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

Abstract

*Recent progress in the interdisciplinary studies of computer vision (CV) and natural language processing (NLP) has enabled the development of intelligent systems that can describe what they see and answer questions accordingly. However, despite showing usefulness in performing these vision-language tasks, existing methods still struggle in understanding real-life problems (*i.e.*, how to do something) and suggesting step-by-step guidance to solve them. With an overarching goal of developing intelligent systems to assist humans in various daily activities, we propose VisualHow, a free-form and open-ended research that focuses on understanding a real-life problem and deriving its solution by incorporating key components across multiple modalities. We develop a new dataset with 20,028 real-life problems and 102,933 steps that constitute their solutions, where each step consists of both a visual illustration and a textual description that guide the problem solving. To establish better understanding of problems and solutions, we also provide annotations of multimodal attention that localizes important components across modalities and solution graphs that encapsulate different steps in structured representations. These data and annotations enable a family of new vision-language tasks that solve real-life problems. Through extensive experiments with representative models, we demonstrate their effectiveness on training and testing models for the new tasks, and there is significant scope for improvement by learning effective attention mechanisms. Our dataset and models are available at* https://github.com/formidify/VisualHow.

1. Introduction

The remarkable progress in vision-language studies has developed visual systems with the ability to understand and generate natural language information. Existing vision-language models mainly focus on the understanding of visual input in task-free (*i.e.*, Image Captioning [3, 12, 17] and Visual Storytelling [28]) or question-directed (*i.e.*, Visual Question Answering [2, 24] and Visual Dialog [18]) settings. In other words, their aim is to develop visual systems that can "look and tell", by describing or answering questions about what is observed in a scene. On large-scale vision-language datasets [1, 2, 12, 14, 18, 24, 28, 37, 39, 62], state-of-the-art models have obtained promising achievements in understanding and predicting visual and textual information. Although achieving significant progress, these methods only perform well on standardized vision-language inference benchmarks and do not generalize to real-life situations to solve problems, which makes their scope of application relatively limited.

We believe that the next generation of visual intelligence systems will need to develop the ability to help humans

Figure 1. VisualHow is a vision-language task aiming to infer the solution to a real-life problem. The solution consists of multiple steps each described with an image and a caption.

*Equal contributions.

solve real-life problems more directly. Achieving the goal requires them to provide step-by-step solutions with both textual descriptions and visual illustration. Applications of such systems may include: 1. Teaching people everyday and/or domain-specific skills, such as to tie a tie, to make a sandwich, or to change a bicycle tire. 2. Helping people decompose an abstract goal into actionable items, such as to improve social skills, to sleep better, or to become a soccer player. To this end, we introduce a novel research problem – VisualHow – along with a large-scale dataset and a systematic evaluation of various modeling approaches. The main objective of VisualHow is to generate a step-by-step vision-language description of how to solve a problem, where a step will be described using an image and a caption. An example of VisualHow data is shown in Fig. 1. To "involve a pet in Christmas", one may need to take a series of different actions. While people may still find it difficult to understand how to involve a pet in Christmas by only reading the textual descriptions, looking at the visual illustrations will offer great help in the process. Therefore, given the description of the problem and the previous steps, the specific goal of VisualHow is to predict a pair of well-matched and complementary image and caption to describe what to do next. Achieving the goal requires the ability to understand three types of relationships: the relationship between the problem and the solution, the relationships between different steps of the solution, and the relationships between the visual and textual information.

Our goal is to enable the development of intelligent systems for tackling various real-life problems. Compared to conventional vision-language tasks, our proposed VisualHow task has the following differentiating factors: 1. **Real-life problems and multimodal solutions.** Rather than focusing on specific vision-language tasks [2, 13, 18, 24, 28, 37], our dataset contains 18 categories and 317 subcategories of real-life problems. Solutions to these problems are described in multiple steps, each with an image-caption pair, enabling the understanding of the decision-making process in problem solving. 2. **Fine-grained annotations.** Our VisualHow dataset offers two types of annotations that are absent from existing studies: the solution graphs describing dependencies between different steps, and multimodal attention that highlights and associates important keywords and regions of interest. They play an essential role in developing a structured understanding of the problem-solving procedure and closing the semantic gaps between vision and language. 3. **New vision-language tasks.** Our dataset enables several new vision-language tasks for various aspects of problem solving. Our experiments lead to several interesting observations and suggestions on improving the model performance.

To summarize, the contributions of this work are:

1. A new VisualHow study aiming to provide the foundation for developing novel vision-language methods and pushing the boundaries of multimodal understanding of real-life problems and solutions;

2. A new dataset that consists of diverse categories of problems, multimodal descriptions of solutions, and fine-grained annotations;

3. Experiments on multiple new tasks on different aspects of the VisualHow problem and extensive analyses of various baseline models.

2. Related Work

This paper is related to a series of studies including visual captioning and storytelling, visual question answering and dialog, multimodal instructions and multimodal representation learning.

