Supplementary Material for
ACRE: Abstract Causal REasoning Beyond Covariation

Chi Zhang Baoxiong Jia Mark Edmonds Song-Chun Zhu Yixin Zhu
UCLA Center for Vision, Cognition, Learning, and Autonomy
{chi.zhang,baoxiongjia,markedmonds}@ucla.edu, sczhu@stat.ucla.edu, yixin.zhu@ucla.edu

1. Symbolic Summary for Blicket Experiments

Query Type	Trials	Covar. of A	Covar. of B
D.R.	E$^+$ | A$^+$B$^-$ E$^+$ | A$^-$B$^+$	100%	100%
I.D.	E$^+$ | A$^+$B$^+$ E$^-$ | A$^+$B$^-$	50%	100%
S.O.	E$^+$ | A$^+$B$^-$ E$^-$ | A$^-$B$^+$ E$^+$ | A$^+$B$^+$	100%	50%
B.B.	E$^+$ | A$^+$B$^+$ E$^+$ | A$^+$B$^-$	100%	100%

Table 1: A symbolic summary for Figure 1 in the main text. In each of the query type, *i.e*., *direct* (D.R), *indirect* (I.D.), *screening-off* (S.O.), and *backward-blocking* (B.B.), we list the trials' configurations and covariation (Covar.) of each object with an activated machine. A trial's configuration is denoted as the combination of variables, where E represents the activation of the Blicket machine, A the attendance of object A, and B the attendance of object B, with $^+$ indicating activation or presence and $^-$ inactivation or absence. Covariance is computed as $P(E^+|X^+), X \in \{A, B\}$.

Table 1 symbolically summarizes the Blicket experiments demonstrated in Figure 1 in the main text. The simplest one conducted in Sobel *et al*. [7] is shown in the *direct* setting, where both objects are independently and always associated with an activated Blicket machine, and hence believed to be Blickets. Such a conclusion could be derived from the covariation of each object with an activated machine. Similarly, in the *indirect* setting, object B also shows perfect covariation with an activated machine, though its Blicketness is indirectly verified from the inactivation of object A. The behavior in object A in the *indirect* query and that of object B in the *screening-off* query are consistent: Despite half the chance of being associated with activation, their Blicketness is screened-off by another object from probabilistically setting the machine off. Note that the indirect setting is also referred to as indirect screening-off [2, 7]. In the *backward-blocking* query, both objects show perfect association with activation. However, object B's Blicketness is actually blocked by object A and cannot be solely determined from the observation. This is the case where we find most models, either purely neural or neuro-symbolic, catastrophically fail.

2. Label and Query Type Distributions

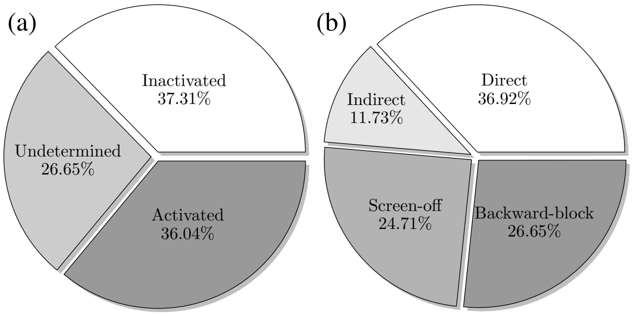


Figure 1: Distributions of (a) labels and (b) query types in the I.I.D. split of ACRE.

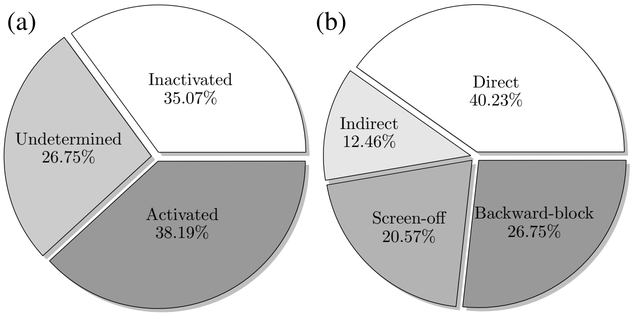


Figure 2: Distributions of (a) labels and (b) query types in the compositionality split of ACRE.

