# CLEVR-X: A Visual Reasoning Dataset for Natural Language Explanations

**Leonard Salewski[1], A. Sophia Koepke[1], Hendrik P. A. Lensch[1], Zeynep Akata[1,3,4]**

[1]University of Tübingen, [2]MPI for Informatics, [4]MPI for Intelligent Systems

{leonard.salewski, a-sophia.koepke, hendrik.lensch, zeynep.akata}@uni-tuebingen.de

## Motivation

**Task:** Given an input image and question, generate an answer (Visual Question Answering) and a textual explanation.

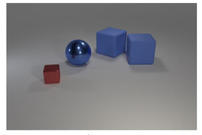
**Challenges:** Textual explanations in existing datasets for vision-language tasks

- require general knowledge
- do not explain the visual reasoning

Our **CLEVR-X dataset**:

- extends CLEVR *[1]* with **natural language explanations**;
- contains explanations that are by design **correct and complete**;
- is **larger** than all existing textual explanation datasets for vision-language problems.

**Question**: How many tiny red things are the same material as the big sphere?

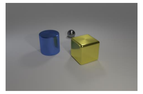


**GT Answer | Explanation**:
1 | The tiny red metal block has the same material as a big sphere.

## CLEVR-X dataset generation

CLEVR-X contains multiple **grammatically correct sentences** per CLEVR sample with the same attribute/shape synonyms used in the question and no redundancies.

### 1. CLEVR inputs



CLEVR scene graph

**Question:** What number of other objects are there of the same material as the tiny thing?

### 2. Tracing the functional program

| Program | Parameters | Tracing result |
|---|---|---|
| filter unique **<size>** | <tiny> | **<obj1>**: |
| same **<attribute>** | <material> | **<obj2>** : |
| count | | **<verb2> : are** |

### 3. Filtering relevant properties

× Size

× Color

✓ Material

✓ Shape

### 4. CLEVR-X explanation generation

**Template:** There **<verb2>** a **<obj2>** {that, which} have the {same, identical} **<attribute>** as the **<obj1>**.

**Explanation:** There **are** a **large yellow metallic cube and cylinder** that have the same **material** as the **tiny sphere**.

## CLEVR-X dataset analysis

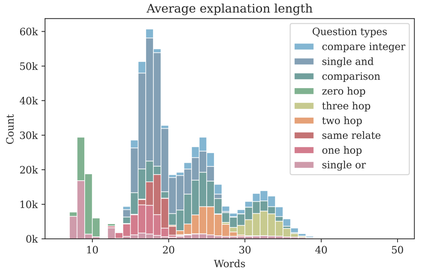
**Dataset properties:**

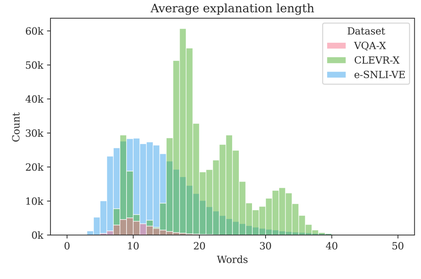
- Textual explanations:
  - 7 to 53 words
  - 21.5 words on average
- Vocabulary: 96 words
- 9 question types

**Dataset size:**

- CLEVR-X has
  - 85k images
  - 850k questions
  - 3.6m explanations

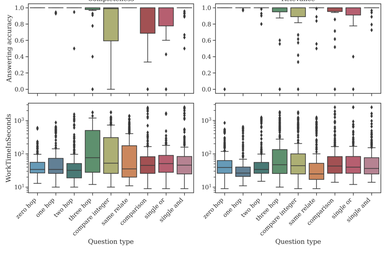
  → Larger than VQA-X *[2]* and e-SNLI-VE *[4]*
- Around 4 explanations for each CLEVR question




## User study

Human user study confirmed that the CLEVR-X explanations are **complete** (with all the evidence for answering) and **relevant**.

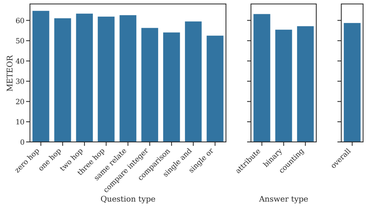



## Results

**Baseline results:** We provide results with PJ-X *[2]* and FM *[3]* on CLEVR-X.

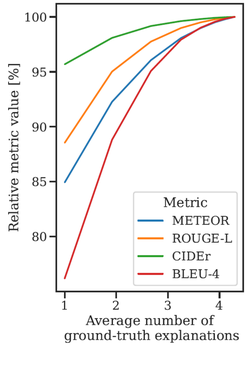
**Results for different question/answer types:**

- Counting answers have worse explanations than attribute answers.
- Explanations are worse for questions that require multiple reasoning steps.




**Multiple explanations per sample in CLEVR-X:**

- allow to analyze how the evaluation metrics behave depending on the number of ground-truth explanations used.
- A single ground-truth explanation for evaluation is not sufficient to capture whether a generated explanation is correct.




## Conclusion

- The **CLEVR-X dataset extends CLEVR** with textual explanations.
- CLEVR-X explanations are **correct and relevant by design** (unlike human explanations).
- We provide **baseline performances** with two standard VQA models.
- CLEVR-X allows to analyze the performance of trained models for different question/answer types.

**Paper, code & dataset**

*[1] J. Johnson et al.: CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In CVPR 2017*
*[2] D. H. Park et al.: Multimodal explanations: Justifying decisions and pointing to the evidence. In CVPR 2018*
*[3] J. Wu and R. J. Mooney: Faithful multimodal explanation for visual question answering. arXiv preprint arXiv:1809.02805*
*[4] M. Kayser et al.: e-ViL: A Dataset and Benchmark for Natural Language Explanations in Vision-Language Tasks. In ICCV 2021*