2.1. Visual Captioning and Storytelling

There is a large body of research centering around generating textual descriptions of visual inputs. For example, the image captioning task [13, 26, 37, 44, 61] focuses on describing a single image with natural language, while visual storytelling [28] aims to generate a narrative with a sequence of sentences about multiple images. The shared goal of these studies is to develop methods to effectively encode the input images into representative features and transform them into a sequence of words that naturally and fluently describes the images. Therefore, in their standard configuration, image captioning and visual storytelling are image-to-sequence prediction tasks whose inputs are pixels and outputs are a sequence of words decoded according to a given vocabulary. While they focus on passively describing visual inputs without being directed by a specific purpose, the VisualHow task is different: First, it jointly predicts the images and captions that complement each other for the description of a solution, and second, the prediction is conditioned on the problem to solve. These differences make VisualHow a distinct and challenging research problem.

2.2. Visual Question Answering and Dialog

Previous studies have attempted to solve simple problems. For example, visual question answering [2, 24] and visual dialog [18, 40, 45] aim to answer questions about visual information based on the understanding of multimodal inputs. A number of recent studies have proposed large-scale datasets [2, 18, 24, 29] and neural network models [16, 23, 30, 51, 55] for free-form and open-ended VQA and visual dialog. However, these studies typically have restricted categories of questions, and their answers are in simplified forms (*i.e.*, categories or short phrases) [18]. On the contrary, the goal of our VisualHow task is to provide step-by-step description of the solution for various types of real-life problems. It not only requires the ability to understand both visual and textual information, but also involves

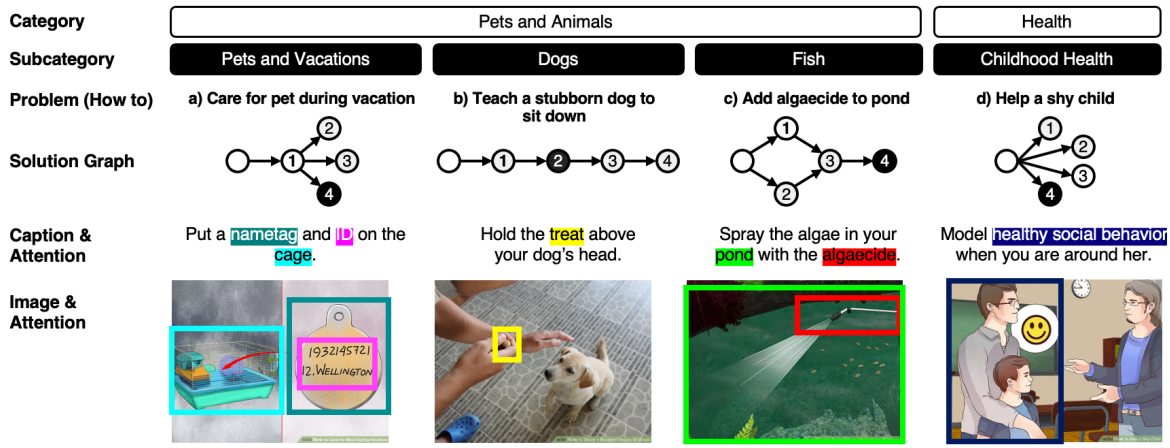

Figure 2. An overview of the VisualHow dataset. We provide a hierarchical structure that organizes our data into categories, sub-categories, problems, solution graphs, steps with image-caption pairs, and multimodal attention. Example steps are highlighted in the solution graph. Steps without a dependency are connected to an empty node.

constructing a reasonable structure of solutions to represent the relationship between different steps.

2.3. Multimodal Instructions

Our work is also related to existing studies on multimodal instructions. Datasets of instructional images [7, 58] and videos [49, 50, 63, 65, 66] provide step-by-step instructions about specific tasks. These datasets either focus on specific tasks or do not consider complex attention or structure in solutions. However, understanding the textual description of problems and providing step-by-step solutions each with a pair of well-matched caption and image have not been considered. Our work is different by considering diverse contents, multimodal attention, and solution structures, where the captions and images jointly describe the solution rather than each other. It contributes a large dataset with diverse and challenging problems, multimodal attention annotations, and non-sequential solutions.

2.4. Multimodal Representation Learning

There has been a long line of studies aiming to learn vision-language representations [10, 20, 35, 46, 53, 56]. They improve the representation learning using advanced attention mechanisms [59], better multimodal fusion methods [31, 48], multistep reasoning [11, 22], incorporation of object relations [35, 46, 64] and compositional reasoning models [27, 47]. Our study is most related with visual semantic embedding (VSE) [10, 20, 21, 32, 34, 56, 57], a typical category of approaches that learn a joint embedding space for visual and language representations. With VSE, compatibility score of visual features and language features can be computed as a simple inner-product. Specifically, DeViSE [21] learns to match the visual embeddings and semantic embeddings for zero-shot image recognition [9]. LSTM-SCNLM [32] encodes the sentence as the semantic embedding via bi-directional LSTMs. VSE++ [20] is a fundamental VSE method that uses average pooling as the feature aggregator with online hard-negative mining. VSRN [34] captures key objects and semantic concepts of a scene to generate visual representations. Global pooling operation (GPO) [10] learns to automatically adapt itself to the best pooling strategy for different features while staying effective and efficient. These studies have provided suitable baselines for the proposed VisualHow task, and inspired the development of computational models for problem-solving in real-life scenarios.