Figs. 1 to 3 show the label and query type distributions in the three splits of the ACRE dataset. Note that we keep the

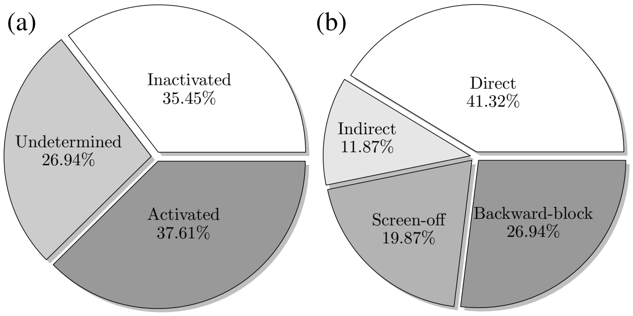


Figure 3: Distributions of (a) labels and (b) query types in the systematicity split of ACRE.

label distributions to be roughly uniform in order to avoid statistical bias. Around half of all queries are on screening-off and backward-blocking; these cases cannot be solved by simply calculating covariation.

3. Model Details

Operator	Params
Convolution	3-2-32
BatchNorm	32
ReLU	
Convolution	3-2-32
BatchNorm	32
ReLU	
Convolution	3-2-32
BatchNorm	32
ReLU	
Convolution	3-2-32
BatchNorm	32
ReLU	

Table 2: Network architecture used for the CNN module.

Operator	Params
Linear	512
ReLU	
Dropout	0.5
Linear	3

Table 3: Network architecture used for the MLP module.

Operator	Params
LSTM	128
Linear	3

Table 4: Network architecture used for the LSTM module.

Operator	Params
Transformer	8-1024-12-0.1
Linear	3

Table 5: Network architecture used for the BERT module.

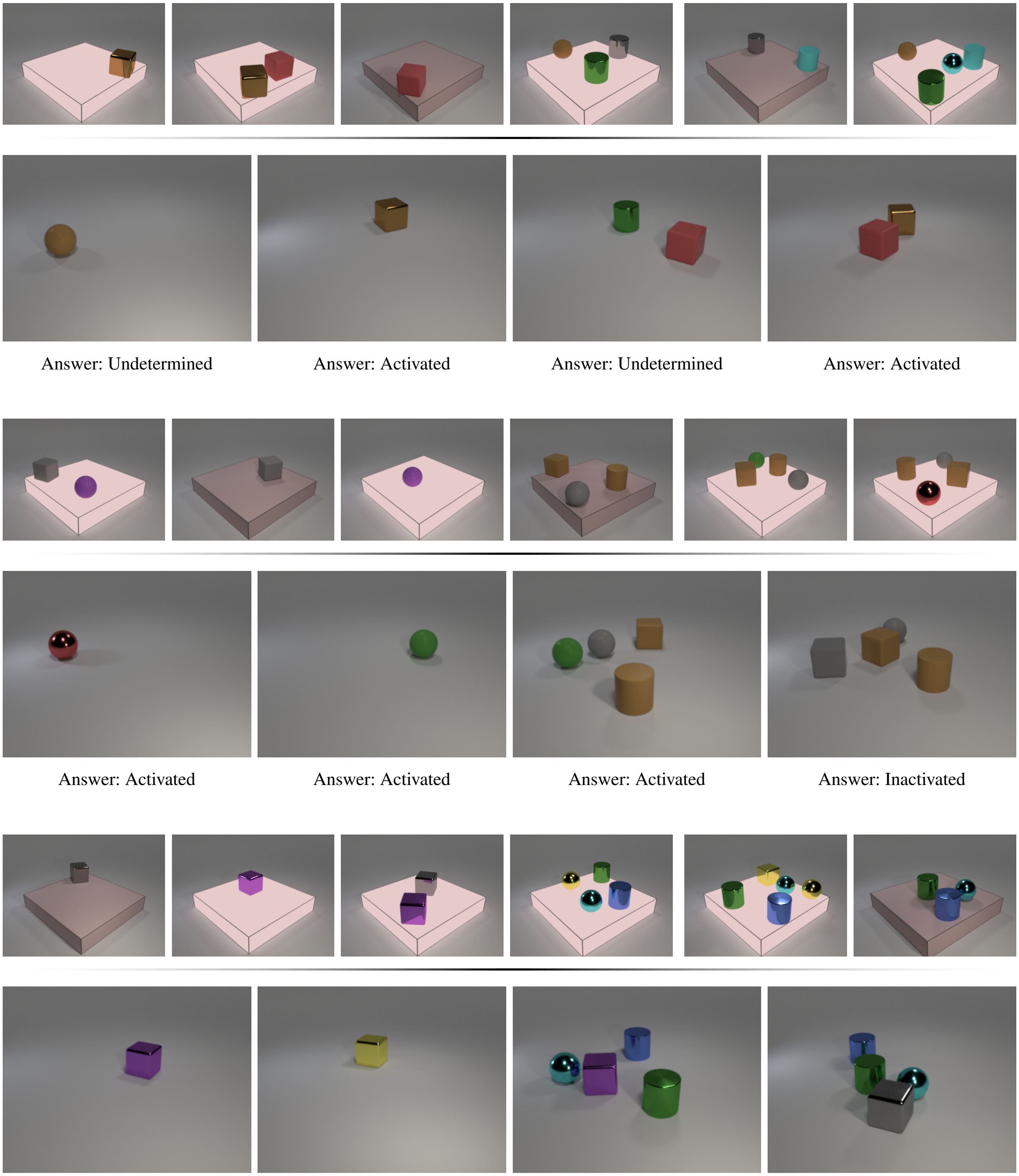
Operator	Params
Linear	10
Sigmoid	
Linear	1

