

Examining Consistency of Visual Commonsense Reasoning based on Person Grounding

Huiju Kim¹, Youjin Kang¹, SangKeun Lee^{1,2}

¹Department of Computer Science and Engineering, ²Department of Artificial Intelligence,
Korea University, Seoul, Republic of Korea
{haena0320, yjkang10, yalph}@korea.ac.kr

Abstract

Given an image depicting multiple individuals, humans are capable of inferring each individual’s emotions, intentions, and social norms based on commonsense understanding. However, a machine’s ability of commonsense reasoning about distinct individuals in images remains underexplored. In this study, we examine the consistency of visual commonsense reasoning based on person grounding. We introduce a novel test dataset called **Visual Commonsense Reasoning-Contrast Sets** (VCR-CS) to evaluate whether models can reason about individual people in an image by changing the person tags in the questions and answers. We benchmark various vision-language models on VCR-CS and observe that they fail in consistent commonsense reasoning about different people in one image, showing a performance decrease of up to 31.5%. To mitigate such failures, we propose a multi-task learning framework called **Person-centric groundIng eNhanced Tuning** (PINT). Our framework enhances a model’s ability to perform person-grounded commonsense reasoning by leveraging two novel person-centric pretraining tasks: Image Person-based Text Matching and Person-Masked Language Modeling. The experimental results revealed the effectiveness of PINT by showing the lowest performance degradation on VCR-CS and the improvements in consistency and sensitivity metrics. Our dataset and code are publicly available ¹.

1 Introduction

Commonsense reasoning from visual scenes involves inferring about people’s emotions, intentions, and social norms based on a commonsense understanding of the given image (Zellers et al., 2019). It plays a crucial role when machines are required to operate in person-centric scenarios by

¹<https://github.com/Haena0320/consistency-pg>

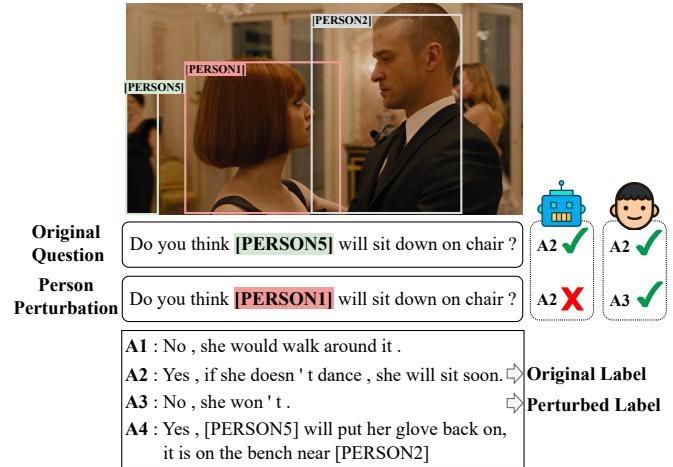


Figure 1: Examples of questions about perturbed person tags with different labels. Existing models face challenges in commonsense reasoning about a person when person tags are perturbed.

leveraging commonsense knowledge about people’s thoughts, behaviors, and interactions in dynamic situations (You et al., 2022). Such an ability should be consistently applied, even though commonsense reasoning processes may differ for different individuals in different situations. For example, Figure 1 depicts the different situations for [PERSON1] and [PERSON5], which leads to varying reasoning processes for individuals. This type of person-grounded commonsense reasoning often shapes the inferences about people in a given situation.

In the fields of vision and language, several datasets (Zellers et al., 2019; Park et al., 2020; Lei et al., 2020; Dong et al., 2022; You et al., 2022) that focus on the reasoning about individuals using visual commonsense knowledge have been proposed. In one of the notable datasets, VCR (Zellers et al., 2019), the models are required to provide answers with justifications for commonsense questions related to the individuals in the given images. In such

scenarios, the ability of vision-language (VL) models to leverage accurate person grounding is a significant factor in commonsense reasoning (Zellers et al., 2019, 2021, 2022), however, this has not received enough research attention.

A recently released dataset called HumanCog (You et al., 2022) focuses on person-centric visual grounding, which requires reasoning regarding which person in an image is being referred to in the commonsensical description. However, we argue that in general situations where such commonsense explanations are not explicitly provided, appropriate commonsense reasoning about individuals should be performed. Moreover, evaluating the consistent commonsense reasoning abilities of various individuals depicted in images remains challenging. Interestingly, in our pilot experiment, we observed that various Transformer-based VL models (Lu et al., 2019; Gan et al., 2020; Chen et al., 2020b; Zellers et al., 2022; Cho et al., 2021) trained on VCR achieved an accuracy greater than 40% on VCR subsets where person-grounding information was not provided.