3. The VisualHow Dataset

The goal of this work is to introduce a new benchmark with a focus on real-life problems and high-quality annotations to the community of vision-language understanding. It consists of 18 categories of real-life problems and step-by-step solutions described with images and captions. The diversity and generality of problems and solutions also make VisualHow a more challenging dataset. In addition to the image-caption pairs, VisualHow provides annotations for solution graph and multimodal attention, which are essential for the understanding of problem-solution relationships and aligning the semantics between vision and language. An example of VisualHow data is shown in Fig. 2. In this section, we describe the data collection method, annotations, and the data statistics. Additional analyses and visualizations are provided in the supplementary materials.

3.1. Problems and Solutions

Building a general problem-solving dataset brings a series of unprecedented challenges. First, with the diversity and generality of real-life problems, manually defining and

categorizing problems is impractical. Next, since many of the problems require domain expertise (*e.g.*, those about health or finance), general online contents or non-expert workers can hardly generate high-quality solutions. To address these challenges, we collect real-life problems and solutions from the wikiHow [5,33,60] knowledge base, which is known for its high-quality instructional articles. The wikiHow articles are generated by a pool of well-qualified experts with the help of a rigorous quality screening process. All articles come with detailed step-by-step descriptions and very relevant visual illustrations in high resolution. Specifically, each problem consists of a language description (*e.g.*, a question starting with "How to") and is provided with a step-by-step solution that describes a method to solve it. The solution is composed of multiple steps described with an image and a caption. To control the data quality, VisualHow focuses on the proportion of wikiHow data with higher user ratings and popularity. A distinction from other wikiHow-based datasets is that for VisualHow we only select contents created by domain experts and with more than 50% of the users who voted and find it helpful, which ensures the quality of VisualHow contents. For problems with multiple solution methods, we consider each method a unique sample with the method title appended to the problem description.

3.2. Data Annotation

Learning to solve general problems is a challenging task, which requires knowledge to be learned from a variety of visual and textual information and organized in a structured representation. To tackle these challenges and benefit the development of future vision-language understanding methods, VisualHow offers fine-grained annotations on the solutions. As shown in Tab. 1, as distinguished from related studies, we collect these annotations with crowdsourcing and implement an effective quality control mechanism.

Crowdsourcing. The annotations are conducted in Amazon Mechanical Turk (AMT) with a custom annotation paradigm and a user interface (see Fig. 3). First, an overview of the problem and solution (*i.e.*, the wikiHow article) is presented to the workers. Next, they browse through all steps one at a time. In each step, they select the important phrases from the caption and annotate the corresponding image regions, which reflects their attention towards the multimodal information when performing different actions. Finally, they are asked to annotate the dependency between each pair of steps, which will formulate a directed solution graph to provide a structured representation of the problem-solving process. This research does not collect personal data from crowd workers and is exempt from IRB review.

Quality control. Our dataset requires an effective mechanism for quality control, so crowd workers can generate high quality annotations. Collecting high-quality multi-

	VisualHow	ViPT [7]	COIN [49,50]
Data Source	wikiHow	Snapguide; instructables	YouTube
Multimodal Attention	Yes	No	No
Solution Graph	Yes	No	No
Solution Types	Various	Procedure	Procedure

Table 1. Comparison between VisualHow and related datasets.

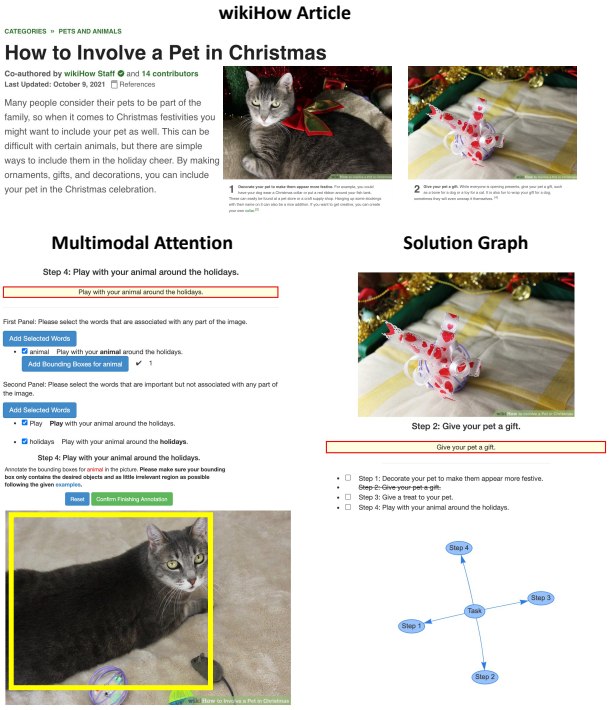

Figure 3. Crowdsourcing interface of the VisualHow task, which contains 1) an overview of the wikiHow Article, 2) annotation of the multimodal attention, and 3) annotation of the solution graph.

modal annotations is challenging. For example, determining what are important and need to be annotated can be subjective. To control the data quality, objectiveness, and consistency, we implement a series of quality control methods including qualification, correction, and expert review. We first compose a list of specific guidelines and require each worker to complete two qualification Human Intelligence Tasks (HITs), and examine the quality and consistency of their annotations to make sure that both the multimodal attention and the solution graphs are reasonably correct. At the end of each HIT, the workers are asked to review their annotations and correct what they find problematic. We review the HITs with automatic anomaly detection and manual examination, and problematic HITs will be sent back to the workers for correction. Through these steps, we ensure all workers follow the same quality standard.

