

# Supplementary Materials for VisualHow: Multimodal Problem Solving

Jinhui Yang\*    Xianyu Chen\*    Ming Jiang    Shi Chen    Louis Wang    Qi Zhao  
University of Minnesota

{yang7004, chen6582, mjiang, chen4595, wangx723}@umn.edu, qzhao@cs.umn.edu

## 1. Introduction

The main text has introduced VisualHow – a free-form and open-ended research that focuses on understanding real-life problems and deriving solutions by integrating information over multiple modalities. The supplementary materials support our main findings with further evidence and report additional details of the VisualHow dataset and baseline models:

- (1) Sec. 2 presents more detailed analyses and visualizations of our VisualHow dataset;
- (2) Sec. 3 presents the Amazon Mechanical Turk (AMT) interface for annotating the VisualHow dataset.
- (3) Sec. 4 presents the implementation details of all baseline models used in the experiments.

## 2. Dataset Visualizations

In this section, we present example problems and solutions, visualizations of problem categories, as well as the distribution of word tokens in the VisualHow dataset.

### 2.1. Data Examples

Fig. 1 shows the extended examples (see Fig. 2 of our main paper) with complete solutions. For illustration purpose, all of the presented examples have four solution steps. The image and caption of each step are shown along with their corresponding attention annotations. As shown in these examples, our dataset consists of a variety of attention annotations, such as objects (*e.g.*, mice, food, water, dog, protective gloves, algaecide), abstract concepts (*e.g.*, “sit” command, social skill words, healthy social behavior), and actions (*e.g.*, take, give, reward, wear, role play). The corresponding visual attention (image regions) and textual attention (highlighted phrases) share the same color. For example, in Fig. 1b, the dog in the images is associated with the word “dog” in the captions. It is noteworthy that the dependencies between different steps form the solution graph, which is unique for each example. As shown in Fig. 1a,

to “care for pet during vacation”, step 1 must be completed first, because all the following steps depend on the decision made in step 1. However, steps 2-4 do not depend on each other. Differently, in Fig. 1b, all steps must be completed following a sequential order. These examples demonstrate the diverse annotations of solution graphs that enable the understanding of different types of solutions.

### 2.2. Categorical Analyses

Fig. 2 shows the number of subcategories in each category, ordered by the number of samples in each category. We find that higher number of samples per category does not necessarily indicate more subcategories. For example, Health has the highest number of subcategories, followed by Pets and Animals, while Sports and Fitness has the least. The different amount of subcategories reflects the original distribution of the wikiHow knowledge base.

### 2.3. Word Distributions

Fig. 3 presents the word distributions for frequent tokens used in the VisualHow problems, solutions, and textual annotations. It shows the overall distributions as word clouds (first row), while highlighting the top-20 most frequent tokens (second row) and the distribution of (nouns, verbs, and other POS (parts of speech)). We find that all three types of text data share a similar distribution of nouns, verbs, and other types of POS, with slightly more nouns in the annotation text as well as less amount of other POS types. It is also noteworthy that although nouns are almost twice as many as verbs, a majority of the top-20 most frequent words are verbs. It suggests that both actions and objects are important for understanding and solving problems.

## 3. AMT Interface

Fig. 4 presents the complete AMT (Amazon Mechanical Turk) interface for the data annotation. Given a wikiHow problem, we first ask the crowd to annotate multimodal attention in the images and captions. These annotations include 1) important words with a corresponding region in

\*Equal contributions.

the image, 2) important words without any image correspondence, and 3) important image regions without corresponding words in the caption. The workers then review the multimodal attention annotations. The full solution graph is annotated by selecting dependencies of each step. In addition, they review the images and annotate the image types, and finally provide their feedback.

