Supplementary Materials for IconQA: A New Benchmark for Abstract Diagram Understanding and Visual Language Reasoning

A The IconQA Dataset

The following datasheet follows the format suggested in this paper [13].

A.1 More Examples

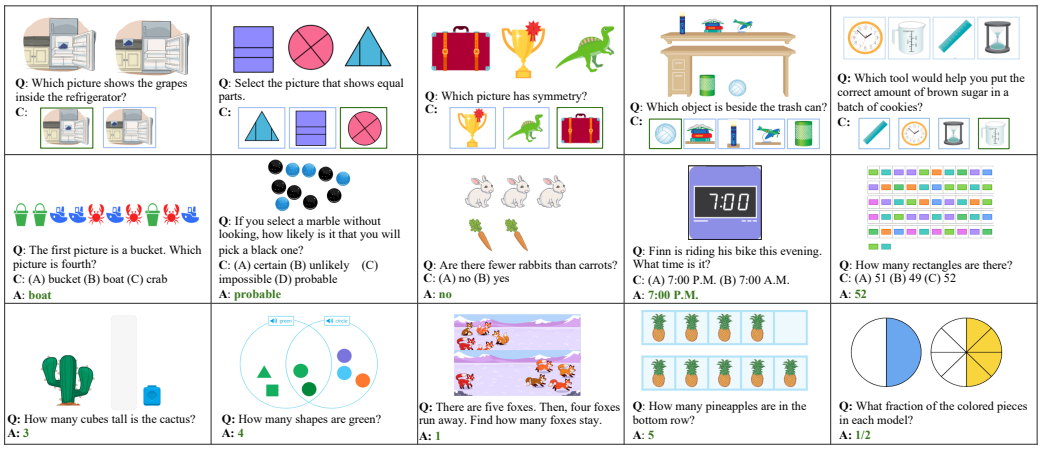

Figure 10: More examples in the IconQA dataset. **Top**: The *multi-image-choice* sub-task. **Middle**: The *multi-text-choice* sub-task. **Bottom**: The *filling-in-the-blank* sub-task.

A.2 Dataset Label

The IconQA dataset label is shown in Figure 11.

A.3 Question Skill Categories

The questions we collected contain meta-information including question topics, chapter names, image names, etc. After extensive data exploration by well-informed individuals, we designed a set of rules that map each question to 1-3 of the 13 categories based on trigger words in metadata. The rules for trigger words are list in Table 9.

A.4 Links

The link to download the IconQA dataset can be found at `iconqa.github.io`.

35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks.

IconQA Dataset Facts

Website	https://iconqa.github.io
Metadata	
Instances	107,439
Format	.png, .json
License	CC BY-NC-SA
Original Use Case	Training VQA systems
Composition	
Sample or Complete	Sample from ixl.com
Missing Data	No data is missing
Collection	
Sampling Strategy	See the main paper
Author Consent	None needed
Cleaning and Labeling	
Cleaning Done	Repetitions and redundancies removed
Labeling Done	Yes
Uses and Distribution	
Notable Uses	Training VQA systems
Original Distribution	Check dataset website
Maintenance and Evolution	
Corrections or Erratum	None as of now
Methods to Extend	Contact the author
Breakdown	% of Example
multi-image-choice 57,672 items	53.7%
multi-text-choice 31,578 items	29.4%
fill-in-the-blank 18,189 items	16.9%

Figure 11: IconQA dataset label, created with the template from the paper [4].

Table 9: Trigger words in metadata for skill categories.

Skill types	Trigger words in metadata
Geometry	name the shape, shapes of, classify shapes, solid, corners, faces, edges, vertices, sides, dimensional, rectangle, circle, triangle, square, rhombus, sphere, cylinder, cone, cubes, hexagon, perimeter, area, curved, open and close, flip turn, symmetry
Counting	count, tally, a group, ordinal number, area, even or odd, place value, represent numbers, comparing review, equal sides, square corners, one more, one less, fewer, enough, more.
Comparing	compare, comparing, more, less, fewer, enough, wide and narrow, light and heavy, long and short, tall and short, match analog and digital
Spatial	top, above, below, beside, next to, inside and outside, left
Scene	problems with pictures, beside, above, inside and outside, wide and narrow, objects
Pattern	the next, comes next, ordinal number, different
Time	clock, am or pm, elapsed time, times
Fraction	equal parts, halves, thirds, fourths, fraction
Estimation	estimate, measure
Algebra	count to fill, skip count, tally, even or odd, tens and ones, thousands, of ten, elapsed time, perimeter, area, divide
Measurement	measure
Commonsense	light and heavy, compare size, holds more or less, am or pm, times of, tool
Probability	likely

