A Simple Baseline for Visual Commonsense Reasoning

Jingxiang Lin, Unnat Jain, Alexander G. Schwing
University of Illinois at Urbana-Champaign

Abstract

Reasoning is an important ability that we learn from a very early age. To develop models with better reasoning abilities, recently, the new visual commonsense reasoning (VCR) task has been introduced. Not only do models have to answer questions, but also do they have to provide a reason for the given answer. Baselines achieve compelling results via a meticulously designed model composed of LSTM modules and attention nets. Here we show that a much simpler model can perform better with half the number of trainable parameters. By associating visual features with attribute information and better text to image grounding, we obtain further improvements for our simpler & effective baseline, **TAB-VCR**. Our approach results in a 5.3%, 4.4% and 6.5% absolute improvement over previous state-of-the-art [37] on question answering, answer justification and holistic VCR. The extended version [26] of this workshop paper and the code is available at `https://deanplayerljx.github.io/tabvcr`.

1 Introduction

Recently, respectable results have been achieved for vision & language tasks. For instance, for visual question answering [4, 10] and visual dialog [8], compelling results have been reported. Many models achieve results well beyond random guessing on challenging datasets [9, 23, 38, 15]. However, it is also known that algorithm results aren't stable at all and trained models often leverage biases to answer questions. Thus, it is important to shed light into the decision process of models, and reasoning is one aspect of it.

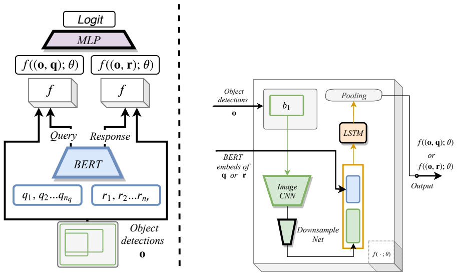
Many efforts have been made to address the multi-modal reasoning problem by either incorporating text-image grounding [34, 35] or increasing the interpretability of models [3, 6, 12, 21, 30, 14, 17, 33, 11, 13, 16, 20]. Recently, a new "visual commonsense reasoning" [37] challenge was posed. In addition to visual question answering, the algorithm has to justify the answer. In this new dataset the questions, answers and rationale are expressed using natural language containing references to objects. The baseline, which achieves compelling results, leverages those cues by combining a LSTM module based deep net with attention over objects to obtain grounding and context.

However, the baseline is also very intricate. We revisit this baseline and show that a much simpler model with less than half the trainable parameters achieves much better results. As shown in Fig. 2, we incorporate visual attribute information in VCR detections and augment object-word grounding provided in the VCR dataset. We refer to our developed tagging and attribute baseline as **TAB-VCR**.

2 Attribute-based Visual Commonsense Reasoning (VCR)

Given an input image, the VCR task is divided into two subtasks: (1) **question answering** ($Q{\to}A$): given a question (Q), select the correct answer (A) from four candidate answers; (2) **answer justification** ($QA{\to}R$): given a question (Q) and correct answer (A), select the correct rationale (R) from four candidate rationales. Importantly, both subtasks can be unified: choosing a *response* from four options given a *query*. For $Q{\to}A$, the query is a question and the options are candidate answers. For $QA{\to}R$, the query is a question appended by its correct answer and the options are candidate rationales. Note, the $Q{\to}AR$ combines both, *i.e.*, a model needs to succeed at $Q{\to}A$ and $QA{\to}R$.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.


Logit
MLP
$f((\mathbf{o},\mathbf{q});\theta)$
$f((\mathbf{o},\mathbf{r});\theta)$
f
f
Query
Response
BERT
$q_1, q_2 ... q_{n_q}$
$r_1, r_2 ... r_{n_r}$
Object detections $\mathbf{o}$
Object detections $\mathbf{o}$
b_1
Pooling
LSTM
BERT embeds of $\mathbf{q}$ or $\mathbf{r}$
Image CNN
Downsample Net
$f((\mathbf{o},\mathbf{q});\theta)$ or $f((\mathbf{o},\mathbf{r});\theta)$
Output
$f(\cdot\,;\theta)$

(a) Overview (b) Joint image & language encoder

Figure 1: (a) **Overview of the proposed TAB-VCR model**: Inputs are the image (with object bounding boxes), a query and a candidate response. Sentences (query & response) are represented using BERT embeddings and encoded jointly with the image using a deep net module $f(\cdot;\theta)$. The representations of query and response are concatenated and scored via a multi-layer perceptron (MLP); (b) **Details of joint image & language encoder** $f(\cdot;\theta)$: BERT embeddings of each word are concatenated with their corresponding local image representation.This information is pass through an LSTM and pooled to give the output $f((I,\mathbf{w});\theta)$. The network components outlined in black, *i.e.*, MLP, downsample net and LSTM are the only components with trainable parameters.

The proposed method focuses on choosing a response given a query, for which we introduce notation next.

We are given an *image*, a *query* and four candidate *responses*. The words in the query and responses are grounded to objects in the image. The query and response are collections of words, while the image data is a collection of object detections. One of the detections also corresponds to the entire image, symbolizing a global representation. The image data is denoted by the set $\mathbf{o} = (o_i)_{i=1}^{n_o}$, where each o_i, $i \in \{1, \ldots, n_o\}$, consists of a bounding box b_i and a class label $l_i \in \mathcal{L}$[1]. The query is composed of a sequence $\mathbf{q} = (q_i)_{i=1}^{n_q}$, where each q_i, $i \in \{1, \ldots, n_q\}$, is either a word in the vocabulary $\mathcal{V}$ or a tag referring to a bounding box in $\mathbf{o}$. A data point consists of four responses and we denote a response by the sequence $\mathbf{r} = (r_i)_{i=1}^{n_r}$, where r_i, $i \in \{1, \ldots, n_r\}$, (like the query) can either refer to a word in the vocabulary $\mathcal{V}$ or a tag.

We develop a conceptually simple joint encoder for language and image information, $f(\,\cdot\,;\theta)$, where θ is the catch-all for all the trainable parameters. Our proposed approach is outlined in Fig. 1(a), and the joint language and image encoder is illustrated in Fig. 1(b). Note that for non-*tag* words, *i.e.*, words without an associated object detection, the object detection for the entire image is utilized.