3.3. Data Statistics

The VisualHow dataset consists of 20,028 real-life problems and their solutions that vary in number of steps and

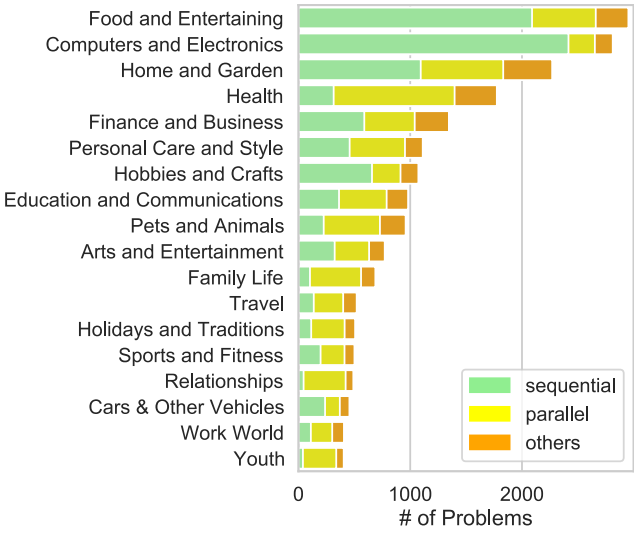

Figure 4. Number of problems in each category and the three types of solution graphs.

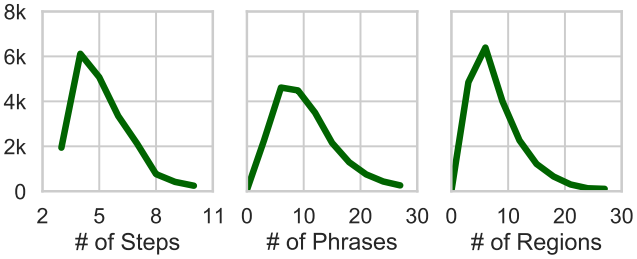

Figure 5. Distribution of solution steps and attention annotations.

fine-grained annotations. In this section, we conduct in-depth analyses and report key statistics of the dataset.

Problems. VisualHow contains 20,028 problems grouped in a hierarchy of 18 categories and 317 subcategories. Some of our major categories, such as Family Life, Computers and Electronics, Health, Finance and Business, have been rarely explored in previous vision-language studies. As shown in Fig. 4, the number of problems in each category ranges from 405 to 2,952, reflecting a naturally skewed distribution of wikiHow data. Despite that, VisualHow is still much more diverse than related datasets such as ViPT [7] and COIN [49], where a vast majority of samples are cooking or other household problems.

Solutions. As shown in Fig. 5, each solution consists of 3 to 10 steps described with images and captions. On average, each solution consists of 5.14 steps. The images and captions are more diverse than existing datasets, thanks to the wide variety of wikiHow data. Of all the images, 36.5% are realistic photos, and 58.6% are abstract images such as cartoons, drawings, handwritings, charts, *etc*. The rest 4.9% are mixed with both realistic and abstract contents. The captions also have a vocabulary of 30k tokens. Tab. 2 shows

Nouns		Verbs		Other POS	
water	icon	click	want	new	right
minutes	account	tap	open	around	first
time	hair	use	take	one	small
button	area	make	choose	sure	away
app	oil	add	remove	next	dry

Table 2. Most common tokens in the caption among 1) nouns; 2) verbs; 3) other parts of speech (POS).

Nouns		Verbs		Other POS	
water	oven	click	remove	online	overnight
doctor	bowl	tap	select	ok	comfortable
hair	child	open	choose	together	inside
settings	oil	add	check	regularly	daily
ingredients	food	use	make	outside	warm

Table 3. Most common tokens in the annotated phrases among 1) nouns; 2) verbs; 3) other parts of speech (POS).

the most common nouns (52.3%), verbs (30.0%), and other parts of speech (17.7%) in the captions.

Multimodal attention. We have collected abundant multimodal annotations about important image regions and phrases, which enables fine-grained learning of visual semantic alignment. In Fig. 5, on average, 9.13 image regions and 11.69 phrases are annotated for each solution. Over 98% of all steps have at least one instance of multimodal attention in both the image and caption, and around 99.5% of all steps have at least one annotated phrase. For each step of a solution, an average of 1.56 instances of multimodal attention are annotated in both the image and the caption. In addition, each step has an average of 0.72 important phrases annotated without specific image regions, and 0.13 image regions are annotated without their corresponding phrases. In Tab. 3, the tokens in the annotated phrases include nouns (60.8%), verbs (31.6%), and other parts of speech (7.6%). Compared to their distributions in the captions (see Tab. 2), the annotations contain more nouns, corresponding to various object instances in the images. The abundance of verbs and others allow to infer a variety of semantics in both modalities. On average, each annotated image region takes about 36.0% of the image size, while an image region without textual correspondence takes about 30.7%. For incomplete annotations, we assign an empty placeholder for their counterparts that allow them to be used in model training or completed in later versions of the dataset.