Table 6: Network architecture used for each g_j in NS-Opt.

Table 2 details the CNN architecture used in various models we benchmarked. We use A-B-C to denote a convolution layer's parameters, where A refers to the kernel size, B the stride, and C the channel number. Table 3 shows the shared MLP architecture, where the final linear layer predicts the state of the Blicket machine, either inactivated, undetermined, or activated. For the LSTM module in Table 4, we use a single-layer LSTM and connects it with a linear layer to predict the final state. In Table 5, the BERT module [1] reuses the bidirectional Transformer layer [8], which is denoted by the number of heads, the size of the hidden space, the number of layers, and the rate of dropout. For ResNet [4] and WReN [6], we keep their network architectures as initially proposed. Modifications for LEN [11] and MXGNet [9] have been discussed in the main text. For neuro-symbolic models, we use the Mask RCNN [3] with ResNet-50 FPN [4, 5] in Detectron 2 [10] for scene parsing. The MLP module used for each g_j in NS-Opt is shown in Table 6. Note that during actual implementation, we combine all g_j into a single model and jointly optimize.

4. Additional Examples

Figs. 4 to 9 show additional examples of ACRE problems in the training sets and test sets of the I.I.D. split, the compositionality split, and the systematicity split, respectively.



Answer: Undetermined

Answer: Activated

Answer: Undetermined

Answer: Activated

Answer: Activated

Answer: Activated

Answer: Activated

Answer: Inactivated

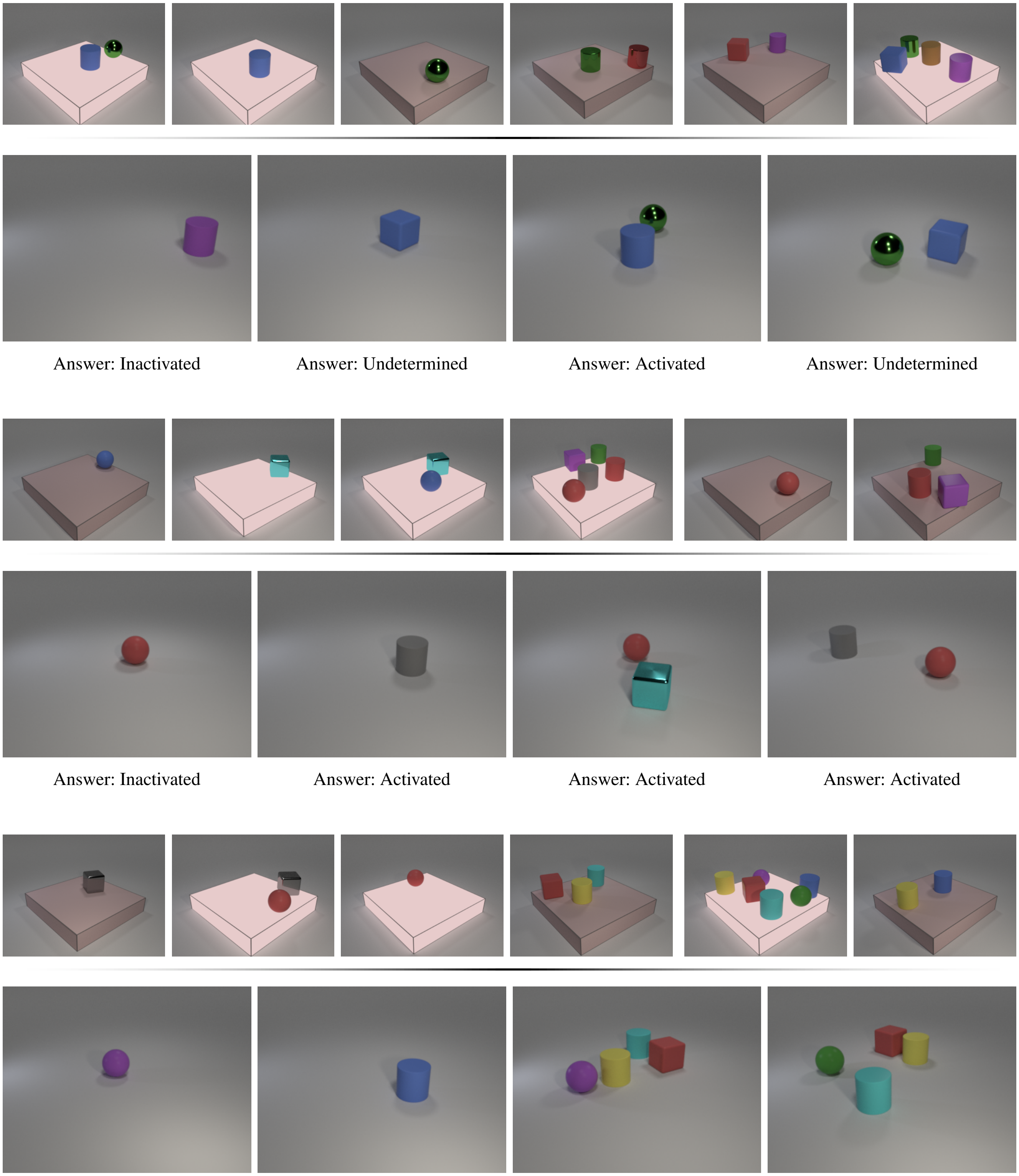
Answer: Activated

Answer: Undetermined

Answer: Activated

Answer: Inactivated

Figure 4: Examples in the training set of the I.I.D. split of ACRE. In each problem, we first show six context trials followed by four query trials.



Answer: Inactivated Answer: Undetermined Answer: Activated Answer: Undetermined

Answer: Inactivated Answer: Activated Answer: Activated Answer: Activated

Answer: Undetermined Answer: Inactivated Answer: Undetermined Answer: Undetermined

Figure 5: Examples in the test set of the I.I.D. split of ACRE. In each problem, we first show six context trials followed by four query trials.