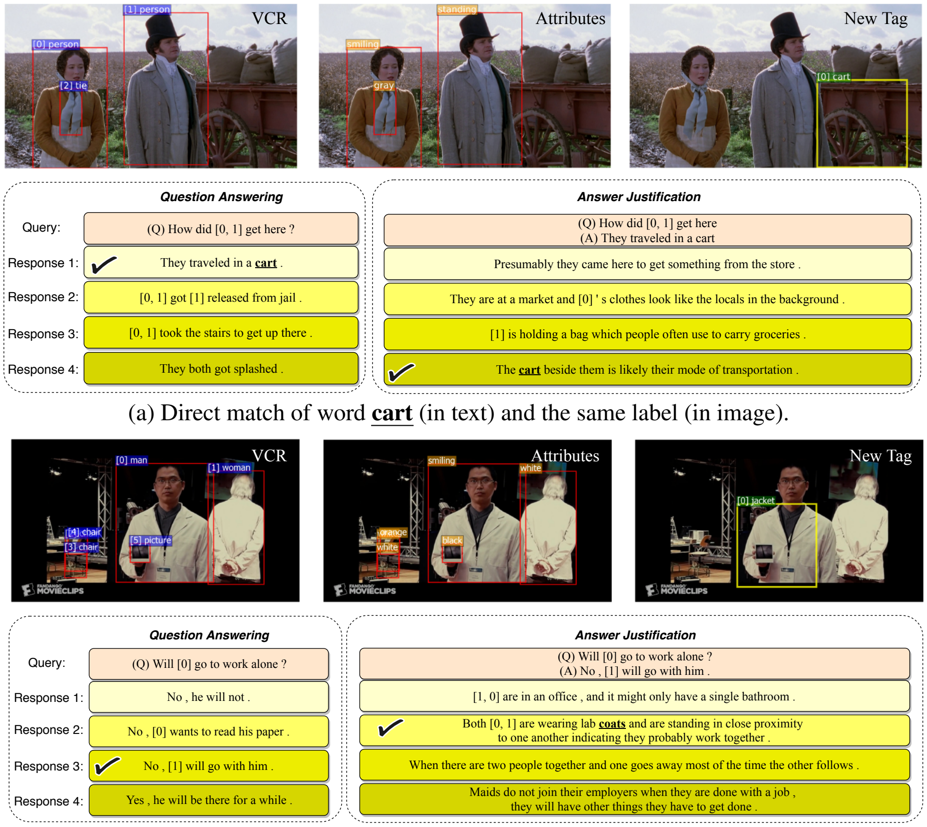
Attributes capturing visual features: We hypothesize that visual question answering and reasoning benefits from information about object characteristics and attributes. This intuition is illustrated in Fig. 2 where attributes add valuable information to help reason about the scene, such as '*gray* tie' and '*standing* man.'

To validate this hypothesis we deploy a pretrained attribute classifier which augments every detected bounding box b_i with a set of attributes such as colors, texture, size and emotions. We show the attributes predicted by our model's image CNN in Fig. 2. For this, we take advantage of work by Anderson et al. [2] as it incorporates attribute features to improve performance on language and vision tasks.

***New tags* for better text to image grounding:** Associating a word in the text with an object detection in the image, *i.e.*, $o_i = (b_i, l_i)$ is what we commonly refer to as text-image grounding. Any word serving as a pointer to a detection is referred to as a *tag* by Zellers et al. [37]. Importantly, we found many nouns in the text (query or responses) aren't grounded with their appearance in the image. To overcome this shortcoming, we develop Algorithm 1 to find new text-image groundings or *new tags*. A qualitative example is illustrated in Fig. 2. Nouns such as 'cart' and 'coat' wasn't tagged by VCR, while our TAB-VCR model can tag them.

Specifically, for text-image grounding we first find detections $\hat{\mathbf{o}}$ (in addition to VCR provided $\mathbf{o}$) using the image CNN. The set of unique class labels in $\hat{\mathbf{o}}$ is assigned to $\hat{\mathcal{L}}$. Both $\mathbf{q}$ and $\mathbf{r}$ are modified such that all *tags* (pointers to detections in the image) are remapped to natural language (class label of the detection). This is done via the `remap` function. We follow Zellers et al. [37] and associate a gender neutral name for the 'person' class. For instance, "How did [0,1] get here?" in Fig. 2 is remapped to "How did Adrian and Casey get here?". Next, the POS tagging function (`pos_tag`) parses a sentence $\mathbf{w}$ and assigns POS tags to each word w. For finding *new tags*, we are only interested in words with the POS tag being either singular noun (NN) or plural noun (NNS). For these noun words, we check if a word w directly matches a label in $\hat{\mathcal{L}}$. If such a direct match exists, we associate w to the

[1] The dataset also includes information about segmentation masks, which are neither used here nor by previous methods. Data available at: `visualcommonsense.com`



Question Answering

Query: (Q) How did [0, 1] get here ?

Response 1: ✓ They traveled in a **cart** .

Response 2: [0, 1] got [1] released from jail .

Response 3: [0, 1] took the stairs to get up there .

Response 4: They both got splashed .

Answer Justification

(Q) How did [0, 1] get here
(A) They traveled in a cart

Presumably they came here to get something from the store .

They are at a market and [0] ' s clothes look like the locals in the background .

[1] is holding a bag which people often use to carry groceries .

✓ The **cart** beside them is likely their mode of transportation .

(a) Direct match of word **cart** (in text) and the same label (in image).

Question Answering

Query: (Q) Will [0] go to work alone ?

Response 1: No , he will not .

Response 2: No , [0] wants to read his paper .

Response 3: ✓ No , [1] will go with him .

Response 4: Yes , he will be there for a while .

Answer Justification

(Q) Will [0] go to work alone ?
(A) No , [1] will go with him .

[1, 0] are in an office , and it might only have a single bathroom .

✓ Both [0, 1] are wearing lab **coats** and are standing in close proximity to one another indicating they probably work together .

When there are two people together and one goes away most of the time the other follows .

Maids do not join their employers when they are done with a job , they will have other things they have to get done .

(b) Word sense based match of word **coats** and label ‘jacket’ with the same meaning.