A.5 Motivation

- For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled?
  - IconQA is created to provide researchers with a wide range of VQA data on the abstract image domain. Currently, existing datasets 1) are limited to natural images, or 2) contain diagrams generated with templates, and therefore lack linguistic variation, or 3) include too much domain specific knowledge. We believe that no other abstract diagram QA dataset exists that covers such a wide range of perceptive and cognitive abilities without requiring complicated domain knowledge.

- Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?
  - This dataset was created under the combined effort of multiple researchers from University of California, Los Angeles, Sun Yat-sen University, East China Normal University, and Columbia University.
- Who funded the creation of the dataset?
  - The project received no funding or associated grant.

A.6 Composition

- What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)?
  - Each instance is a complete icon question answering task.
- How many instances are there in total (of each type, if appropriate)?
  - There are a total of 107,439 instances. 57,672 are *multi-image-choice* questions, 31,578 are *multi-text-choice* questions, and 18,189 are *filling-in-the-blank* questions.
- Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?
  - The dataset does not contain all possible instances.
- What data does each instance consist of?
  - Each instance in IconQA includes a textual question, an image, and multiple optional visual / textual choices. We also included some metadata about each question, such as the skill type, question type, etc.
- Is there a label or target associated with each instance?
  - Yes, each question is associated with a ground truth answer.
- Is any information missing from individual instances?
  - No. All related information is included in the dataset.
- Are there recommended data splits (e.g., training, development/validation, testing)?
  - Yes. Following conventions in the field, we have splitted the dataset into a training set, a validation set, and a test set with a 0.6:0.2:0.2 ratio.
- Are there any errors, sources of noise, or redundancies in the dataset?
  - We randomly selected 1,000 questions from each sub-task and ask an experienced expert to double check the answers carefully. Among the 1,000 *multi-image-choice* questions, only 1 error was found. Among the 2,000 questions of the other two sub-tasks, no error was found.
  - In the *multi-image-choice* sub-task, questions that ask "Which two are exactly the same?" might be a source of noise for certain use cases, as in the data label, only one correct answer out of the two is given. We intend to address this problem in the later versions of the dataset.
- Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?
  - The dataset is self-contained. All related information is included in the dataset.
- Does the dataset contain data that might be considered confidential?
  - No, the dataset does not contain anything related to any individual.
- Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?
  - No, the dataset does not contain anything offensive.

A.7 Collection Process

- How was the data associated with each instance acquired?
  - The data is publicly available on ixl.com. More details are included in the main paper.
- What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?
  - We implemented an integrated graphic user interface tool using Python to help crowd workers to collect the data.
- Over what timeframe was the data collected?
  - The dataset was finally completed in March, 2021 after three months of data collection, cleaning, and prepossessing.
- Were any ethical review processes conducted (e.g., by an institutional review board)?
  - No, we did not conduct an ethical review under the assumption that math and science questions designed for young children do not contain any discriminative or offensive content.
- Does the dataset relate to people?
  - No, the dataset does not relate to people.

A.8 Preprocessing

- Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?
  - We cropped white space from each diagram in IconQA to tighten it up. Questions with invalid diagrams, answers, or choices were filtered out. Redundant instances were removed based on the metrics of exact question text matching and diagram similarity.
- Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?
  - Yes, each QA data is accompanied with reasoning skill types and a grade level for comprehensive analysis of different benchmarks.
- Is the software used to preprocess/clean/label the instances available?
  - The data preprocessing and cleaning was done using Python.

A.9 Use Cases

- Has the dataset been used for any tasks already?
  - Yes, we developed a baseline model of cross-modal Transformers and multiple benchmarks for icon question answering, and we trained the models on the IconQA dataset. For more details, refer to Section 5 of the main paper.
- Is there a repository that links to any or all papers or systems that use the dataset?
  - Yes, you can access the code to our model at `github.com/lupantech/IconQA`.
- What (other) tasks could the dataset be used for?
  - Currently, the dataset is intended for training visual question answering systems to access the abilities of diagram upstanding and visual reasoning. More uses could be explored in research of computer vision, natural language processing, and multimodal learning, as well as applications in smart education like tutoring systems.
- Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?
  - No.

