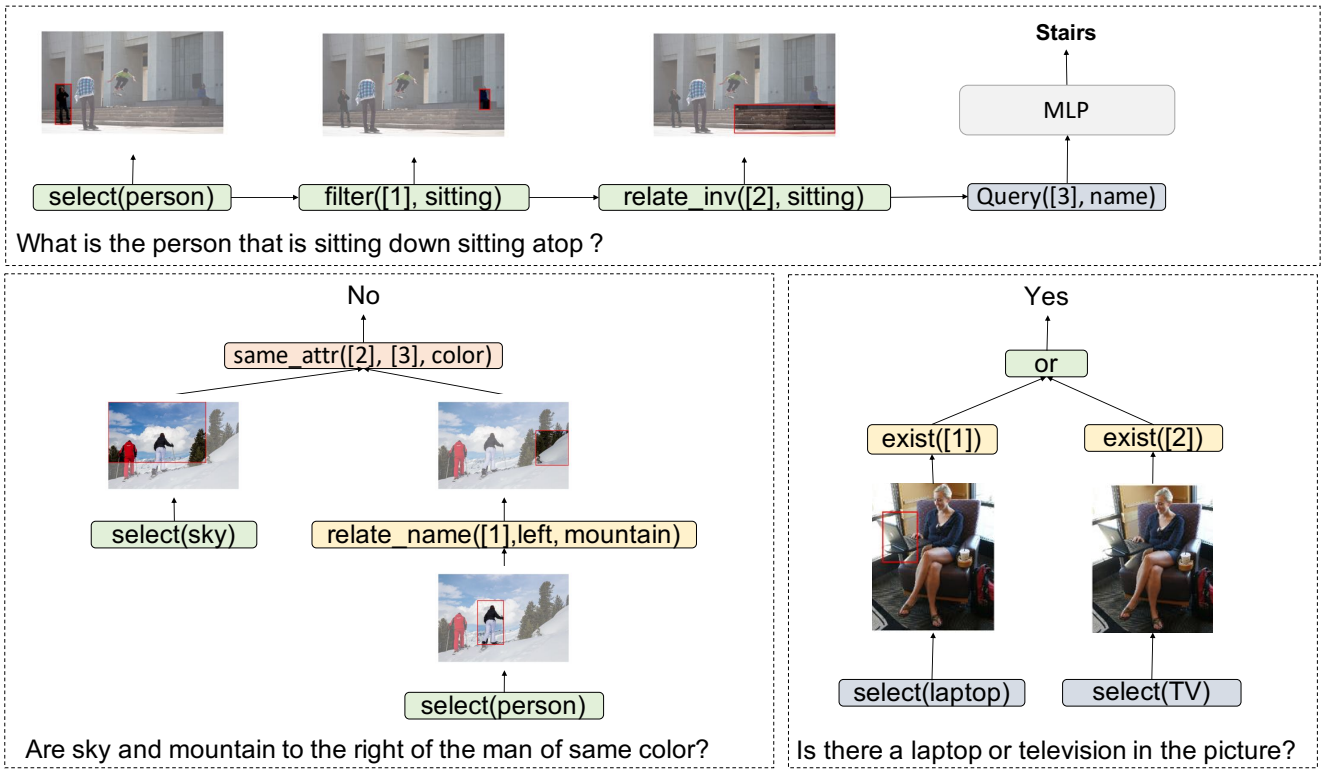

Figure 10. More examples on visualization of the inferential chains learned by our model.

A.6. Detailed Error Analysis

Furthermore, we conduct a detailed analysis of function-wise execution accuracy to understand the limitation of MMN. Results are shown in Table 6. Below are the observed main bottlenecks: (i) relation-type functions such as `relate`, `relate_inv`; and (ii) object/attribute recognition functions such as `query_name`, `query_color`. We hypothesize that this might be attributed to the quality of visual features from standard object detection models [1], which does not capture the relations between objects well. Besides, the object and attribute classification network are not fine-tuned on GQA. This suggests that scene graph modeling for visual scene understanding is critical to surpassing NSM [18] on performance.

A.7. Function Description

The detailed function descriptions for CLEVR and GQA are provided in Figure 12 and Figure 13, respectively.