Figure 2: **Qualitative results:** Two types of *new tags* found by our method are (a) direct matches and (b) word sense based matches. Note that the images on the left show the object detections provided by VCR. The images in the middle show the attributes predicted by our model and thereby captured in visual features. The images on the right show *new tags* detected by our proposed method. Below the images are the question answering and answer justification subtasks.

detections of the matching label. As shown in Fig. 2(a), this direct matching associates the word **cart** in the text (response 1 of the $Q \rightarrow A$ subtask and response 4 of the $QA \rightarrow R$ subtask) to the detection corresponding to label ‘cart’ in the image, creating a *new tag*.

If there is no such direct match for w, we find matches based on word sense. This is motivated in Fig. 2(b) where the word ‘coat’ has no direct match to any image label in $\hat{\mathcal{L}}$. Rather there is a detection of ‘jacket’ in the image. Notably, the word ‘coat’ has multiple word senses, such as ‘an outer garment that has sleeves and covers the body from shoulder down’ and ‘growth of hair or wool or fur covering the body of an animal.’ Also, ‘jacket’ has multiple word senses, two of which are ‘a short coat’ and ‘the outer skin of a potato’. As can be seen, the first word senses of ‘coat’ and ‘jacket’ are similar and would help match ‘coat’ to ‘jacket.’ Having said that, the second word senses are different from common use and from each other. Hence, for words that do not directly match a label in $\hat{\mathcal{L}}$, choosing the appropriate word sense is necessary. To this end, we adopt a simple approach, where we use the most frequently used word sense of w and of labels in $\hat{\mathcal{L}}$. This is obtained using the first synset in Wordnet in NLTK [29, 27]. Then, using the first synset of w and labels in $\hat{\mathcal{L}}$, we find the best matching label ‘best_label’ corresponding to the highest Wu-Palmer similarity between synsets [36]. Additionally, we lemmatize w before obtaining its first synset. If the Wu-Palmer similarity between word w and the ‘best_label’ is greater than a threshold k, we associate the word to the detections of ‘best_label.’ Overall this procedure leads to *new tags* where text and label aren’t the same but have the same meaning. We found $k = 0.95$ was apt for our experiments. While inspecting, we found this algorithm missed to match the word ‘men’ in the text to the detection label ‘man.’ This is due to the ‘lemmatize’ function provided by NLTK [27]. Consequently, we additionally allow *new tags* corresponding to this ‘men-man’ match.

	$Q{\rightarrow}A$	$QA{\rightarrow}R$	$Q{\rightarrow}AR$	Params (Mn)	
	(val)	(val)	(val)	(total)	(trainable)
R2C (Zellers et al. [37])	63.8	67.2	43.1	35.3	26.8
Improving R2C					
R2C + Det-BN	64.49	67.02	43.61	35.3	26.8
R2C + Det-BN + Freeze (R2C++)	65.30	67.55	44.41	35.3	11.7
R2C++ + Resnet101	67.55	68.35	46.42	54.2	11.7
R2C++ + Resnet101 + Attributes	68.53	70.86	48.64	54.0	11.5
Ours					
Base	66.39	69.02	46.19	28.4	4.9
Base + Resnet101	67.50	69.75	47.51	47.4	4.9
Base + Resnet101 + Attributes	69.51	71.57	50.08	47.2	4.7
Base + Resnet101 + Attributes + New Tags (**TAB-VCR**)	**69.89**	**72.15**	**50.62**	47.2	4.7

Table 1: Comparison of our approach to the current state-of-the-art R2C [37] on the validation set. Legend: **Det-BN**: Deterministic testing using train time batch normalization statistics. **Freeze**: Freeze all parameters of the image CNN. **ResNet101**: ResNet101 backbone as image CNN (default is ResNet50). **Attributes**: Attribute capturing visual features by using [2] (which has a ResNet101 backbone) as image CNN. **Base**: Our base model, as detailed in Fig. 1(a) and Fig. 1(b)**New Tags**: Augmenting object detection set with *new tags* (as detailed in Sec. 2), *i.e.*, grounding additionnal nouns in the text to the image.

Algorithm 1 Finding *new tags*

```
Forward pass through image CNN to obtain object detections $\hat{\mathbf{o}}$
$\hat{\mathcal{L}} \leftarrow$ set(all class labels in $\hat{\mathbf{o}}$)
for $w \in \mathbf{w}$ where $\mathbf{w} \in \{\mathbf{q}, \mathbf{r}\}$ do
    if $w$ is tag then $w \leftarrow$ remap($w$)
new_tags ← {}
for $w \in \mathbf{w}$ where $\mathbf{w} \in \{\mathbf{q}, \mathbf{r}\}$ do
    if (pos_tag($w|\mathbf{w}$) $\in$ {NN, NNS}) and (wsd_synset($w, \mathbf{w}$) has a noun) then
        if $w \in \hat{\mathcal{L}}$ then                                  ▷ Direct match between word and detections
            new_detections ← detections in $\hat{\mathbf{o}}$ corresponding to $w$
            add ($w$, new_detections) to new_tags
        else                                                       ▷ Use word sense to match word and detections
            max_wup ← 0
            word_lemma ← lemma($w$)
            word_sense ← first_synset(word_lemma)
            for $\hat{l} \in \hat{\mathcal{L}}$ do
                if wup_similarity(first_synset($\hat{l}$), word_sense) > max_wup then
                    max_wup ← wup_similarity(first_synset($\hat{l}$), word_sense)
                    best_label ← $\hat{l}$
            if max_wup > $k$ then
                new_detections ← detections in $\hat{\mathbf{o}}$ corresponding to best_label
                add ($w$, new_detections) to new_tags
```

This algorithm permits to find *new tags* in 7.1% answers and 32.26% rationales. A split over correct and incorrect responses is illustrated in Fig. 3. If there is more than one detection associated with a *new tag*, we average the visual features at the step before the LSTM in the joint encoder.

3 Experiments[2]

Dataset: We train our models on the visual commonsense reasoning dataset [37] which contains over 212k (train set), 26k (val set) and 25k (test set) questions on over 110k unique movie scenes.