In this study, we propose a novel test dataset called **Visual Commonsense Reasoning-Contrast Sets** (VCR-CS) to investigate the consistent commonsense reasoning abilities of individuals depicted in images. VCR-CS is a challenging dataset that leads the model to predict incorrect answers when the model ignores the person referred to in the text description. The dataset comprises original VCR validation examples and manually edited contrast examples in which the person mentioned in the original question is changed to another person in such a manner that the gold label changes. We then benchmark six visual commonsense reasoning models on VCR-CS and observed a significant performance decrease ($\sim 31.5\%$) on the suggested hard-negative examples. We then evaluate the models on VCR-CS using three metrics: accuracy, consistency, and sensitivity. Consistency estimates the model’s ability to predict correct answers across the original and contrast examples, and sensitivity measures whether predictions change after perturbations in the person tags.

We further present **P**erson-centric ground**I**ng eNhanced Tuning (PINT), which is a novel multi-task learning framework that enhances the model’s ability in commonsense reasoning about different individuals within an image. PINT comprises two person-centric pre-training tasks: (i)

Image Person-based Text Matching (IPTM) and (ii) Person-Masked Language Modeling (PMLM). IPTM task guides the model to learn the alignment between images and text queries by focusing on person links in the text. In PMLM task, the model is trained to reconstruct masked person links using a cross-modal context. Extensive experiments revealed that PINT achieved the best performance for most metrics on VCR-CS. Specifically, our experimental results show that PINT improved the consistency by more than 25%, and sensitivity by 15% on VCR-CS dataset. To summarize, our contributions are as follows:

- In this study, we examine the consistency of visual commonsense reasoning (VCR) systems based on person grounding.
- We propose a test dataset called VCR-CS, to evaluate whether VCR models can reason about individual people in an image. We benchmark six VL models and observe a performance decrease (up to 31.5%).
- Furthermore, we introduce PINT, which is an effective multi-task learning framework, to enhance a model’s ability in person-grounded commonsense reasoning.

2 Preliminaries

2.1 Vision-Language Model

Transformer-based VL models (Chen et al., 2020b; Gan et al., 2020; Lu et al., 2019; Li et al., 2019; Yu et al., 2021) benefit from multimodal pre-training to learn universal image-text representation. Given a single image-text pair (I, T) in a pre-training dataset, the model first encodes the image and text inputs as feature vectors, and the vectors pass through the multilayer transformer to learn cross-modal representations. To learn rich multimodal representations, previous studies (Gan et al., 2020; Lu et al., 2019; Chen et al., 2020b; Zellers et al., 2021, 2022; Cho et al., 2021) mainly employed two representative multimodal pre-training tasks: (i) Image-Text Matching (ITM) task and (ii) Masked Language Modeling (MLM) task. ITM task learns to predict whether I and T are aligned, and MLM task learns to reconstruct the corrupted text inputs with [MASK] tokens given the multimodal context inputs.

| Mode | Original example | Masked example |
|------|--|---|
| Q | Did [PERSON1] cause all this destruction? | Did [MASK] cause all this destruction? |
| GT | No, [PERSON1] appears to have had no role. | No, [MASK] appears to have had no role. |
| DT | Yes It was [PERSON1] who started it. | Yes It was [MASK] who started it. |

Table 1: Example of person-masked modification. The original example is changed into a masked version (Masked example). In this modification, we replaced all references of persons with [MASK] tokens in question (Q), ground truth answer (GT), and distractors (DT). Here, we report on one of the distractors for reference.

2.2 Visual Commonsense Reasoning

Task Given an image, I , VCR task can be decomposed into two subtasks: (1) $q \rightarrow a$: Given question Q , choose the correct answer, A^+ , out of the four candidate answers. (2) $qa \rightarrow r$: Given question Q with the correct answer, A^+ , select the correct rationale, R^+ , to justify A^+ from the four rationale candidates. By integrating these two subtasks, the $q \rightarrow ar$ metric measures whether the model chooses the correct answer with the proper rationale.