Solution graphs. The solution graphs are diverse, complicated, and important for characterizing the relationships between solution steps. They broadly fall into three types based on their structures, including sequential (Fig. 2b, all steps are performed in a sequential order), parallel (Fig. 2d, steps can be performed independently in any order), and others (Fig. 2a and Fig. 2c, some of the steps depend on another). As shown in Fig. 4, the distribution of the three types

Method	Image					Caption				
	MRR	R@1	R@3	R@5	Mean	MRR	R@1	R@3	R@5	Mean
GAP	0.495	28.583	62.313	77.882	3.758	0.535	34.449	64.785	79.284	3.558
GPO	0.501	29.441	62.249	78.676	3.695	0.549	35.695	67.165	81.203	3.392
ATT	**0.505**	**29.589**	**63.420**	**79.579**	**3.649**	**0.572**	**38.563**	**69.240**	**83.213**	**3.186**

Table 4. Quantitative results of Task 1: solution steps prediction.

of solution graphs varies across categories. For some categories (*e.g.*, Food and Entertaining, Computers and Electronics, Hobbies and Crafts), a majority of the graphs are sequential because they require to follow a certain procedure. For other categories (*e.g.*, Health, Pets and Animals, Relationships), the solutions often contain multiple steps that address different aspects of the problem (*i.e.* parallel) or have complex dependencies between steps.

These data and annotations enable fine-grained studies of understanding multimodal information in problem solving.

4. Experiments

Our VisualHow dataset enables new developments of intelligent problem-solving models that understand and generate solutions to real-life problems. In this section, we systematically analyze a series of baseline models that address new vision-language tasks based on the VisualHow dataset: 1) predicting the solution steps of paired images and captions, 2) predicting the dependencies of different solution steps, 3) describing the problem based on a given solution, and 4) generating captions of images in solution steps. These experiments demonstrate the success of benchmarking baseline models on the proposed VisualHow dataset. They also provide interesting analyses and observations and shed light on new research areas in multimodal understanding and real-life problem solving.

4.1. Baseline Models

In our experiments, we adopt state-of-the-art pretrained models to extract features from the visual and language modalities. In particular, the visual features are extracted from a ResNeXT-101 (32×8d) [25] pretrained on Instagram (WSL) [38], while the language features are obtained with a pretrained BERT model [19]. We explore three baseline methods to transform these features for downstream tasks: 1) GAP – a global average pooling method that independently processes features from different regions and words without considering their importance, 2) GPO – a generalized pooling operator [10] that aligns visual and language features and jointly considers them during feature aggregation, and 3) ATT – an attention mechanism to highlight the important semantic region of each modality and then aggregate them by the learned weights. Implementation details of these methods are introduced in the supplementary materials. Based on these methods, we develop baseline models for each of our four experiment tasks.

4.2. Task 1: Solution Steps Prediction

The main research objective of our work is to enable the learning of intelligent models that can predict step-by-step solutions to real-life problems with both visual illustrations and language descriptions. The joint prediction of multimodal descriptions has not been fully explored by existing vision-language studies. We achieve the goal by carrying out demonstrative experiments on the proposed VisualHow dataset with baseline models that simultaneously generate the multimodal solutions.

Implementation. Specifically, given the problem description and the previous solution steps, the models are asked to predict the image and caption of the next solution step by sorting two sets of candidate images and captions. We encode the problem, images, and captions using three encoders. The encoded features are dynamically integrated with a bidirectional GRU [15, 54]. To predict the next step of the solution, we develop a triplet network [10,20] to maximize the cosine similarity between the features of a positive candidate and the GRU features integrated from all previous steps, and to minimize that of a negative candidate.

Evaluation. At evaluation time, the candidates are sampled from the validation set following [18], which includes three sets of correct or incorrect solution steps: 1) the correct next step of the ground-truth solution, 2) 'hard negative' steps from solutions to the 10 most similar problems, 3) random solution steps from the same problem category. To capture this, all questions are embedded into a vector space by concatenating the averaged GloVe [42] embeddings of all words in the problem description. To generate 20 candidates, we first find the union of the correct and hard negative steps, and include other random steps until a unique set of 20 is found. The model is evaluated with three metrics: 1) mean reciprocal rank (MRR) of the correct step, 2) Recall@K, *i.e.*, existence of the correct step in top-K ranked steps, and 3) mean rank of the correct step.

Results. Tab. 4 shows the evaluation results of this task. First, we observe that conventional vision-language methods such as GPO [10] achieve mediocre performance, although better than the GAP baseline, suggesting that solving a real-life problem is more challenging than existing vision-language tasks. Further, the results show that atten-

Method	Image			Caption		
	MRR	SIM	KLD	MRR	SIM	KLD
ATT	0.505	0.293	1.937	0.572	0.371	1.586
ATT+CE	**0.507**	**0.520**	**0.852**	**0.580**	**0.665**	**0.543**

Table 5. Quantitative results of Task 1: solution steps prediction (with attention supervision).

tion mechanisms (ATT) can effectively improve the model performance even without explicit supervision, suggesting the importance of focused attention for understanding and solving real-life problems. Finally, it is noteworthy that the performance rankings across all evaluation metrics are consistent, which suggests that our dataset offers a fair benchmark for evaluating solution step prediction models.