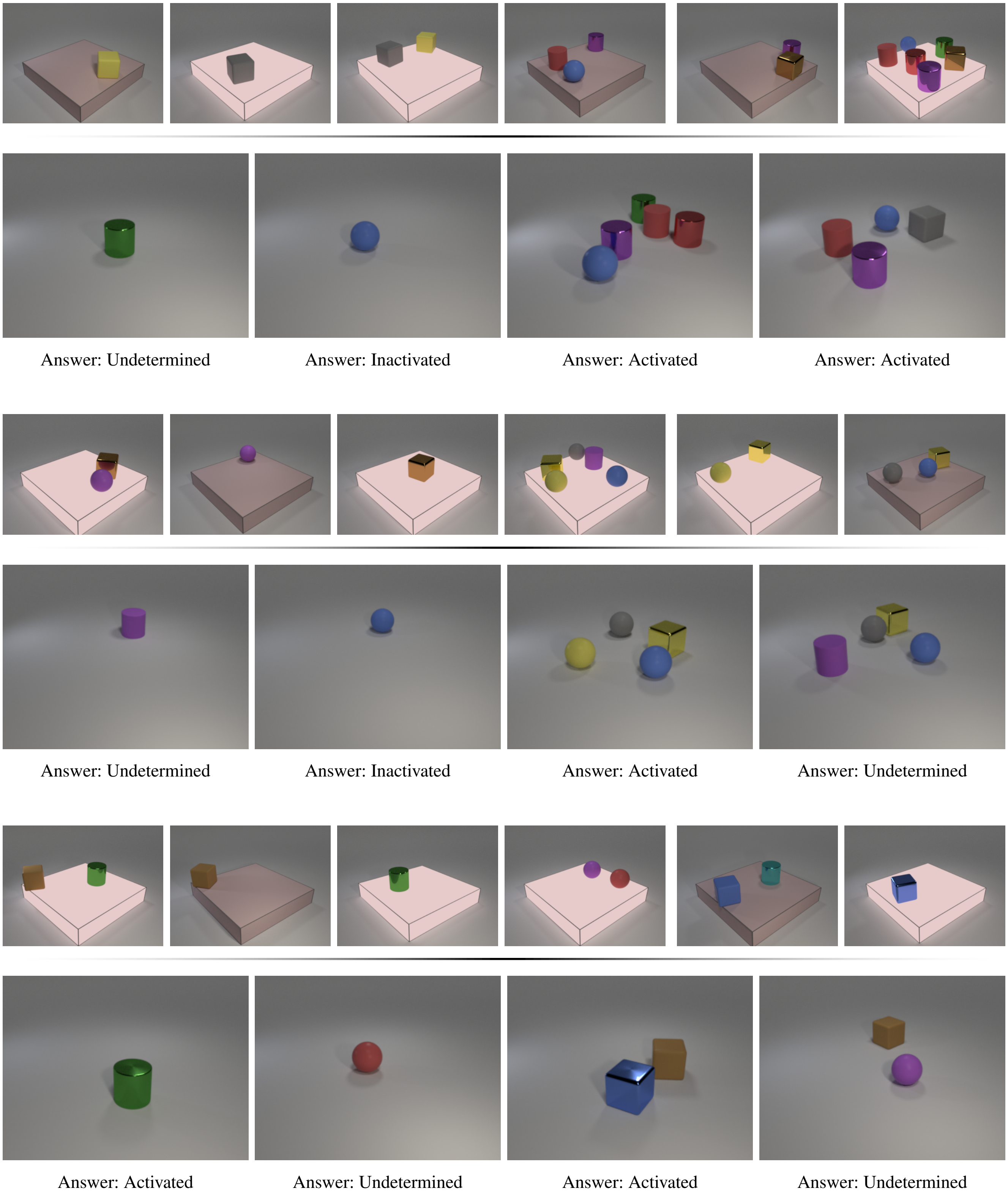