Fine-tuning Strategy Following recent studies (Chen et al., 2020b; Gan et al., 2020; Zellers et al., 2022, 2021), we fine-tuned the model to both subtasks simultaneously by decomposing the multi-choice settings into binary classification problems. Mathematically, the objective function for a single VCR example is:

$$L_{q \rightarrow a}(\theta) = -\log P_\theta(a_i | I, Q, A_i) \quad (1)$$

$$L_{qa \rightarrow r}(\theta) = -\log P_\theta(r_i | I, Q, A^+, R_i) \quad (2)$$

where A_i and R_i denote the i -th answer and rationale candidate and $a_i \in \{0, 1\}$ and $r_i \in \{0, 1\}$ are binary labels representing whether A_i and R_i are correct. θ represents a VL model with a softmax classifier that outputs a predicted probability distribution. Finally, the model is trained to minimize the objective function, L_{VCR} , as follows:

$$L_{VCR}(\theta) = L_{q \rightarrow a}(\theta) + L_{qa \rightarrow r}(\theta) \quad (3)$$

3 Pilot Experiments

We conducted a pilot experiment to investigate whether models trained on VCR can predict correct answers without seeing the person referred to in the queries.

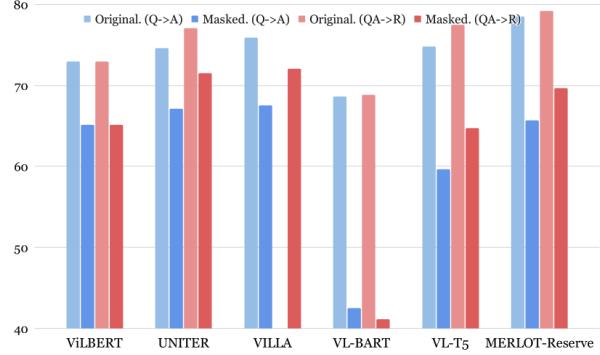


Figure 2: The existing VCR model’s performance on original (Origin.) subset and masked modification (Masked.) settings.

Person-Masked Modification As shown in Table 1, we conduct an experiment employing person-masked modifications. In this setting, we replaced the person links with [MASK] tokens. Intuitively, if the model cannot predict an answer without seeing the person about whom the question is being asked, the accuracy is similar to that of random settings. From Figure 2, it is evident that the entire model could predict the correct answers with a probability greater than 40%, without seeing any individual.

4 Dataset Construction

To examine the grounded commonsense reasoning of VL models, we proposed a new evaluation dataset called VCR-CS, which is built on top of VCR dataset. VCR-CS comprises pairs of “original” and “contrast” examples, which are denoted by $\{I, q_1, a, g_1\}$ and $\{I, q_2, a, g_2\}$, respectively. Each example includes an image, a question, multiple-choice answers, and a gold label. The two examples within a pair are distinguished from each other in a simple yet carefully designed manner to investigate person-grounded reasoning.

Instance candidate selection The creation process begins by selecting the original examples from VCR validation split. In the candidate selection phrase, we excluded instances that either non-person tags in the questions or belonged to the “why” and “where” question types, because these are deemed to inappropriate for generating contrast examples. Moreover, we only consider the examples in which the number of individuals detected in the image ranged from 2 to 15.

Perturbed instance generation In this phase, the original question, q_1 , is manually selected from

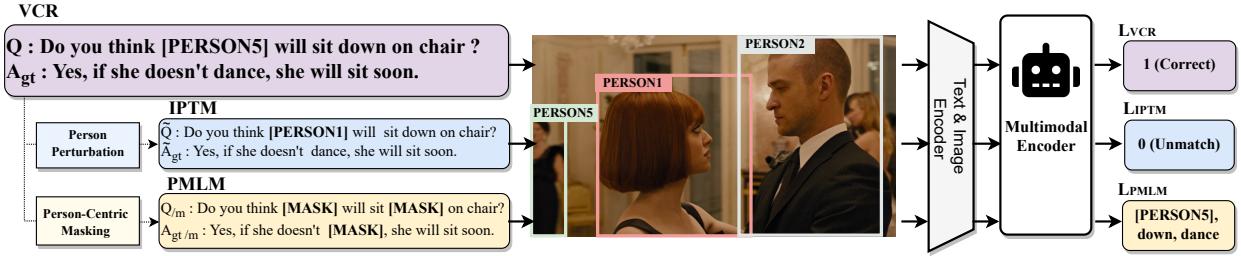


Figure 3: PINT, the proposed multi-task learning framework for the person-grounded VCR model. In PINT, models are trained by sharing global parameters in three tasks (VCR, IPTM, and PMLM).