A.10 Distribution

- Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?
  - The dataset is free to all under the condition that the dataset is used for non-commercial purposes only.
- How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?
  - You can find our dataset both on the IconQA website iconqa.github.io, or the github repository github.com/lupantech/IconQA
- Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?
  - The dataset will be distributed under the CC BY-NC-SA (Attribution-NonCommercial-ShareAlike) license[3].
- Have any third parties imposed IP-based or other restrictions on the data associated with the instances?
  - The source of the data instances, IXL, does not allow the data to be used commercially.

A.11 Maintenance

- Who is supporting/hosting/maintaining the dataset?
  - The dataset is maintained by the paper's authors.
- How can the owner/curator/manager of the dataset be contacted?
  - The contact information of the authors can be found at the beginning of the main paper.
- Is there an erratum?
  - Currently, little errors have been found in the dataset. However, if errors were to be found, an erratum will be included in the repository.
- Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?
  - If the dataset were to be updated, all versions will be available on the dataset website.
- If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?
  - Please contact the author through email.

A.12 Novelty

IconQA presents new challenges in icon understanding and cognitive reasoning to many existing visual reasoning methods. 1) Icons feature intrinsic natures of abstract symbolism, varied styles, and ambiguous semantics, which differs from natural images significantly. 2) Since there is a lack of high-quality annotation data for icon diagrams, it restricts current mainstream data-driven visual methods to transfer smoothly to the icon domain. 3) As 107,439 questions in IconQA stem from real-world math word problems, it has made 13 different cognitive reasoning skills essential, including spatial reasoning, commonsense reasoning, estimation, and arithmetic calculation.

A.13 Limitations and Future Work

Dataset Expansion. As discussed in Section 3, IconQA focuses on colored abstract diagrams and questions of third grade and below to simplify the context scenarios and attract the community's attention on diagram understanding and visual reasoning. We would like to expand the dataset to provide greater diversity of diagram formats, grade levels, icon classes and reasoning skill types.

Fine-grained Annotations. IconQA benchmarks the visual question answering task in the icon domain and releases a dataset of questions, diagrams and answers. But it would be beneficial to include the object-level parsing annotations and textual explanations for each diagram and question, which facilitates future research on semantic diagram parsing and transparent visual reasoning.

[3] https://creativecommons.org/licenses/by-nc-sa/4.0

B The Icon645 Dataset

The following datasheet follows the format suggested in this paper [13].

B.1 Dataset Statistics

Table 10: Statistics for the Icon645 dataset.

Data	#Classes	#Icons	Min Size	Max Size	Colored
Icon645	377	645,687	64×64	256×256	✓

B.2 Dataset Label

Icon645 Dataset Facts

Website	https://iconqa.github.io
Metadata	
Classes	377
Instances	645,687
Format	.png
Image Sizes	64× 64 - 256 × 256
License	CC BY-NC-SA
Original Use Case	Training visual encoder
Composition	
Sample or Complete	Sample from flaticon.com
Missing Data	No data is missing
Collection	
Sampling Strategy	See the main paper
Author Consent	None needed
Cleaning and Labeling	
Cleaning Done	Repetitions and redundancies removed
Labeling Done	Yes
Uses and Distribution	
Notable Uses	Pre-training abstract icon image encoder
Other Uses	Abstract image classification, transfer learning
Original Distribution	Check dataset website
Maintenance and Evolution	
Corrections or Erratum	None as of now
Methods to Extend	Contact the author

Figure 12: Icon645 dataset label, created with the template from the paper [4].

B.3 Links

The link to download the IconQA dataset can be found on iconqa.github.io.

B.4 Motivation

- For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled?
  - Icon645 was created for the purpose of pre-training image encoders on the icon image classification task. Presently, no other dataset provides such a large variety of abstract icons with appropriate labels.
- Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?
  - This dataset was created under the combined effort of multiple researchers from University of California, Los Angeles, Sun Yat-sen University, East China Normal University, and Columbia University.

- Who funded the creation of the dataset?
  - The project received no funding or associated grant.