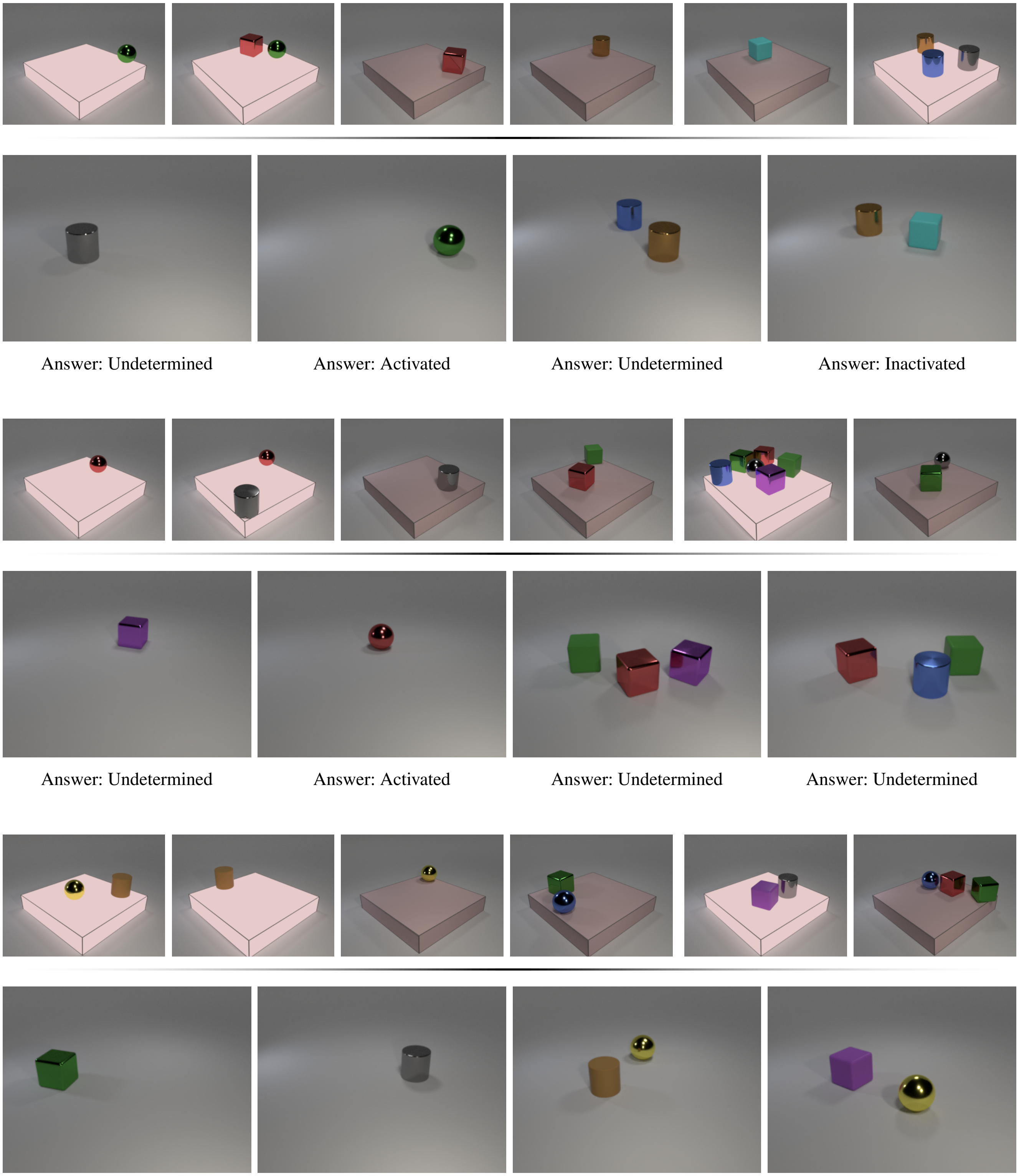