the candidates. The two authors then manually substitute the person reference in q_1 with another person observed in an image to create a contrast question, q_2 . The people queried in each contrast instance are deliberately selected to have different intentions or actions compared to the people described in the original question. Therefore, the expected gold label, which is denoted by g_1 , for the original instance is different from the gold label, which is denoted by g_2 , for the contrast instance ($g_1 \neq g_2$). For example, as shown in Figure 1, if [PERSON5] in the question of the original example is changed to [PERSON1] in the question of the contrast example, the gold label for the contrast example is changed from A2 in the original example to A3. Therefore, we can expect VCR models to predict A2 as the correct answer when the person mentioned in the question is [PERSON5] and predict A3 when [PERSON1].

Validation To ensure high quality, annotators from Amazon Mechanical Turk (AMT) verify the labels of the instance pairs. Each instance is evaluated by five annotators, and the final label for the example is determined by a majority vote. If the label of the original instance, provided in advance by VCR, does not match the result of the majority vote, the latter is adopted as the final gold label. The agreement score between the annotators measured using Fleiss’s kappa (Fleiss, 1971) is 0.64 (the indicating “Substantial agreement” degree of agreement). Consequently, VCR-CS dataset contains 159 instance pairs². We offer various analyses and distributions of VCR-CS in Appendix A.

5 The Person-Grounded VCR Model

To enhance person-grounded reasoning ability, we present a new framework, PINT, which improves the model’s ability to perform reasoning about different individuals within an image. PINT is a multi-task learning framework consisting of two pre-training tasks: i) IPTM and ii) PMLM. In PINT, the model learns from both suggested tasks and VCR at the same time. Figure 3 depicts our suggested framework.

5.1 Image Person-based Text Matching

IPTM task aims to determine whether a given question is relevant to the image context by focusing on the person mentioned in the question. We design this task as an image-text matching task that includes hard negatives, where the person links mentioned in the questions are perturbed.

The training set for our IPTM task is built by reconstructing VCR training set and employing this person-perturbation strategy to obtain hard-negative examples. For each epoch, we randomly select 50% of VCR training examples and apply the following algorithm to the question of these examples to generate hard negatives. First, we create a set of $\{[\text{PERSON}\#]\}_{\#=1}^{81}$ person links. We consider the person links mentioned in each question as the target links that need to be perturbed. If the image of the example depicts two or more people, the target link is swapped for a person link referring to another person in the image. If only one person appears in the image, the target link is changed to another person link, randomly selected from the set of person links. If the above conditions were not satisfied, the question is replaced by a question of randomly selected example.

The person perturbation strategy uses the question Q , the ground truth answer A , the ground truth rationale R in each example of VCR to generate

²We experimentally determine the evaluation scale by observing the convergence of the performance of models (see Appendix C).

a perturbed question, \tilde{Q} , a perturbed ground truth answer \tilde{A} , and a perturbed ground truth rationale \tilde{R} . The text sequence $S \in \{S^{qa}, S^{qar}\}$ is produced by either concatenating $\{Q, A\}$ or $\{Q, A, R\}$ and is assumed to align with image V . The text sequence $\tilde{S} \in \{\tilde{S}^{qa}, \tilde{S}^{qar}\}$ is then generated by concatenating either $\{\tilde{Q}, \tilde{A}\}$ or $\{\tilde{Q}, \tilde{A}, \tilde{R}\}$, and is considered unaligned with image V . The model performs binary classification on both the aligned pair (S, V) and the misaligned pair (\tilde{S}, V) , where it is trained to minimize the objective function described below:

$$L_{IPTM} = -E_{(S,V) \sim D} [y \log f_\theta(S, V) + (1 - y) \log(1 - f_\theta(\tilde{S}, V))] \quad (4)$$

where f_θ is a vision-language model with a sigmoid classifier that outputs a normalized probability vector indicating whether S or \tilde{S} and V are aligned.

5.2 Person-Masked Language Modeling

The goal of PMLM task is to recover corrupted tokens (mainly person links) based on observations of their surrounding tokens and visual regions. This task forces the model to learn the fine-grained connections between the person links and the location of the persons in the image during training.