B.5 Composition

- What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)?
  - Each instance is a single colored icon image with size between 64×64 and 256×256 pixels.
- How many instances are there in total (of each type, if appropriate)?
  - There are a total of 645,687 instances categorized into 377 classes.
- Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?
  - The dataset is a sample of the Flaticon library. Only 377 classes of icons that satisfy our requirements outlined in the paper are included in the dataset.
- What data does each instance consist of?
  - Each instance is a single icon image
- Is there a label or target associated with each instance?
  - Yes, Each image is given a text label, specifying its class.
- Is any information missing from individual instances?
  - No. All related information is included in the dataset.
- Are there recommended data splits (e.g., training, development/validation, testing)?
  - No. The user can decide how they want to split the dataset.
- Are there any errors, sources of noise, or redundancies in the dataset?
  - No.
- Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?
  - The dataset is self-contained. All related information is included in the dataset.
- Does the dataset contain data that might be considered confidential?
  - No, the dataset does not contain anything related to any individual.
- Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?
  - No, the dataset does not contain anything offensive.

B.6 Collection Process

- How was the data associated with each instance acquired?
  - The data is publicly available on flaticon.com. More details are included in the main paper.
- What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?
  - We implemented a program to retrieve the target 377 icon classes using Python.
- Over what timeframe was the data collected?
  - The dataset was finally completed in March, 2021 after three months of data collection, cleaning and prepossessing.
- Does the dataset relate to people?
  - No, the dataset does not relate to people.

B.7 Preprocessing

- Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?
  - We cropped white space from each icon diagram in Icon645 to tighten it up. Black and white icons were filtered out. Redundant instances were removed based on the metric of diagram similarity.
- Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?
  - No.
- Is the software used to preprocess/clean/label the instances available?
  - The data preprocessing and cleaning was done using Python.

B.8 Use Cases

- Has the dataset been used for any tasks already?
  - Yes, we have used the dataset to pre-train an abstract image encoder to act as the backbone network in our Patch-TRM model.
- Is there a repository that links to any or all papers or systems that use the dataset?
  - Yes, you can access the code to our model at github.com/lupantech/IconQA.
- What (other) tasks could the dataset be used for?
  - Currently, the dataset is intended for training abstract icon image classifiers. Other possibilities could be explored in the future.
- Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?
  - No.

B.9 Distribution

- Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?
  - The dataset is free to all under the condition that the dataset is used for non-commercial purposes only.
- How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?
  - The dataset will be accessible on github.com/lupantech/IconQA
- Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?
  - The dataset will be distributed under the CC BY-NC-SA (Attribution-NonCommercial-ShareAlike) license[4].
- Have any third parties imposed IP-based or other restrictions on the data associated with the instances?
  - The source of the data instances, Flaticon, does not allow the data to be used commercially.

B.10 Maintenance

- Who is supporting/hosting/maintaining the dataset?
  - The dataset is maintained by the paper's authors.
- How can the owner/curator/manager of the dataset be contacted?
  - The contact information of the authors can be found at the beginning of the main paper.

[4] https://creativecommons.org/licenses/by-nc-sa/4.0

- Is there an erratum?
  - Currently, no error has been found in the dataset. However, if errors were to be found, an erratum will be included in the repository.
- Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?
  - If the dataset were to be updated, all versions will be available on the dataset website.
- If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?
  - Please contact the author through email.

C Details of Baseline Patch-TRM

We develop a patch cross-modal Transformer model (Patch-TRM) as a strong baseline for the IconQA task as illustrated in Figure 7. We will introduce the details of Patch-TRM as follows.

C.1 Diagram Encoder

Similar to natural images in most VQA datasets, abstract diagrams also have rich visual and semantic information that is critical to answering questions. Current dominant VQA methods [3, 2, 28, 11, 54, 19, 1] either extract high-level visual representations from a pre-trained ResNet backbone network [16] in a top-down fashion, or apply a bottom-up mechanism to extract semantic representations via a object detector, such as a model based on Faster R-CNN [42]. However, these methods depend heavily on the backbone network, which is pre-trained on natural images. When processing diagrams in IconQA, they are likely to fail to extract meaningful representations or reasonable object proposals. Inspired by the early progress in using hierarchical scene layout to parse images [32, 56, 51] and the recent advances in Transformer-based image encoding [36, 33, 53], we develop a method that splits diagrams into hierarchical patch sequences from a pyramid structure and learns their visual representations using a visual Transformer.