Metrics: Models are evaluated with classification accuracy on the $Q{\rightarrow}A$, $QA{\rightarrow}R$ subtasks and the holistic $Q{\rightarrow}AR$ task. For train and validation splits, the correct labels are available for development. To prevent overfitting, the test set labels were not released. Since evaluation on the test set is a manual effort by Zellers et al. [37], we provide numbers for our best performing model on the test set and illustrate results for the ablation study on the validation set.

Quantitative Evaluation: Tab. 1 compares performance of variants of our approach to the current state-of-the-art R2C [37]. While we report validation accuracy on both subtasks ($Q{\rightarrow}A$ and $QA{\rightarrow}R$) and the joint ($Q{\rightarrow}AR$) task in Tab. 1, in the following discussion we refer to percentages with reference to $Q{\rightarrow}AR$.

We make two modifications to improve R2C. The first, is `Det-BN` where we calculate and use train time batch normalization [18] statistics. This makes evaluation independent of batch size and ordering. Second, we `freeze` all the weights of the image CNN in R2C, whereas Zellers et al. [37] keep the last layer trainable. With these two minor but useful changes we obtain an improvement (1.31%) in performance and a significant reduction in trainable parameters (15Mn). We use the shorthand `R2C++` to refer to this improved variant of R2C.

[2] Details in the supplementary material.

Model	$Q \rightarrow A$	$QA \rightarrow R$	$Q \rightarrow AR$
Revisited [19]	57.5	63.5	36.8
BottomUp [2]	62.3	63.0	39.6
MLB [22]	61.8	65.4	40.6
MUTAN [5]	61.0	64.4	39.3
R2C [37]	65.1	67.3	44.0
TAB-VCR (ours)	**70.4**	**71.7**	**50.5**

Table 2: **Evaluation on test set.** Accuracy on the three VCR tasks. Comparison with top VQA models + BERT performance (source: [37]). Our best model outperforms R2C [37] on the test set by a significant margin.

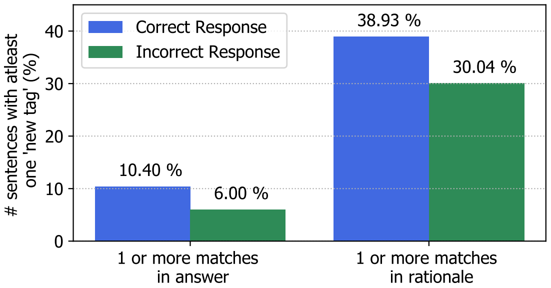


Figure 3: **New tags:** Percentage of response sentences with a *new tag*, *i.e.*, a new grounding for noun and object detection. Correct responses more likely have new detections than incorrect ones.

VCR subtask	Avg. no. of *tags* in query+response		
	(a) all	(b) correct	(c) errors
$Q \rightarrow A$	2.673	2.719	2.566
$QA \rightarrow R$	4.293	4.401	4.013

Table 3: Average number of tags in the query+response for the two subtasks for (a) all datapoints (b) datapoints where TAB-VCR was correct (c) datapoints where TAB-VCR made errors. Our model performs better on datapoints with more tags, i.e., richer association of image and text.

Ques. type	Matching patterns	Counts	$Q \rightarrow A$	$QA \rightarrow R$
what	what	10688	72.30	72.74
why	why	9395	65.14	73.02
isn't	is, are, was, were, isn't	1768	75.17	67.70
where	where	1546	73.54	73.09
how	how	1350	60.67	69.19
do	do, did, does	655	72.82	65.80
who	who, whom, whose	556	86.69	69.78
will	will, would, wouldn't	307	74.92	73.29

Table 4: Accuracy analysis by question type (with at least 100 counts) of `TAB-VCR` model. *Why* and *how* questions are most challenging for the $Q \rightarrow A$ subtask.

Our `base` model (described in Sec. 2) which includes (`Det-BN`) and `Freeze` improvements, significantly improves over `R2C++` by 1.78%, while having half the number of trainable parameters.

By using a more expressive ResNet as image CNN model (`Base + Resnet101`), we obtain another 1.32% improvement. We obtain another big increase of 2.57% by leveraging attributes capturing visual features (`Base + Resnet101 + Attributes`). Our best performing variant incorporates *new tags* during training and inference (`TAB-VCR`) with a final 50.62% on the validation set. We ablate `R2C++` with `ResNet101` and `Attributes` modifications, which leads to better performance too. This suggests our improvements aren't confined to our particular net.

In Tab. 2 we show results evaluating the performance of TAB-VCR on the private test set, set aside by Zellers et al. [37]. We obtain a 5.3%, 4.4% and 6.5% absolute improvement over R2C on the test set. We perform much better than top VQA models which were adapted for VCR in [37]. Models evaluated on the test set are posted on the leaderboard[3]. We appear as 'TAB-VCR' and outperform prior peer-reviewed work. At the time of submitting this camera-ready (31th Oct 2019), TAB-VCR ranked seventh among single models on the leaderboard. Based on the available reports [25, 32, 1, 24, 28, 7], most of these seven methods capture the idea of re-training BERT with extra information from Conceptual Captions [31]. This, in essence, is orthogonal to our *new tags* and attributes approach to build simple and effective baselines with significantly fewer parameters.

Qualitative Evaluation: We illustrate qualitative results in Fig. 2 and error analysis in Tab. 3 and Tab. 4. The error mode analysis will be provided in supplementary material.

4 Conclusion

We develop an effective baseline for visual commonsense reasoning. We leverage additional object detections to better ground noun-phrases. We show that the proposed approach outperforms state-of-the-art, despite significantly fewer trainable parameters, providing a basis for future development.

Acknowledgements This work is supported in part by NSF under Grant No. 1718221 and MRI #1725729, UIUC, Samsung, 3M, Cisco Systems Inc. (Gift Award CG 1377144) and Adobe. We thank NVIDIA for providing GPUs used for this work and Cisco for access to the Arcetri cluster. The authors thank Prof. Svetlana Lazebnik for insightful discussions and Rowan Zellers for releasing and helping us navigate the VCR dataset & evaluation.

[3] `visualcommonsense.com/leaderboard`

References

[1] C. Alberti, J. Ling, M. Collins, and D. Reitter. Fusion of detected objects in text for visual question answering. *arXiv preprint arXiv:1908.05054*, 2019.