Analyses of attention. The rich multimodal attention annotations of VisualHow dataset may act as a guidance for semantic alignment between the two modalities, which allows us to learn more accurate attention with explicit supervision and improve the prediction of solutions. To demonstrate this, we introduce auxiliary cross entropy (CE) losses to supervise the visual attention and language attention of models, and analyze the improvement of attention accuracy as well as task performance. We measure the attention accuracy with two popular evaluation metrics, Similarity (SIM) and KL-Divergence (KLD) [6].

Tab. 5 shows the quantitative results of models learned with (*i.e.*, ATT+CE) or without (*i.e.*, ATT) attention supervision. Consistent with past observations [11], we find that explicit attention supervision during training may help models focus on important visual and language features, resulting in improved SIM and KLD scores. It also improves their image and caption prediction performance (*i.e.*, MRR). Fig. 6 further compares the attention output of the two models learned with and without explicit supervision. They show that explicit attention supervision not only helps the model locate important regions and words in the multimodal solutions, but also plays an essential role in correlating key components across the two modalities (*e.g.*, fish oil, steak) and deriving more accurate solutions. These observations highlight the important role of multimodal attention for deriving comprehensive solutions to real-life problems.

4.3. Task 2: Solution Graph Prediction

Next, given the problem and solution descriptions, we develop models to predict the solution graph. This experiment aims to demonstrate the solution graph as a fine-grained annotation for developing a better understanding about the order and dependency of different solution steps.

Implementation. To capture the relationships between different steps, we concatenate the features extracted from images, captions and the problem description, and learn a single linear layer with a sigmoid activation function to predict the dependency matrix that indicates the dependencies between every two steps.

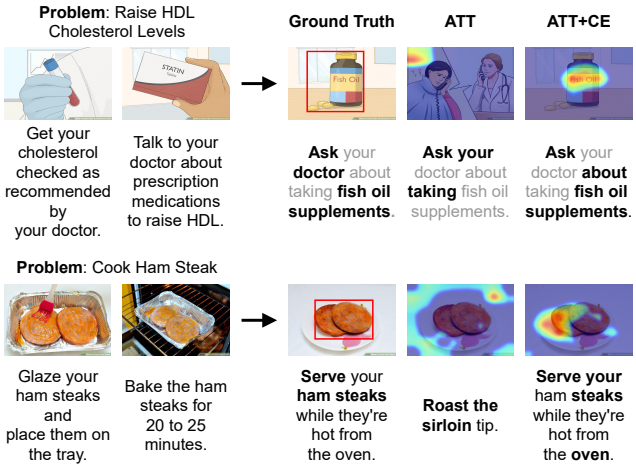

Figure 6. Qualitative results for attention supervision. Important regions and keywords are highlighted with red and black colors.

Method	IoU@0.25	IoU@0.5	IoU@0.75
GAP	0.484	0.377	0.268
GPO	0.468	0.380	0.302
ATT	0.473	0.389	0.319
ATT+CE	**0.494**	**0.434**	**0.376**

Table 6. Quantitative results of Task 2: solution graph prediction.

Evaluation. Evaluation of solution graph prediction is an open problem. In this work, we calculate the intersection over union (IoU) [8, 43] given specific thresholds to compare the similarity between the predicted probability matrix and the ground-truth solution graph. Specifically, we apply a threshold (*e.g.*, 0.25, 0.5, 0.75) to the model output to determine the graph edges and count the edges for the intersection and union between the graph and the ground truth to compute the IoU score.

Results. As shown in Tab. 6, understanding and predicting the dependencies between solution steps is a challenging task for the baseline models, while the ranks of different models remain similar to Task 1. Similarly, the IoU performance can be improved with the attention mechanism and explicit supervision. These results demonstrate the potential of learning fine-grained solution structures based on the understanding of vision and language descriptions.

4.4. Task 3: Problem Description Generation

To further demonstrate the usage of our VisualHow as a general vision-language benchmark, we present a demonstrative experiment for the generation of problem description based on the visual and textual descriptions of a solution. This experiment resembles those for the conventional vision-language tasks (*e.g.*, image captioning and vi-

Method	B-1	B-2	B-3	B-4	M.	R.	C.
ATT (I)	16.7	8.5	4.7	2.9	6.7	16.5	22.9
ATT (C)	22.1	11.4	6.3	3.9	9.8	22.1	44.5
ATT (I+C)	22.7	12.0	6.8	4.4	9.9	22.4	46.7
ATT+CE (I)	16.9	9.5	5.3	3.7	7.3	18.5	24.6
ATT+CE (C)	22.8	11.7	6.3	3.8	9.9	22.3	47.0
ATT+CE (I+C)	**24.1**	**13.1**	**7.7**	**4.8**	**10.7**	**23.2**	**50.8**

Table 7. Quantitative results of Task 3: problem description generation.

sual question answering) and focus on estimating the models' capability of understanding multimodal contents and performing language generation.

Implementation. For this task, the visual and language features are directly concatenated across all steps, and a BUTD captioning model [3] is adapted to generate the problem description. We implement the attention-based methods (*i.e.*, ATT and ATT+CE) using different inputs: images only (I), captions only (C), and both (I+C).