As diagrams in IconQA have more varied aspect ratios than natural images, we add blank paddings at the bottom or on the right side of the images to ensure that they are square-shaped. Each padded diagram is then cropped into a set of patch sequences with different scales. The padding operation and the hierarchical scene layout can facilitate extracting complete objects that retain specific semantics. Let $p = [p_1, p_2, \ldots, p_n]$ denote the patch sequence in the splitting order from the original diagram. From each patch sequence, we extract the visual features using a ResNet model and represent the features as $f_p = [f_{p_1}, f_{p_2}, \ldots, f_{p_n}]$. The representation for each patch, f_{p_i}, is then summed up with its positional embedding with respect to its sequencial index i. Finally, the updated visual patch embeddings pass through a standard multi-layer Transformer [50] to learn high-level visual representations $h_p = [h_{\texttt{[CLS]}}, h_{p_1}, h_{p_2}, \ldots, h_{p_n}]$. Here, the trainable token $[\texttt{CLS}]$, which is added to the Transformer inputs, learns the global meaning of these sequences. As mentioned before, it is not feasible to use existing pre-trained ResNet to process abstract diagrams due to a lack of similar resources for pre-training. So we pre-train the ResNet on icon classification with the icon dataset we compiled (Section 4). More details of the pre-training task are discussed in Section 6.4.

C.2 Language Encoder

Questions in IconQA have a wide distribution of question lengths, so we follow the recent approaches [50, 20, 49, 33, 36] that apply the BERT model [9] to embed question texts, rather than using traditional LSTM [17] or GRU [7] for long sequence encoding. Given the question $w_0, w_1, \ldots, w_t$, the input is formatted as $[[\texttt{CLS}], w_0, w_1, \ldots, w_t]$. We use the WordPiece [45] subword tokenizer and the resulting sequence is padded to the maximum length. Similar to other methods that use BERT as sentence encoders, we consider the output corresponding to the first token $[\texttt{CLS}]$ as the embedding of the entire question, noted as h_q.

C.3 Answer Reasoning

Given the image patch representation $h_p \in \mathcal{R}^{n \times k}$, and question embedding $h_q \in \mathcal{R}^k$, where n denotes the number of diagram patches and k denotes the learned embedding size of the patches, we

apply a cross-modal attention to learn their joint representation:

$$a = \text{softmax}\left(W_p h_p \circ W_q h_q\right), \tag{1}$$

$$h_v = \sum_i^n a(i) \times h_{p_i}, \tag{2}$$

where W_p and W_q are learnable mapping parameters, and $\circ$ is the element-wise product operator. The joint representation h_v is calculated as the weighted sum over all diagram patches.

Before predicating the answer, multiple choice candidates need to be encoded. Taking the *multi-image-choice* task as an example, each image choice is encoded as the output of the last pooling layer of the pre-trained ResNet. The encoded image choice is denoted as $h_c \in \mathcal{R}^{m \times k}$, where m is the number of the candidates. The choice embeddings are concatenated with the diagram-question representation, and then the resulted embeddings are fed to a classifier over the candidates:

$$p_{ans} = \text{softmax}\left(W_a\left([h_v; h_c]\right) + b_a\right), \tag{3}$$

where W_a and b_a are classifier parameters, and p_{ans} is the probability of the predicated answer choice.

Similarly, in the *multi-text-choice* sub-task, the answer is predicated over text choices, except that each text choice is embedded with LSTM layers first. We formulate the *filling-in-the-blank* sub-task as a multi-class classification problem from all possible answers in the training data, as most VQA works do. After generating the joint encoding for the input diagram and question, a linear classifier is trained to predict the final answer.

D User Study

D.1 Crowd Sourcing Method

Using Amazon Mechanical Turk (AMT), we ask people to provide answers to the questions in the test set along with their age group. We also strongly encourage parents who have young children to let their children complete the questionnaires, as their answers give us insights to how the designed audience of these questions perform. The test set is split into batches of 20 questions, which we call a task, with each task assigned to 3 crowd workers on AMT. This amounts to a total of 64,467 effective test set answers.

D.2 Quality Assurance

To ensure the truthfulness of the age information, we ask the participants to select their age at both the beginning and the end of the questionnaire, with the age choices appearing in 2 different orders.

To ensure the quality of the answers, we include 4 attention check questions: 3 of which are about the instructions, making sure that the participants read the instructions carefully. We also add an extra fake question in the middle for each *choosing an image choice* and *choosing a text choice* task, instructing them to choose the fourth choice despite what the choices are. Figure 13 shows the instructions and the first three attention check questions. Figure 14 shows the fake question along with the age confirmation. Figure 15, 16, and 17 are example questions for three sub-tasks respectively. We also make sure that the workers answering our tasks have a history HIT approval rate of at least 95% and a previous approval count of 1,000.