[2] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proc. CVPR*, 2018.

[3] L. Anne Hendricks, R. Hu, T. Darrell, and Z. Akata. Grounding visual explanations. In *ECCV*, 2018.

[4] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. VQA: Visual question answering. In *Proc. ICCV*, 2015.

[5] H. Ben-younes, R. Cadene, M. Cord, and N. Thome. Mutan: Multimodal tucker fusion for visual question answering. In *Proc. ICCV*, 2017.

[6] A. Chandrasekaran, V. Prabhu, D. Yadav, P. Chattopadhyay, and D. Parikh. Do explanations make vqa models more predictable to a human? *arXiv preprint arXiv:1810.12366*, 2018.

[7] Y.-C. Chen, L. Li, L. Yu, A. E. Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu. Uniter: Learning universal image-text representations. *arXiv preprint arXiv:1909.11740*, 2019.

[8] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. Moura, D. Parikh, and D. Batra. Visual Dialog. In *Proc. CVPR*, 2017.

[9] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu. Are you talking to a machine? Dataset and Methods for Multilingual Image Question Answering. In *Proc. NeurIPS*, 2015.

[10] Y. Goyal, T. Khot, A. Agrawal, D. Summers-Stay, D. Batra, and D. Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. *IJCV*, 2017.

[11] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. 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*, pages 6904–6913, 2017.

[12] L. A. Hendricks, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele, and T. Darrell. Generating visual explanations. In *Proc. ECCV*, 2016.

[13] R. Hu, M. Rohrbach, J. Andreas, T. Darrell, and K. Saenko. Modeling relationships in referential expressions with compositional modular networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1115–1124, 2017.

[14] R. Hu, J. Andreas, T. Darrell, and K. Saenko. Explainable neural computation via stack neural module networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 53–69, 2018.

[15] D. A. Hudson and C. D. Manning. Gqa: a new dataset for compositional question answering over real-world images. In *Proc. CVPR*, 2019.

[16] D. Huk Park, L. Anne Hendricks, Z. Akata, A. Rohrbach, B. Schiele, T. Darrell, and M. Rohrbach. Multimodal explanations: Justifying decisions and pointing to the evidence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8779–8788, 2018.

[17] D. Huk Park, L. Anne Hendricks, Z. Akata, A. Rohrbach, B. Schiele, T. Darrell, and M. Rohrbach. Multimodal explanations: Justifying decisions and pointing to the evidence. In *Proc. CVPR*, 2018.

[18] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proc. ICML*, 2015.

[19] A. Jabri, A. Joulin, and L. van der Maaten. Revisiting visual question answering baselines. In *Proc. ECCV*, 2016.

[20] Y. Jang, Y. Song, Y. Yu, Y. Kim, and G. Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2758–2766, 2017.

[21] J. Kim, A. Rohrbach, T. Darrell, J. Canny, and Z. Akata. Textual explanations for self-driving vehicles. In *ECCV*, 2018.

[22] J.-H. Kim, K. W. On, W. Lim, J. Kim, J.-W. Ha, and B.-T. Zhang. Hadamard Product for Low-rank Bilinear Pooling. In *The 5th International Conference on Learning Representations*, 2017.

[23] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.

[24] G. Li, N. Duan, Y. Fang, D. Jiang, and M. Zhou. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. *arXiv preprint arXiv:1908.06066*, 2019.

[25] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

[26] J. Lin, U. Jain, and A. G. Schwing. TAB-VCR: Tags and Attributes based VCR Baselines. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

[27] E. Loper and S. Bird. Nltk: the natural language toolkit. *arXiv preprint cs/0205028*, 2002.

[28] J. Lu, D. Batra, D. Parikh, and S. Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*, 2019.

[29] G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[30] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017.

[31] P. Sharma, N. Ding, S. Goodman, and R. Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*, 2018.

[32] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019.

[33] T. Tommasi, A. Mallya, B. Plummer, S. Lazebnik, A. C. Berg, and T. L. Berg. Combining multiple cues for visual madlibs question answering. *International Journal of Computer Vision*, 127(1):38–60, 2019.

[34] R. Vedantam, C. Lawrence Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. In *Proc. CVPR*, 2015.

[35] L. Wang, Y. Li, and S. Lazebnik. Learning Deep Structure-Preserving Image-Text Embeddings. In *Proc. CVPR*, 2016.

[36] Z. Wu and M. Palmer. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, 1994.

[37] R. Zellers, Y. Bisk, A. Farhadi, and Y. Choi. From recognition to cognition: Visual commonsense reasoning. In *Proc. CVPR*, 2019.

[38] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei. Visual7W: Grounded Question Answering in Images. In *Proc. CVPR*, 2016.

5 Supplementary Material for: A Simple Baseline for Visual Commonsense Reasoning

We structure the supplementary into three subsections.

1. Details about implementation and training routine, including hyperparamters and design choices.
2. Additional qualitative results including error modes

5.1 Implementation and training details

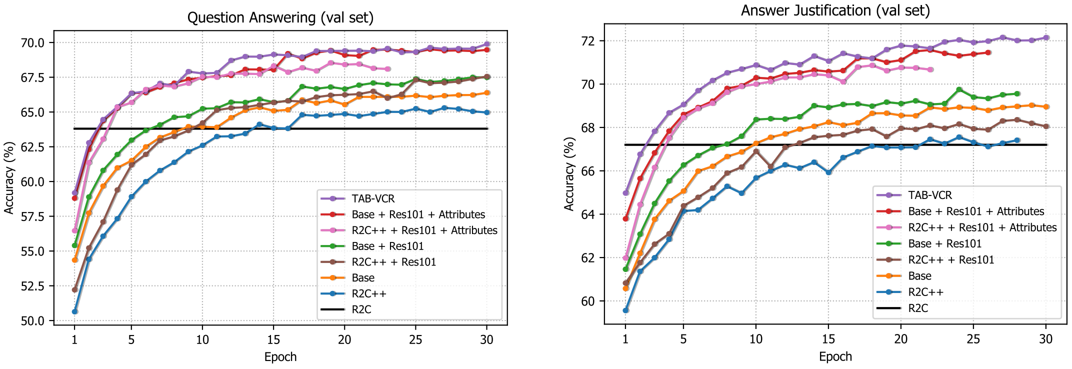


Figure 4: **Accuracy on validation set.** Performance for $Q{\rightarrow}A$ (left) and $QA{\rightarrow}R$ (right) tasks.