Evaluation. To evaluate problem description models, we adopt four automatic metrics that are widely used for captioning evaluation, including BLEU [41], METEOR [4], ROUGE-L [36], and CIDEr [52].

Results. Tab. 7 presents the results of problem description generation. We observe that leveraging both images and captions (I+C) leads to a clear improvement over single-modality models (*i.e.*, image-only (I) and caption-only (C)). Furthermore, the results show that attention supervision (ATT+CE) has a positive impact on the task performance. Notably, the improvement is bigger with both modalities compared with single modality, suggesting the usefulness of the attention data and supervision methods that highlight multimodal attention alignment.

4.5. Task 4: Solution Captions Generation

Our proposed VisualHow dataset can also serve as a useful testbed for evaluating the models' capability of jointly considering multiple images and generating fluent stories. For the final task in our experiments, we consider generating the solution captions based on the input problem description and solution images. It can be considered as a visual storytelling task, but with additional emphasis on the contextual relationship between the goals of problems and the different steps for achieving them.

Implementation. We adapt the AREL [54] model that achieves the state-of-the-art performance on the ViST [28] dataset. We feed the solution images and a BERT embedding of the problem description to the model, to obtain a sequence of captions corresponding to the images [28, 54].

Evaluation. The training and evaluation of models follow the standard visual storytelling paradigm. BLEU [41], METEOR [4], ROUGE-L [36], and CIDEr [52] are used as evaluation metrics to compare the generated captions with the ground truth.

Method	B-1	B-2	B-3	B-4	M.	R.	C.
GAP	28.2	13.0	7.3	4.5	23.2	24.1	12.7
GPO	33.0	15.7	7.4	5.6	27.0	26.4	23.0
ATT	33.6	16.4	7.4	5.8	27.2	27.1	23.4
ATT+CE	**33.8**	**17.0**	**9.9**	**6.2**	**28.1**	**28.2**	**24.3**

Table 8. Quantitative results of Task 4: solution captions generation.

Results. Quantitative results of this task are demonstrated in Tab. 8. From the results, we observe that the generation of captions is less challenging than the prediction of problem descriptions, as suggested by the higher BLEU [41], METEOR [4] and ROUGE-L [36] scores. However, the CIDEr [52] scores are significantly lower than those of Task 3. It is because the length of solution captions is much longer than that of the problem description and the models are prone to predict the common words that are discounted by CIDEr. Comparing the different models, we observe that ATT+CE obtains the best performance, while the ATT and GPO fall slightly behind, and the GAP achieves the lowest performance. This suggests that learning to focus on important features can help with the understanding of solution images and generating their corresponding captions. These observations suggest that VisualHow is a challenging benchmark for visual storytelling models, and accurate attention is important for generating fluent descriptions.

5. Conclusion

The ability to solve real-world problems is an important step toward human-like intelligence. In this paper, we have introduced VisualHow, a large-scale dataset for solving real-life problems. Utilizing expert-generated internet contents and crowdsourcing, we collected and annotated 20,028 problems and solutions. Dataset statistics demonstrate that the problems, solutions, and annotations contain rich multimodal solutions for a variety of problems in real-life scenarios. Understanding and predicting solutions to real-life problems is an inherently challenging problem. These data and annotations enable a family of new vision-language tasks and computational methods for understanding and solving problems. Our results indicate that there is significant scope for improvement. We hope that this work will facilitate future research to better understand the multimodal information in real-life problem-solving. We envision that this work will spur innovation and encourage developments in problem-solving systems that can positively impact a wide range of applications.

Acknowledgements

This work is supported by NSF Grant 1908711.

References

[1] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[2] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

[3] 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.

[4] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Annual Conference of the Association for Computational Linguistics Workshop (ACLW)*, 2005.

[5] Irshad Bhat, Talita Anthonio, and Michael Roth. Towards modeling revision requirements in wikiHow instructions. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

[6] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell us about saliency models? *IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI)*, 2018.

[7] Khyathi Raghavi Chandu, Ruo-Ping Dong, and Alan Black. Reading between the lines: Exploring infilling in visual narratives. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

[8] Chien-Yi Chang, De-An Huang, Danfei Xu, Ehsan Adeli, Li Fei-Fei, and Juan Carlos Niebles. Procedure planning in instructional videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[9] Soravit Changpinyo, Wei-Lun Chao, Boqing Gong, and Fei Sha. Synthesized classifiers for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[10] 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.

[11] Shi Chen, Ming Jiang, Jinhui Yang, and Qi Zhao. AiR: Attention with reasoning capability. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[12] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

[13] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325v2*, 2015.

[14] Zhenfang Chen, Peng Wang, Lin Ma, Kwan-Yee K. Wong, and Qi Wu. Cops-Ref: A new dataset and task on compositional referring expression comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[15] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

[16] Michael Cogswell, Jiasen Lu, Rishabh Jain, Stefan Lee, Devi Parikh, and Dhruv Batra. Dialog without dialog data: Learning visual dialog agents from VQA data. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[17] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. M^2: Meshed-memory transformer for image captioning. 2020.

[18] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[19] 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.

[20] Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. VSE++: Improving visual-semantic embeddings with hard negatives. In *British Machine Vision Conference (BMVC)*, 2018.