In summary, for each Human Intelligence Task (HIT) on AMT, we have 2 age questions, 4 attention check questions, and 20 real questions from the IconQA test set. Among the 64,467 test answers, we filter out 1) the questionnaires that do not pass the 4 attention check questions, 2) the questionnaires that do not answer consistently for the two age-related questions, 3) the questionnaires that are finished unreasonably slowly/quickly. After filtering, we have 54,896 effective question answers, which we believe is a decently large sample for the human performance study.

D.3 Worker Compensation

For each batch of *multiple choice* questions, we provide a monetary compensation of 10 US cents. For each batch of *filling-in-the-blank* questions, we provide a compensation of 20 cents due to the

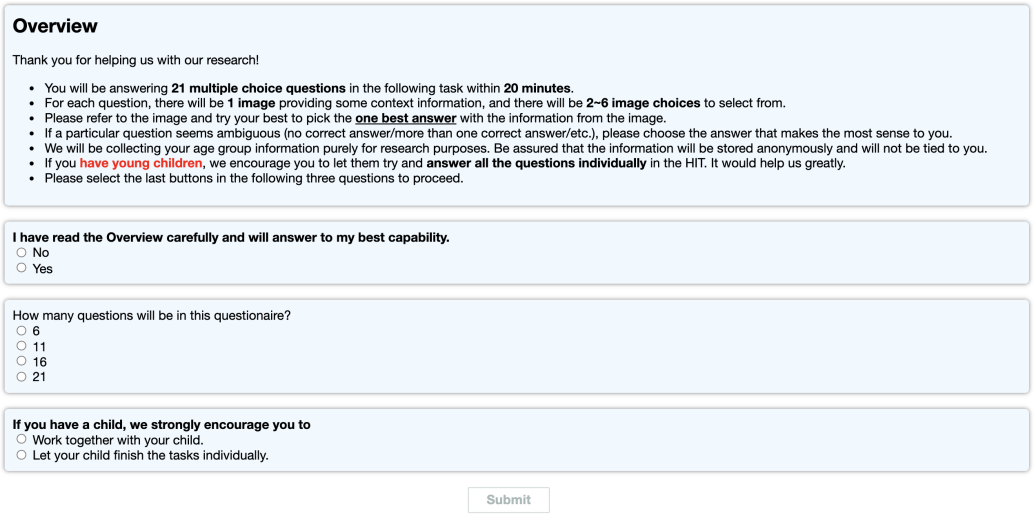

Figure 13: AMT instructions for the user study.

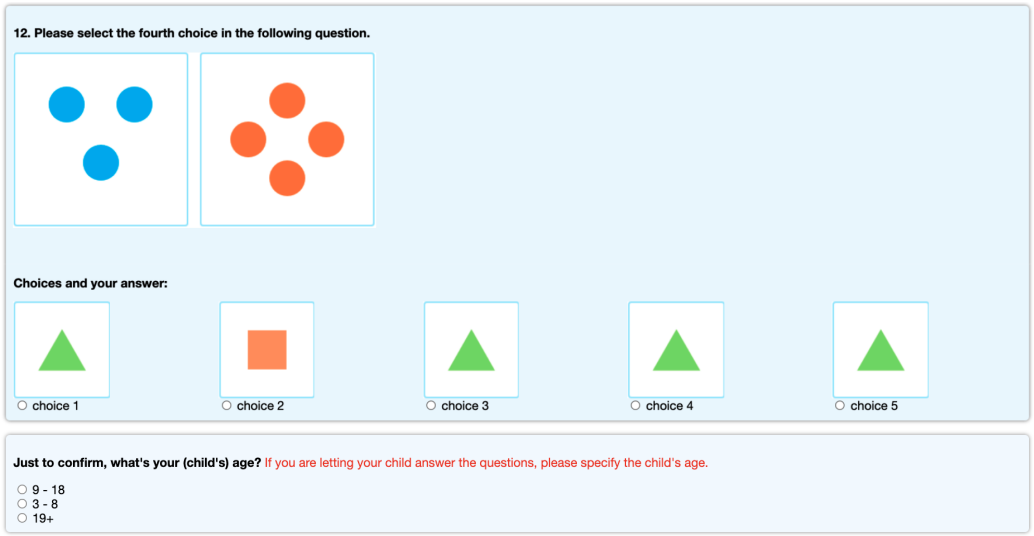

Figure 14: AMT attention check questions.