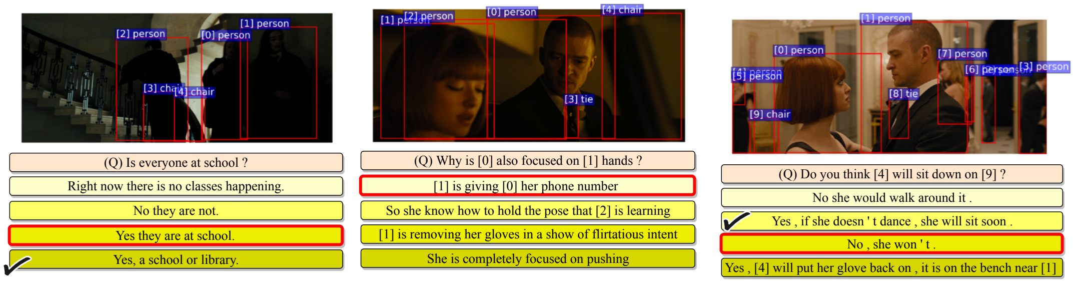
As explained in Fig. 1, our approach is composed of three components. Here, we provide implementation details for each: (1) BERT: Operates over query and response under consideration. The features of the penultimate layer are extracted for each word. Zellers et al. [37] release these embeddings with the VCR dataset and we use them as is. (2) Joint encoder: The output dimension of the CNN is 2048. The downsample net is a single fully connected layer with input dimension of 2048 (from the image CNN) and an output dimension of 512. We use a bidirectional LSTM with a hidden state dimension of $2 \cdot 256 = 512$. The outputs of which are average pooled. (3) MLP: Our MLP is much slimmer than the one from the R2C model. The pooled query and response representations are concatenated to give a $512 + 512 = 1024$ dimensional input. The MLP has a 512 dimensional hidden layer and a final output (score) of dimension 1. The threshold for Wu Palmer similarity k is set to 0.95.

We used the cross-entropy loss function for end-to-end training, Adam optimizer with learning rate $2\mathrm{e}{-4}$, and LR scheduler that reduce the learning rate by half after two consecutive epochs without improvement. We train our model for 30 epochs. We also employ early stopping, *i.e.*, we stop training after 4 consecutive epochs without validation set improvement. Fig. 4 shows validation accuracy for both the subtasks of VCR over the training epochs. We observe the proposed approach to very quickly exceed the results reported by previous state-of-the-art (marked via a solid horizontal black line).

5.2 Additional qualitative results

Examples of TAB-VCR performance on the VCR dataset are included in Fig. 6. They supplement the qualitative evaluation in the main paper (Fig. 2). Our model correctly predicts for each of these examples. Note how our model can ground important words. These are highlighted in **bold**. For instance, for Fig. 6(a), the correct rationale prediction is based on the expression of the **lamp**, which we ground. The lamp wasn't grounded in the original VCR dataset. Similarly grounding the **tag**, and **face** helps answer and reason for the image in Fig. 6(b) and Fig. 6(c). As illustrated via the **couch** in Fig. 6(d), it is interesting that the same noun is present in detections yet not grounded to words in the VCR dataset.

Error modes. We also qualitatively study TAB-VCR's shortcomings by analyzing error modes, as illustrated in Fig. 5. The correct answer is marked with a tick while our prediction is outlined in red.