[21] Andrea Frome, Greg S. Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. DeViSE: A deep visual-semantic embedding model. 2013.

[22] Zhe Gan, Yu Cheng, Ahmed El Kholy, Linjie Li1, Jingjing Liu, and Jianfeng Gao. Multi-step reasoning via recurrent dual attention for visual dialog. In *Annual Conference of the Association for Computational Linguistics (ACL)*, 2019.

[23] Difei Gao, Ke Li, Ruiping Wang, Shiguang Shan, and Xilin Chen. Multi-modal graph neural network for joint reasoning on vision and scene text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[24] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[25] 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.

[26] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and

evaluation metrics. *Journal of Artificial Intelligence Research (JAIR)*, 2013.

[27] Richang Hong, Daqing Liu, Xiaoyu Mo, Xiangnan He, and Hanwang Zhang. Learning to compose and reason with language tree structures for visual grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI)*, 2019.

[28] Ting-Hao (Kenneth) Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. Visual storytelling. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT)*, 2016.

[29] Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[30] Huaizu Jiang, Ishan Misra, Marcus Rohrbach, Erik Learned-Miller, and Xinlei Chen. In defense of grid features for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[31] Aisha Urooj Khan, Amir Mazaheri, Niels da Vitoria Lobo, and Mubarak Shah. MMFT-BERT: Multimodal fusion transformer with bert encodings for visual question answering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

[32] Ryan Kiros and Richard S. Zemel Ruslan Salakhutdinov. Unifying visual-semantic embeddings with multimodal neural language models. *Conference on Neural Information Processing Systems Workshops (NeurIPSW)*, 2014.

[33] Mahnaz Koupaee and William Yang Wang. WikiHow: A large scale text summarization dataset. *arXiv preprint arXiv:1810.09305*, 2018.

[34] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[35] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[36] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Annual Conference of the Association for Computational Linguistics Workshop (ACLW)*, 2004.

[37] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014.

[38] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[39] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. HowTo100M: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[40] Yulei Niu, Hanwang Zhang, Manli Zhang, Jianhong Zhang, Zhiwu Lu, and Ji-Rong Wen. Recursive visual attention in visual dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[41] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Annual Conference of the Association for Computational Linguistics (ACL)*, 2002.

[42] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

[43] Darwin Saire Pilco and Adín Ramírez Rivera. Graph learning network: A structure learning algorithm. In *Proceedings of the International Conference on Machine Learning Workshop (ICMLW)*, 2019.

[44] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[45] Jiaxin Qi, Yulei Niu, Jianqiang Huang, and Hanwang Zhang. Two causal principles for improving visual dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[46] Alec Radford, Jong Wook Kim, Aditya Ramesh Chris Hallacy, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.

[47] Raeid Saqur and Karthik Narasimhan. Multimodal graph networks for compositional generalization in visual question answering. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

[48] Hao Tan and Mohit Bansal. LXMERT: Learning cross-modality encoder representations from transformers. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

[49] Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. COIN: A large-scale dataset for comprehensive instructional video analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[50] Yansong Tang, Jiwen Lu, and Jie Zhou. Comprehensive instructional video analysis: The COIN dataset and performance evaluation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI)*, 2020.

[51] Tao Tu, Qing Ping, Govindarajan Thattai, Gokhan Tur, and Prem Natarajan. Learning better visual dialog agents with

pretrained visual-linguistic representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[52] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[53] Haoran Wang, Ying Zhang, Zhong Ji, Yanwei Pang, and Lin Ma. Consensus-aware visual-semantic embedding for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[54] Xin Wang, Wenhu Chen, Yuan-Fang Wang, and William Yang Wang. No metrics are perfect: Adversarial reward learning for visual storytelling. In *Annual Conference of the Association for Computational Linguistics (ACL)*, 2018.

[55] Yue Wang, Shafiq Joty, Michael R. Lyu, Irwin King, Caiming Xiong, and Steven C.H. Hoi. VD-BERT: A unified vision and dialog transformer with bert. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

[56] Hao Wu, Jiayuan Mao, Yufeng Zhang, Yuning Jiang, Lei Li, Weiwei Sun, and Wei-Ying Ma. Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[57] Hao Wu, Jiayuan Mao, Yufeng Zhang, Yuning Jiang, Lei Li, Weiwei Sun, and Wei-Ying Ma. Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[58] Semih Yagcioglu, Aykut Erdem, Erkut Erdem, and Nazli Ikizler-Cinbis. RecipeQA: A challenge dataset for multimodal comprehension of cooking recipes. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

[59] Xu Yang, Hanwang Zhang, Guojun Qi, and Jianfei Cai. Causal attention for vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[60] Yue Yang, Artemis Panagopoulou, Qing Lyu, Li Zhang, Mark Yatskar, and Chris Callison-Burch. Visual goal-step inference using wikiHow. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

[61] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics (TACL)*, 2014.

[62] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling context in referring expressions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[63] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. MERLOT: Multimodal neural script knowledge models. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

[64] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. VinVL: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[65] Luowei Zhou, Chenliang Xu, and Jason J. Corso. Towards automatic learning of procedures from web instructional videos. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2018.

[66] Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, and Josef Sivic. Cross-task weakly supervised learning from instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.