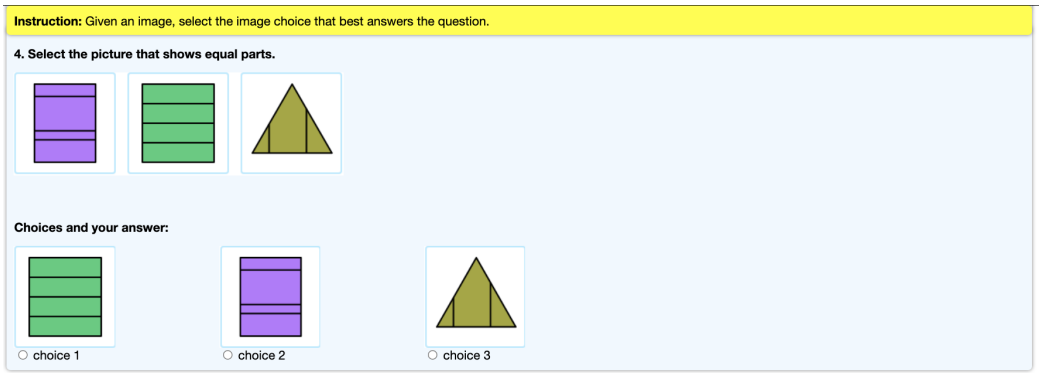

Figure 15: An AMT question example for the *multi-image-choice* sub-task.

increased difficulty. We decide upon these numbers after a few timed test trials run by ourselves. we

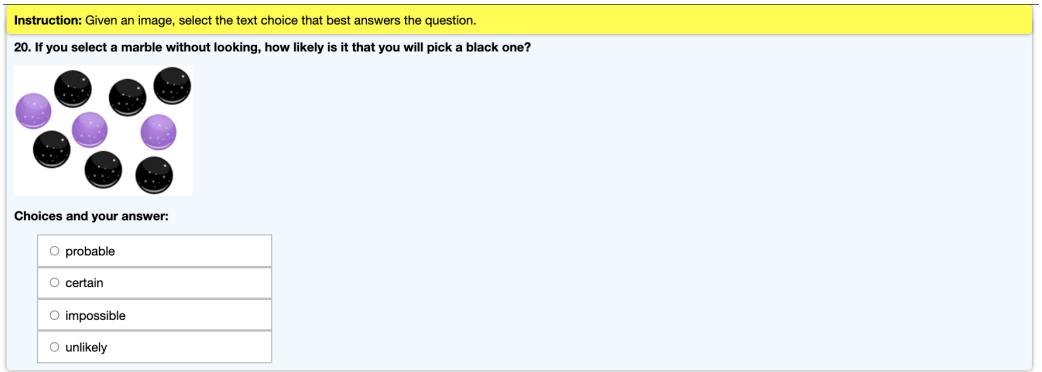

Figure 16: An AMT question example for the *multi-text-choice* sub-task.

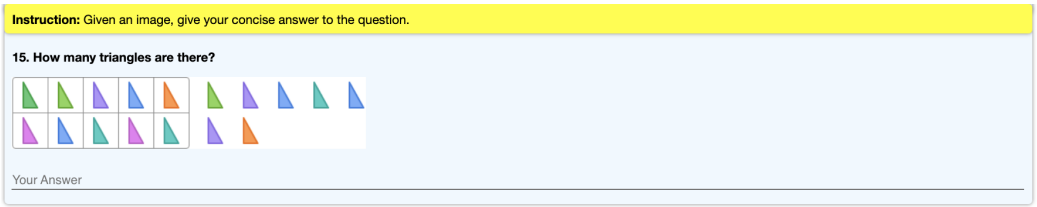

Figure 17: An AMT question example for the *filling-in-the-blank* sub-task.

find that these numbers enable the workers to acquire 6 USD per hour, an above average hourly wage on the AMT platform [15]. The total spending in the end sums up to 452.52 USD.

E Experiments

E.1 Experimental Details

We use the same learning parameters set in Top-Down [2] when evaluating the eight benchmarks listed in Section 5 and our developed baseline Patch-TRM. Some crucial parameters used in our model are clarified below.