#### 4. Experimental Details

In this section, we briefly describe the implementation details of our baseline models. Without loss of generalization, for each modality, *e.g.*, vision or language, we use  $\mathbf{v}_1, \dots, \mathbf{v}_K$  as the grid features from the ResNet [4] or the language features from the transformer [3], where  $K$  is the number of features. Each feature is represented as  $\mathbf{v}_i \in \mathbb{R}^V$ , where  $V$  is the feature dimension. In the following context, we will use superscripts to represent the different sources of the extracted feature. More specifically,  $\mathbf{v}_k^i$ ,  $\mathbf{v}_k^c$  and  $\mathbf{v}_k^g$  represent the  $k$ -th grid feature of the image, the  $k$ -th language feature of the caption and the problem description, respectively.

**GAP.** For the global average pooling baseline, it computes the aggregated feature  $\hat{\mathbf{v}}^i$  by calculating the average feature of each modality as follows:

$$\hat{\mathbf{v}}^i = \frac{1}{K} \sum_{k=1}^K \mathbf{v}_k^i, \quad (1)$$

$$\hat{\mathbf{v}}^c = \frac{1}{K} \sum_{k=1}^K \mathbf{v}_k^c, \quad (2)$$

$$\hat{\mathbf{v}}^g = \frac{1}{K} \sum_{k=1}^K \mathbf{v}_k^g, \quad (3)$$

where  $\hat{\mathbf{v}}^i$ ,  $\hat{\mathbf{v}}^c$  and  $\hat{\mathbf{v}}^g$  are the aggregated image features, caption features and problem features, respectively.

**GPO.** For the generalized pooling operation baseline, we follow the method of [2] (see Sec. 3.2 and Sec. 3.3) to compute the aggregated features  $\hat{\mathbf{v}}^i$ ,  $\hat{\mathbf{v}}^c$  and  $\hat{\mathbf{v}}^g$  for image, caption and problem description, respectively.

**ATT.** The baseline with an attention mechanism is implemented following the Bottom-Up and Top-Down image captioning model [1]. Specifically, we compute a weight vector  $a_k^i$  for each of the  $K$  image features  $\mathbf{v}_k^i$  as follows:

$$a_k^i = \mathbf{w}_{ai}^T \tanh(W_{ai}\mathbf{v}_k^i + W_{gi}\mathbf{v}_k^g), \quad (4)$$

where  $W_{ai} \in \mathbb{R}^{H \times V}$ ,  $W_{gi} \in \mathbb{R}^{H \times V}$  and  $\mathbf{w}_{ai} \in \mathbb{R}^H$  are learned parameters, and  $H$  is the dimension of the projection layer.

Similarly, the attention weights  $a_k^c$  for each of the  $K$  language features  $\mathbf{v}_k^c$  in the solution description and  $a_k^g$  for each of the  $K$  language features  $\mathbf{v}_k^g$  in the problem description

are computed as

$$a_k^c = \mathbf{w}_{ac}^T \tanh(W_{gc}\mathbf{v}_k^g + W_{cc}\mathbf{v}_k^c), \quad (5)$$

and

$$a_k^g = \mathbf{w}_{ag}^T \tanh(W_{ag}\mathbf{v}_k^i + W_{gg}\mathbf{v}_k^g + W_{cg}\mathbf{v}_k^c), \quad (6)$$

where  $W_{gc} \in \mathbb{R}^{H \times V}$ ,  $W_{cc} \in \mathbb{R}^{H \times V}$ ,  $W_{ag} \in \mathbb{R}^{H \times V}$ ,  $W_{gg} \in \mathbb{R}^{H \times V}$ ,  $W_{cg} \in \mathbb{R}^{H \times V}$ ,  $\mathbf{w}_{ac} \in \mathbb{R}^H$  and  $\mathbf{w}_{ag} \in \mathbb{R}^H$  are the learned parameters. For simplicity, in solution steps prediction task, we do not consider  $\mathbf{v}_k^i$  and  $\mathbf{v}_k^c$  to get  $a_k^g$ .

Finally, we normalize the attention weights  $a_k^i$ ,  $a_k^c$ ,  $a_k^g$  using a softmax function, and apply them to the corresponding features.

$$\hat{\mathbf{v}}^m = \sum_{k=1}^K \text{softmax}(\mathbf{a}_k^m) \mathbf{v}_k^m, \quad (7)$$

where  $m$  indicates the corresponding modality, *i.e.*, image ( $i$ ), caption ( $c$ ) or problem description ( $g$ ).