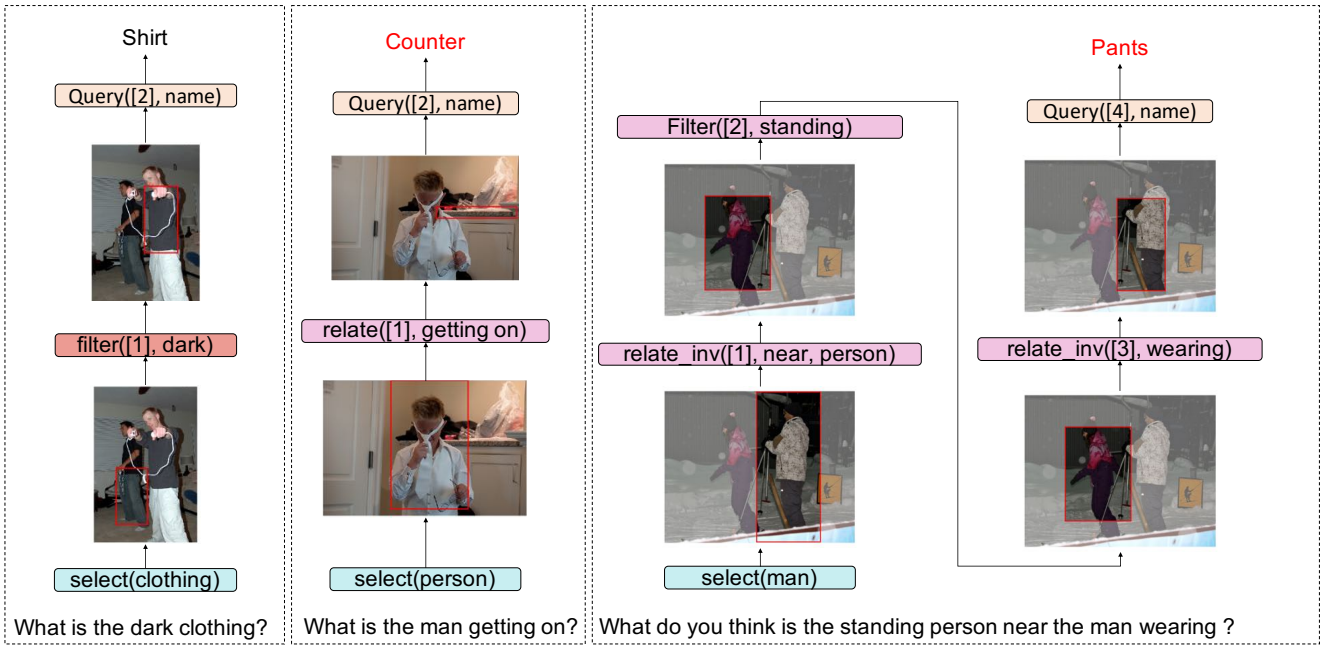

Figure 11. More examples on visualization of the inferential chains learned by our model.

Type	Overrides	Arg0 ? Means Dependecy	arg1	Output
Relate	-	?	Position	Region
Filter	color	?	color	Region
	material	?	material	Region
	shape	?	shape	Region
	size	?	size	Region
Same	color	?	?	Region
	material	?	?	Region
	shape	?	?	Region
	size	?	?	Region
Query	color	?	color	Answer
	material	?	material	Answer
	shape	?	shape	Answer
	size	?	size	Answer
Logic	Union	?	?	Region
	Intersect	?	?	Region
Equal	color	?	?	Answer
	material	?	?	Answer
	shape	?	?	Answer
	size	?	?	Answer
Compare	greater_than	?	?	Answer
	less_than	?	?	Answer
Count	-	?	-	Answer
exist	-	?	-	Answer

Figure 12. The function definitions and their corresponding outputs on CLEVR.

Type	Overrides	arg0 (? Means Dependency)	arg1	arg2	arg3	Output
Relationship	Relate	?	Relation	-	-	Region
	Relate_with_name	?	Relation	Object	-	
	Relate_invese	?	Relation	-	-	
	Relate_inverse_with_name	?	Relation	Object	-	
	Relate_with_same_attribute	?	Relation	Attribute	-	
Selection	Select	-	Object	-	-	Region
Filter	Filter_horizontal_position	?	H-Position	-	-	Region
	Filter_Vertical_position	?	V-Position	-	-	
	Filter_with_color	?	Color	-	-	
	Filter_with_shape	?	Shape	-	-	
	Filter_with_activity	?	Activity	-	-	
	Filter_with_material	?	Material	-	-	
	Filter_with_color_noteq	?	Color	-	-	
	Filter_with_shape_noteq	?	Shape	-	-	
Choose	Choose_name	?	Name1	Name2	-	Answer
	Choose_scene	-	Scene1	Scene2	-	
	Choose_color	?	Color1	Color2	-	
	Choose_shape	?	Shape1	Shape2	-	
	Choose_horizontal_position	?	H-Position1	H-Position2	-	
	Choose_vertical_position	?	V-Position1	V-Position2	-	
	Choose_relation_name	?	Relation1	Relation2	Name	
	Choose_relation_inverse_name	?	Relation1	Relation2	Name	
	Choose_younger	?	?	-	-	
	Choose_older	?	?	-	-	
	Choose_healthier	?	?	-	-	
	Choose_less_healthier	?	?	-	-	
Verify	Verify_color	?	Color	-	-	Answer
	Verify_shape	?	Shape	-	-	
	Verify_scene	-	Scene	-	-	
	Verify_relation_name	?	Relation	Name	-	
	Verify_relation_inv_name	?	Relation	Name	-	
Query	Query_name	?	-	-	-	Answer
	Query_color	?	-	-	-	
	Query_shape	?	-	-	-	
	Query_scene	-	-	-	-	
	Query_horizontal_position	?	-	-	-	
	Query_vertical_position	?	-	-	-	
Common	Common_color	[? ? .. ?]	-	-	-	Answer
	Common_material	[? ? .. ?]	-	-	-	
Different	Different_name	[? ? .. ?]	-	-	-	Answer
	Different_name	?	?	-	-	
	Different_color	?	?	-	-	
Same	Same_name	[? ? .. ?]	-	-	-	Answer
	Same_name	?	?	-	-	
	Same_color	?	?	-	-	
And	And	?	?	-	-	Answer
Or	Or	?	?	-	-	Answer
Exist	Exist	?	?	-	-	Answer

Figure 13. The function definitions and their corresponding outputs on GQA.