Our Baseline Model. For our baseline Patch-TRM, each diagram is split four times by varied scales, resulting in 79 (1+4+9+16+49) patches totally. After resizing them to to 224×224, patch visual features are extracted from the last pooling layer, resulting in a 2048-d feature vector. The ResNet network used to embed the patches is pre-trained on the icon classification task as discussed in Section 6.4. The patch Transformer has one layer of Transformer block with four attention heads and outputs embeddings with a hidden state size of 768. A small pre-trained BERT model [49] is used to encode the question text in the language encoder.

Attention models. For Top-Down, the attention-based baselines use 7×7×2048-d features from the last convolution layer. For BAN [28], DFAF [11], and MCAN [54], image features of dimension 2,048 are extracted from Faster R-CNN [42]. Question words in these attention models are encoded into features of dimension 1,024 by GRU [7]. And the visual and textual features are then embedded into 1,024 dimensions with the corresponding attention mechanisms and fusion methods reported in original works.

Transformer models. For ViLBERT [36] and UNITER [6], we use Faster R-CNN [42] to extract 36 proposal regions as the visual inputs. Both ViL [53] and ViLT [29] use ViT-B/32 pre-trained on ImageNet to encode the image emebeddings. The hidden size is set as 768, the layer depth is 32, and the input image is sliced into patches with a size of 32. For ViL, we use two dependent Transformers to embed the question and image respectively.

E.2 Human Performance

The detailed results for human performance in the IconQA task are shown in Table 11.

Table 11: Human performance in the IconQA task.

	Sub-tasks (3)			Reasoning skills (13)												
Method	Img.	Txt.	Blank	Geo.	Cou.	Com.	Spa.	Sce.	Pat.	Tim.	Fra.	Est.	Alg.	Mea.	Sen.	Pro.
Human	95.69	93.91	93.56	94.63	97.63	94.41	93.31	92.73	95.66	97.94	97.45	87.51	96.29	86.55	97.06	85.67
Human (3-8)	94.58	89.51	89.61	93.02	96.20	91.28	91.24	90.45	95.76	95.32	97.54	78.86	95.33	78.57	93.92	74.76
Human (9-18)	94.63	90.97	93.71	93.28	97.04	93.46	91.47	90.92	94.55	97.59	96.77	86.79	95.83	86.60	96.51	80.56
Human (19+)	97.34	95.83	94.22	96.27	98.44	96.17	96.31	95.85	96.34	98.96	97.95	89.59	96.84	88.00	98.49	90.82

E.3 Quantitative Analysis

Figure 18 presents five examples from the IconQA test set predicted by our Patch-TRM baseline for each sub-task. Although Patch-TRM achieves promising results for most problems in IconQA, it still fails to address some complicated cases. For example, it encounters difficulties in identifying dense objects and making multi-hop reasoning.

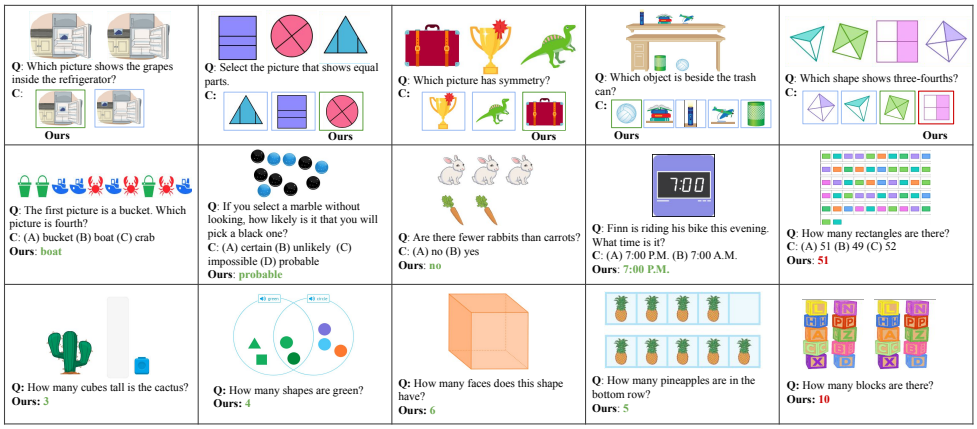

Figure 18: Result examples predicted by our Patch-TRM model in the IconQA test set. **Top**: *Multi-image-choice* sub-task. **Middle**: *Multi-text-choice* sub-task. **Bottom**: *Filling-in-the-blank* sub-task. Correctly predicted answers are highlighted by green, while wrong ones are highlighted by red.