#### References

- [1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [2] Jiacheng Chen, Hexiang Hu, Hao Wu, Yuning Jiang, and Changhu Wang. Learning the best pooling strategy for visual semantic embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- [3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019.
- [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.

Category	Pets and Animals			Health
Subcategory	Pets and Vacations	Dogs	Fish	Childhood Health
Problem (How to)	a) Care for pet during vacation	b) Teach a stubborn dog to sit down	c) Add algaecide to pond	d) Help a shy child
Solution Graph				
Caption & Attention	1. Only <b>take mice</b> with you if necessary.	1. Get your <b>dog</b> to <b>focus</b> on a <b>treat</b> in your <b>hand</b> .	1. Get <b>quaternary ammonia algaecide</b> .	1. <b>Role play</b> social situations with her.
Image & Attention	<p>2. Ensure <b>mice-friendly accommodations</b>.</p> <p>3. Bring extra <b>food</b> and <b>water</b>.</p> <p>4. Put a <b>nametag</b> and <b>ID</b> on the <b>cage</b>.</p>	<p>2. Hold the <b>treat</b> above your dog's head.</p> <p>3. <b>Give</b> the <b>'sit'</b> command.</p> <p>4. <b>Reward</b> the dog as soon as she <b>sits</b>.</p>	<p>2. <b>Wear protective gloves</b> and long sleeves.</p> <p>3. Mix <b>equal part algaecide</b> and water in a <b>tank sprayer</b>.</p> <p>4. Spray the algae in your <b>pond</b> with the <b>algaecide</b>.</p>	<p>2. Be an <b>active listener</b>.</p> <p>3. Help her practice <b>social skill words</b>.</p> <p>4. Model <b>healthy social behavior</b> when you are around her.</p>

Figure 1. An overview of the VisualHow dataset. We provide a hierarchical structure that organizes our data into categories, sub-categories, problems, solution graphs, steps with image-caption pairs, and multimodal attention.

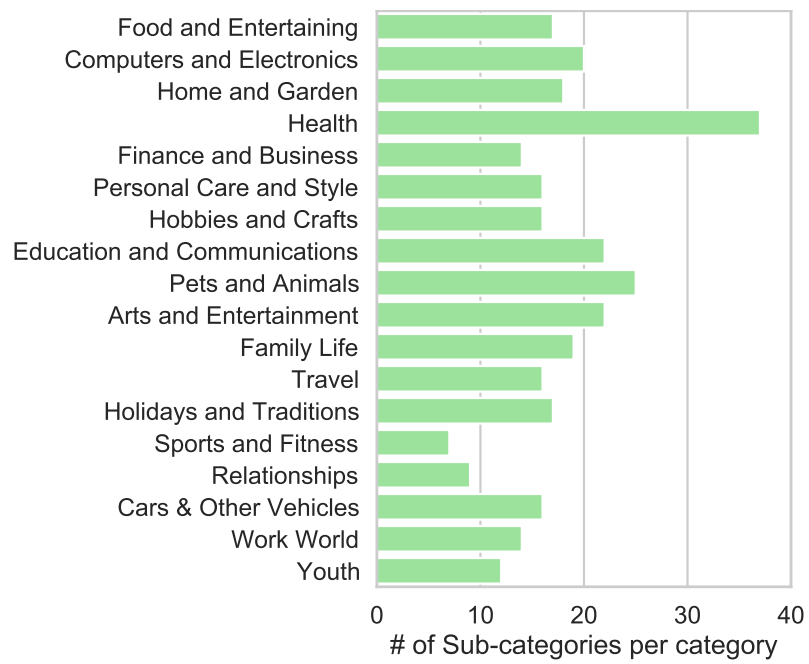


Figure 2. Numbers of subcategories in each category.