Figure 6: Examples in the training set of the compositionality split of ACRE. In each problem, we first show six context trials followed by four query trials.

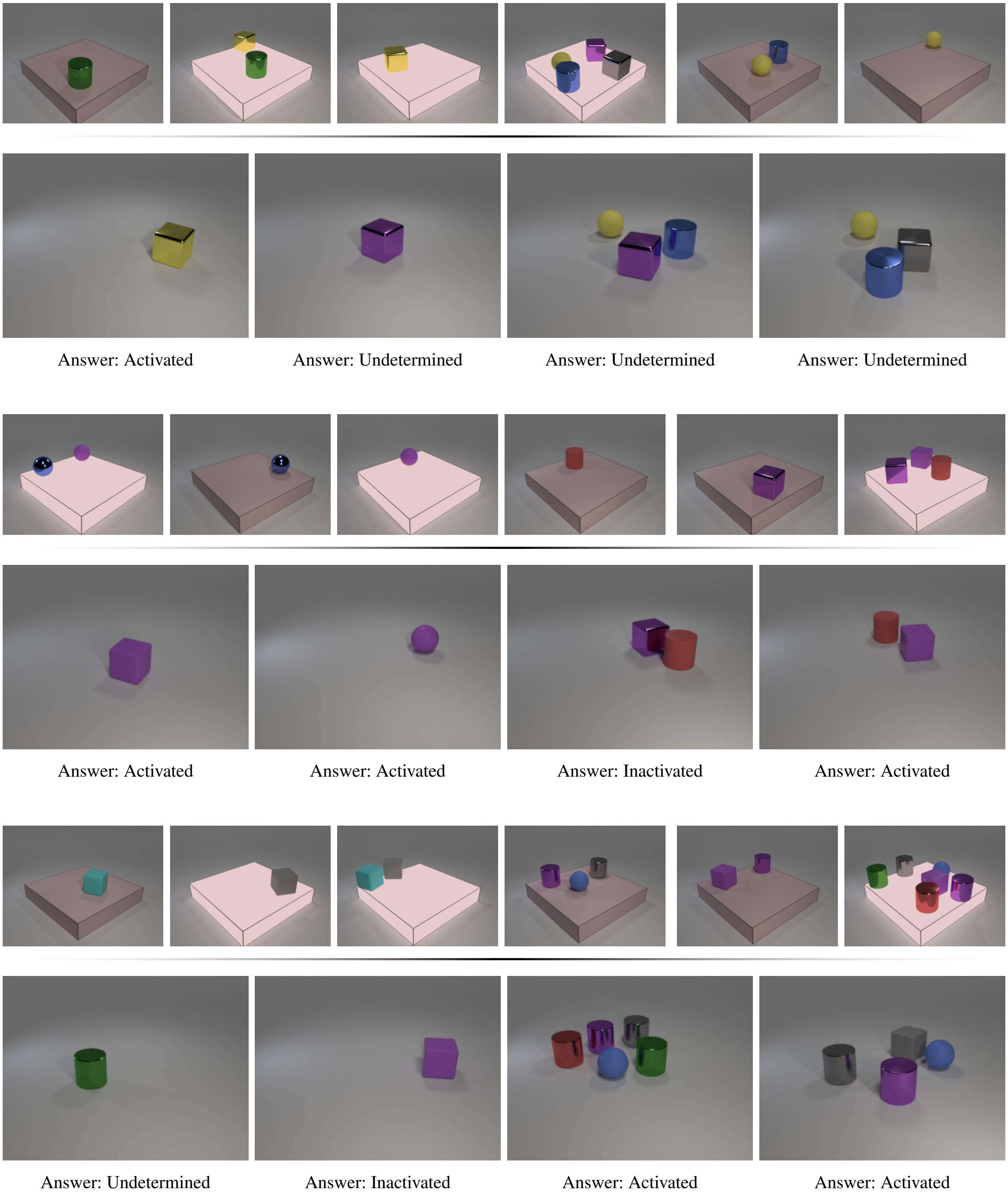


Answer: Undetermined Answer: Activated Answer: Undetermined Answer: Inactivated

Answer: Undetermined Answer: Activated Answer: Undetermined Answer: Undetermined

Answer: Inactivated Answer: Undetermined Answer: Activated Answer: Undetermined

Figure 7: Examples in the test set of the compositionality split of ACRE. In each problem, we first show six context trials followed by four query trials. Note that the attribute combinations in the test set are disjoint with those in the training set.