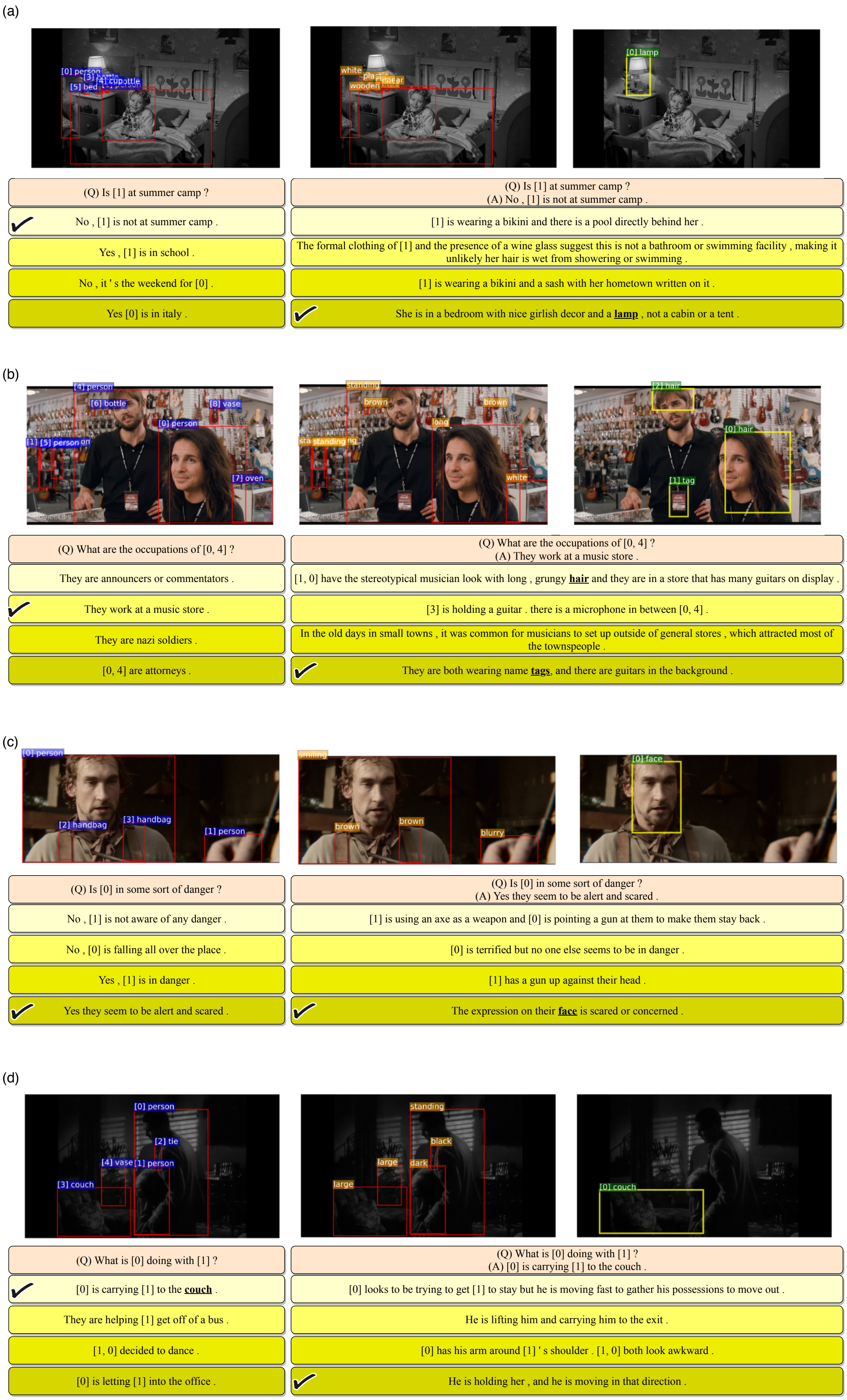

[2] person
[1] person
[0] person
[3] chair
[4] chair
(Q) Is everyone at school ?
Right now there is no classes happening.
No they are not.
Yes they are at school.
Yes, a school or library.
[1] person
[2] person
[0] person
[4] chair
[3] tie
(Q) Why is [0] also focused on [1] hands ?
[1] is giving [0] her phone number
So she know how to hold the pose that [2] is learning
[1] is removing her gloves in a show of flirtatious intent
She is completely focused on pushing
[1] person
[0] person
[5] person
[7] person
[6] person
[3] person
[8] tie
[9] chair
(Q) Do you think [4] will sit down on [9] ?
No she would walk around it .
Yes , if she doesn ' t dance , she will sit soon .
No , she won ' t .
Yes , [4] will put her glove back on , it is on the bench near [1]

(a) Similar responses (b) Missing context (c) Future ambiguity

Figure 5: **Qualitative analysis of error modes.** Correct answers are marked with ticks and our incorrect prediction is outlined in red. (a) shows options with overlapping meaning. Both the third and the fourth answer have similar meaning. (b) shows the error due to objects which aren't present in the image. (c) shows examples that have scenes offer an ambiguous future.

Examples include options with overlapping meaning (Fig. 5(a)). Both the third and the fourth answers have similar meaning which could be accounted for the fact that Zellers et al. [37] automatically curated competing incorrect responses via adversarial matching. Our method misses the 'correct' answer. Another error mode (Fig. 5(b)) is due to objects which aren't present in the image, like the "gloves in a show of flirtatious intent." This could be accounted to the fact that crowd workers were shown context from the video in addition to the image (video caption), which isn't available in the dataset. Also, as highlighted in Fig. 5(c), scenes often offer an ambiguous future, and our model gets some of these cases incorrect. We provide additional examples in Fig. 7. TAB-VCR gets the question answering subtask (left) incorrect, which we detail next.

Once the model knows the correct answer it can correctly reason about it, as evidenced by being correct on the answer justification subtask (right). In Fig. 7(a) both the responses 'Yes, she does like [1]' and 'Yes, she likes him a lot' are very similar, and our model misses the 'correct' response. Since the VCR dataset is composed by an automated adversarial matching, these options could end up being very overlapping and cause these errors. In Fig. 7(b) it is difficult to infer that the the audience are watching a live band play. This could be due to the missing context as video captions aren't available to our models, but were available to workers during dataset collection. In Fig. 7(c) multiple stories could follow the current observation, and TAB-VCR makes errors in examples with ambiguity regarding the future.



(a)

(Q) Is [1] at summer camp ?

✓ No , [1] is not at summer camp .

Yes , [1] is in school .

No , it ' s the weekend for [0] .

Yes [0] is in italy .

(Q) Is [1] at summer camp ?
(A) No , [1] is not at summer camp .

[1] is wearing a bikini and there is a pool directly behind her .

The formal clothing of [1] and the presence of a wine glass suggest this is not a bathroom or swimming facility , making it unlikely her hair is wet from showering or swimming .

[1] is wearing a bikini and a sash with her hometown written on it .

✓ She is in a bedroom with nice girlish decor and a **lamp** , not a cabin or a tent .

(b)

(Q) What are the occupations of [0, 4] ?

They are announcers or commentators .

✓ They work at a music store .

They are nazi soldiers .

[0, 4] are attorneys .

(Q) What are the occupations of [0, 4] ?
(A) They work at a music store .

[1, 0] have the stereotypical musician look with long , grungy **hair** and they are in a store that has many guitars on display .

[3] is holding a guitar . there is a microphone in between [0, 4] .

In the old days in small towns , it was common for musicians to set up outside of general stores , which attracted most of the townspeople .

✓ They are both wearing name **tags**, and there are guitars in the background .

(c)

(Q) Is [0] in some sort of danger ?

No , [1] is not aware of any danger .

No , [0] is falling all over the place .

Yes , [1] is in danger .

✓ Yes they seem to be alert and scared .

(Q) Is [0] in some sort of danger ?
(A) Yes they seem to be alert and scared .

[1] is using an axe as a weapon and [0] is pointing a gun at them to make them stay back .

[0] is terrified but no one else seems to be in danger .

[1] has a gun up against their head .

✓ The expression on their **face** is scared or concerned .

(d)

(Q) What is [0] doing with [1] ?

✓ [0] is carrying [1] to the **couch** .

They are helping [1] get off of a bus .

[1, 0] decided to dance .

[0] is letting [1] into the office .

(Q) What is [0] doing with [1] ?
(A) [0] is carrying [1] to the couch .