(a) wikiHow Article

CATEGORIES + PETS AND ANIMALS

## How to Involve a Pet in Christmas

Co-authored by wikiHow Staff and 14 contributors  
Last Updated: October 9, 2021

- 1 **Decorate your pet to make them appear more festive.** For example, you could...
- 2 **Give your pet a gift.** While receiving a special present, give your pet a gift, such as...

(b) Multimodal Attention

### Task: Involve a Pet in Christmas

Welcome to Amazon Mechanical Turk (MTurk). You can find our task instruction by referring to AMT Instruction Pages. If it is your first time to complete this task, we highly recommend you to watch this short instruction video at first.

Task Categories: Pets and Animals

Object and Bounding Box Annotation

Submit

Step 4: Play with your animal around the holidays.

Play with your animal around the holidays.

First Panel: Please select the words that are associated with any part of the image.

Add Selected Words

- animal Play with your animal around the holidays.

Add bounding boxes for animal

Second Panel: Please select the words that are important but not associated with any part of the image.

Add Selected Words

- Play Play with your animal around the holidays.
- holidays Play with your animal around the holidays.


Last Panel: Please annotate any important parts of the image that are not mentioned in the sentence.

Add New Regions

Step 4: Play with your animal around the holidays.

Annotate the bounding boxes for animal in the picture. Please make sure your bounding box only contains the desired objects and as little irrelevant region as possible following the given constraints.

Next Confirm Finishing Annotation



### Summary

#### Task: Involve a Pet in Christmas

- Decorate your pet to make them appear more festive.
  - pet Decorate your pet to make them appear more festive. ✓ 1
  - festive Decorate your pet to make them appear more festive. ✓ 1
  - Decorate Decorate your pet to make them appear more festive. ✓ 1
- Give your pet a gift.
  - gift Give your pet a gift. ✓ 1
  - pet Give your pet a gift. ✓ 1
- Give a treat to your pet.
  - treat Give a treat to your pet. ✓ 1
  - pet Give a treat to your pet. ✓ 1
- Play with your animal around the holidays.
  - animal Play with your animal around the holidays. ✓ 1
  - Play Play with your animal around the holidays. ✓ 1
  - holidays Play with your animal around the holidays. ✓ 1
  - You annotate the important parts of the image. ✓ 1

Edit Confirm

(c) Solution Graph

### Task: Involve a Pet in Christmas

Welcome to Amazon Mechanical Turk (MTurk). You can find our task instruction by referring to AMT Instruction Pages. If it is your first time to complete this task, we highly recommend you to watch this short instruction video at first.

Task Categories: Pets and Animals

Relations of Dependence

Submit


- Step 1: Decorate your pet to make them appear more festive.
- Step 2: Give your pet a gift.
- Step 3: Give a treat to your pet.
- Step 4: Play with your animal around the holidays.

Please select the most fitting step dependency relation to accomplish this task and click the confirm button.

All the steps to finish this task must be done sequentially.

All the steps to finish this task can be done in any order, or in an order other than sequential.

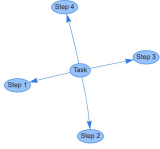
Confirm



Step 2: Give your pet a gift.

Give your pet a gift.

- Step 1: Decorate your pet to make them appear more festive.
- Step 2: Give your pet a gift.
- Step 3: Give a treat to your pet.
- Step 4: Play with your animal around the holidays.

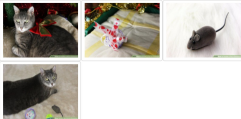


Edit Confirm

(d) Image Type

### Task: Involve a Pet in Christmas

Congratulations! You have almost finish this task. Select the image category of this task and press the Submit button to finish it.



- All real-life images
- All cartoon images
- Real-life and cartoon mixed images

(e) Feedback

If you have any comments or feedbacks, please fill the following text area. Thanks for your help.

Thank you!

Submit

Figure 4. Crowdsourcing interface of the VisualHow task, which contains a) an overview of the wikiHow article, b) annotation of the multimodal attention, and c) annotation of the solution graph, d) annotation of image type, and e) feedback.