Answer: Activated

Answer: Undetermined

Answer: Undetermined

Answer: Undetermined

Answer: Activated

Answer: Activated

Answer: Inactivated

Answer: Activated

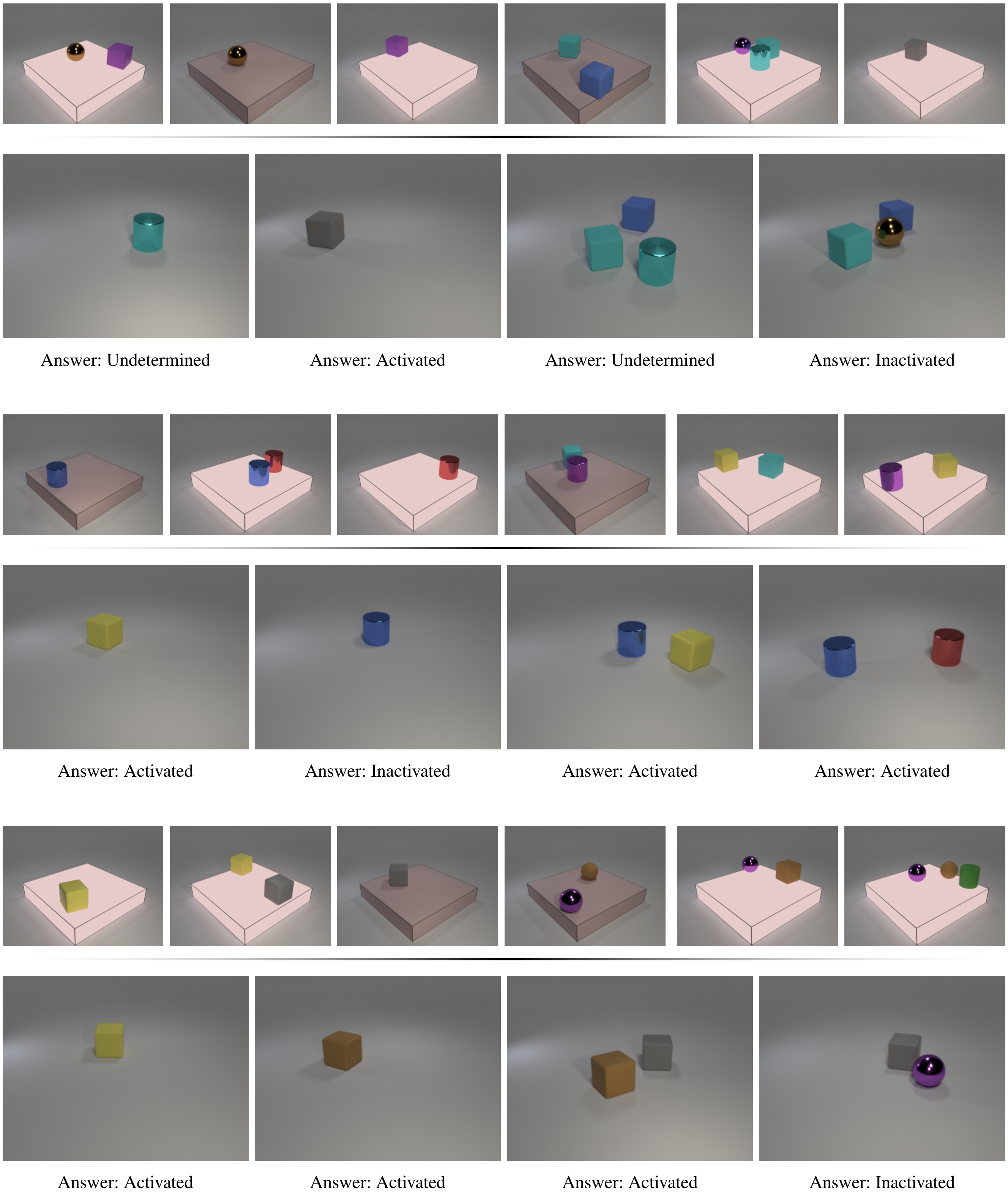
Answer: Undetermined

Answer: Inactivated

Answer: Activated

Answer: Activated

Figure 8: Examples in the training set of the systematicity split of ACRE. In each problem, we first show six context trials followed by four query trials.



Answer: Undetermined
Answer: Activated
Answer: Undetermined
Answer: Inactivated

Answer: Activated
Answer: Inactivated
Answer: Activated
Answer: Activated

Answer: Activated
Answer: Activated
Answer: Activated
Answer: Inactivated

Figure 9: Examples in the test set of the systematicity split of ACRE. In each problem, we first show six context trials followed by four query trials. Note the distributions of an activated machine are different in the training set and the test set, but the causal reasoning strategy remains the same.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019. 2

[2] Alison Gopnik, David M Sobel, Laura E Schulz, and Clark Glymour. Causal learning mechanisms in very young children: Two-, three-, and four-year-olds infer causal relations from patterns of variation and covariation. *Developmental psychology*, 37(5):620, 2001. 1

[3] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2017. 2

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[5] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[6] Adam Santoro, Felix Hill, David Barrett, Ari Morcos, and Timothy Lillicrap. Measuring abstract reasoning in neural networks. In *Proceedings of International Conference on Machine Learning (ICML)*, 2018. 2

[7] David M Sobel, Joshua B Tenenbaum, and Alison Gopnik. Children's causal inferences from indirect evidence: Backwards blocking and bayesian reasoning in preschoolers. *Cognitive science*, 28(3):303–333, 2004. 1

[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 2

[9] Duo Wang, Mateja Jamnik, and Pietro Lio. Abstract diagrammatic reasoning with multiplex graph networks. In *International Conference on Learning Representations (ICLR)*, 2020. 2

[10] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019. 2

[11] Kecheng Zheng, Zheng-Jun Zha, and Wei Wei. Abstract reasoning with distracting features. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 2