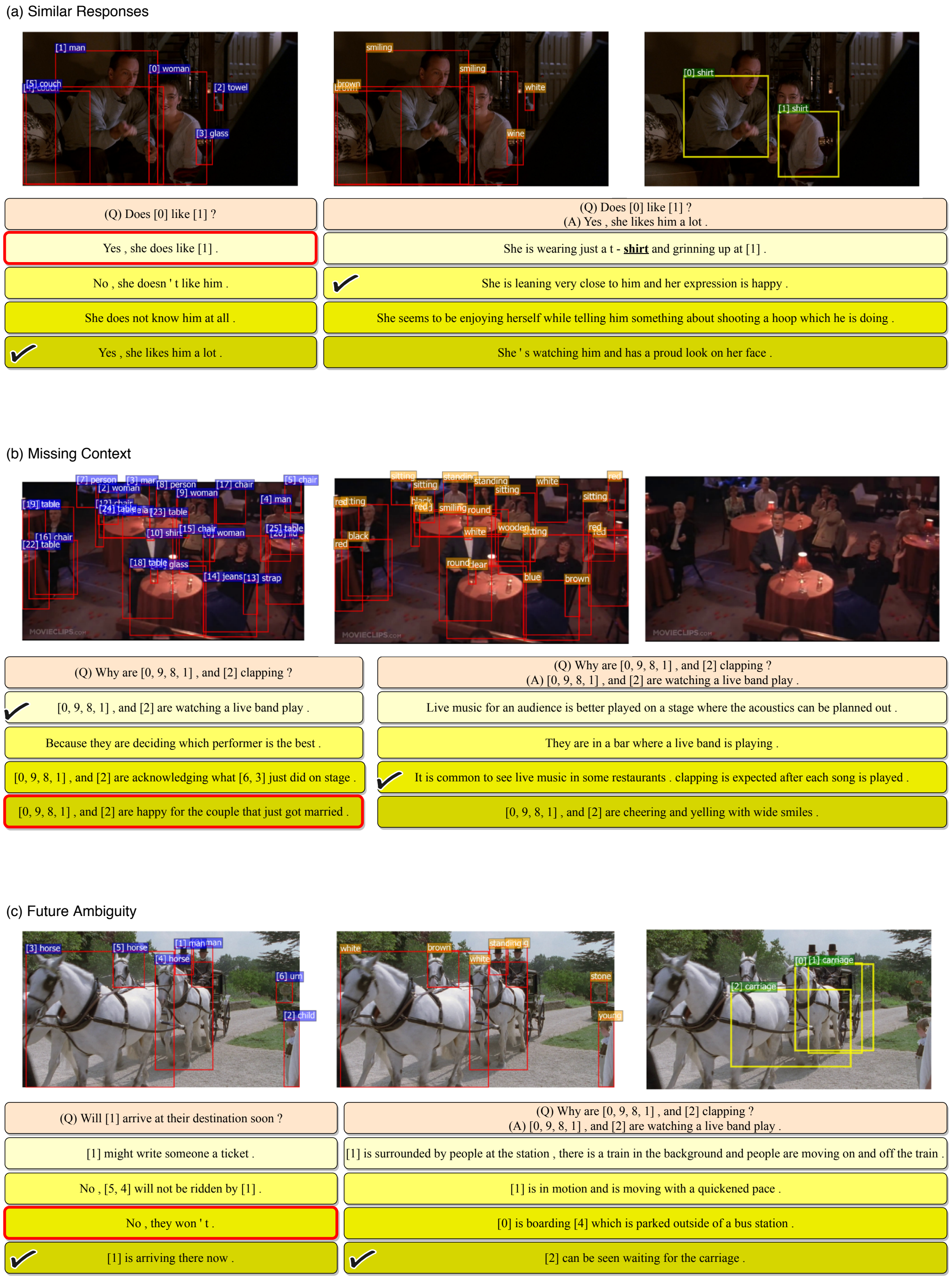
[0] looks to be trying to get [1] to stay but he is moving fast to gather his possessions to move out .

He is lifting him and carrying him to the exit .

[0] has his arm around [1] ' s shoulder . [1, 0] both look awkward .

✓ He is holding her , and he is moving in that direction .

Figure 6: **Qualitative results.** More examples of the proposed **TAB-VCR** model, which incorporates attributes and augments image-text grounding. The image on the left shows the object detections provided by VCR. The image in the middle shows the attributes predicted by our model and thereby captured in visual features. The image on the right shows *new tags* detected by our proposed method. Below the images are the question answering and answer justification subtasks. The *new tags* are highlighted in **bold**.



(a) Similar Responses

(Q) Does [0] like [1] ?

Yes , she does like [1] .

No , she doesn ' t like him .

She does not know him at all .

✓ Yes , she likes him a lot .

(Q) Does [0] like [1] ?
(A) Yes , she likes him a lot .

She is wearing just a t - **shirt** and grinning up at [1] .

✓ She is leaning very close to him and her expression is happy .

She seems to be enjoying herself while telling him something about shooting a hoop which he is doing .

She ' s watching him and has a proud look on her face .

(b) Missing Context

(Q) Why are [0, 9, 8, 1] , and [2] clapping ?

✓ [0, 9, 8, 1] , and [2] are watching a live band play .

Because they are deciding which performer is the best .

[0, 9, 8, 1] , and [2] are acknowledging what [6, 3] just did on stage .

[0, 9, 8, 1] , and [2] are happy for the couple that just got married .

(Q) Why are [0, 9, 8, 1] , and [2] clapping ?
(A) [0, 9, 8, 1] , and [2] are watching a live band play .

Live music for an audience is better played on a stage where the acoustics can be planned out .

They are in a bar where a live band is playing .

✓ It is common to see live music in some restaurants . clapping is expected after each song is played .

[0, 9, 8, 1] , and [2] are cheering and yelling with wide smiles .

(c) Future Ambiguity

(Q) Will [1] arrive at their destination soon ?

[1] might write someone a ticket .

No , [5, 4] will not be ridden by [1] .

No , they won ' t .

✓ [1] is arriving there now .

(Q) Why are [0, 9, 8, 1] , and [2] clapping ?
(A) [0, 9, 8, 1] , and [2] are watching a live band play .

[1] is surrounded by people at the station , there is a train in the background and people are moving on and off the train .

[1] is in motion and is moving with a quickened pace .

[0] is boarding [4] which is parked outside of a bus station .

✓ [2] can be seen waiting for the carriage .

Figure 7: **Qualitative analysis of error modes.** Responses with (a) similar meaning, (b) lack of context and (c) ambiguity in future actions. Correct answers are marked with ticks and our models incorrect prediction is outlined in red.