This strategy aims to construct a corrupted input sequence by masking the person links. Given the text sequence S , the object-region links mentioned in the descriptions are replaced with their object names. This preprocessing results in two types of tokens for the input sequence: person links and common words. Then, we apply the two masking strategies to this sequence. First, person links in a given sequence are randomly selected with a probability of 50%. All selected person links are replaced with [MASK] tokens. This masking strategy increases the sensitivity to person links and guides the model to capture different personal information from distinct tokens. Secondly, if we have 15% of the remaining masking budget, we select common words and decompose this masking budget into 10% random, 10% unconverted, and 80% [MASK] tokens.

The objective function of PMLM is defined as follows:

$$L_{PMLM}(\theta) = -E_{(S,V) \sim D} \log P_\theta(S_m | S_{/m}, V) \quad (5)$$

where $S_{/m}$ is the corrupted token sequence obtained by masking S ; m is the set of masked token indices; and S_m is the masked token sequence.

5.3 Overall Training Objectives

Finally, we train the model with three tasks using an overall loss function defined as follows:

$$L_{total} = L_{VCR} + L_{IPTM} + L_{PMLM} \quad (6)$$

6 Experiments and Results

Human Evaluation To measure the difficulty of our tests, we sampled 50% of the contrast set pairs (79 pairs) and conducted an evaluation per question by five AMT annotators. If the workers failed to achieve majority voting or answered the question incorrectly, the question was considered incorrect.

Baseline We evaluated six re-implemented Transformer-based VL models on VCR-CS. The characteristics of the models differ according to their image encoding choices and Transformer architectures: UNITER (Chen et al., 2020b), VILLA (Gan et al., 2020), ViLBERT (Lu et al., 2019), VL-BART (Cho et al., 2021), VL-T5 (Cho et al., 2021), and MERLOT-Reserve (Zellers et al., 2022). Further details of the model architecture are provided in Appendix B.1

Implementation Details The models were trained on a VCR training set (using the published code and hyper-parameters reported in the original papers; see Appendix B.2). In our experiment, we reported the performance of all base-sized models, except for ViLBERT. We evaluated the six trained models on VCR-CS and VCR validation sets. We fine-tuned VILLA model using PINT scheme on VCR training set. We adopted an additional training schedule based on a previous study (Chen et al., 2020a). Specifically, in the early training process, we focused on training IPTM and PMLM and then shifted to training VCR towards the end of the process (see Appendix B.3).

Evaluation Metrics We introduce two metrics to evaluate the person-grounded commonsense reasoning ability: *Consistency* and *Sensitivity*. Following (Gardner et al., 2020), *Consistency* is defined as whether all the elements of the contrast set pair are accurately predicted. Mathematically, this is expressed as:

$$\text{Consistency} = \frac{1}{N} \sum_{i=1}^N \mathbb{1}[(y_o^i = g_o^i) \wedge (y_p^i = g_p^i)]$$

where the model-predicted labels y_o^i, y_p^i on i^{th} original and perturbed questions are matched to the gold labels, g_o^i and g_p^i , respectively.

| Methods | Model-size | VCR-CS | | | |
|---------------------------------------|------------|---------------|-----------------------|--------------|--------------|
| | | Original Acc. | Contrast Acc. | Consistency | Sensitivity |
| ViLBERT (Lu et al., 2019) | 245M | 62.89 | 41.50 (-21.39) | 21.38 | 50.75 |
| UNITER (Chen et al., 2020b) | 86M | 61.84 | <u>43.42</u> (-18.42) | 19.74 | 46.88 |
| VILLA (Gan et al., 2020) | 86M | 61.18 | 42.76 (-18.42) | 20.39 | 47.69 |
| VL-BART (Cho et al., 2021) | 220M | 70.44 | 38.99 (-31.45) | <u>24.52</u> | 59.09 |
| VL-T5 (Cho et al., 2021) | 400M | 69.18 | 42.13 (-27.05) | 23.89 | 49.35 |
| MERLOT-Reserve (Zellers et al., 2022) | 200M | <u>69.18</u> | 39.62 (-29.56) | <u>24.52</u> | 38.36 |
| PINT (ours.) | 86M | 61.84 | 46.71 (-15.13) | 25.66 | <u>58.21</u> |
| Human | - | 91.64 | 92.40 (+0.75) | 87.84 | 93.78 |

Table 2: Evaluation results of the existing VCR models and the effectiveness of our proposed model, PINT. Boldface text denotes the best scores and the underlined text denotes the second-best scores. In VCR-CS dataset, Original Acc. and Contrast Acc. represent the accuracies of the original and contrast examples, respectively. In the bracket under Contrast Acc., we show a decrease in accuracy from the original to the contrast examples

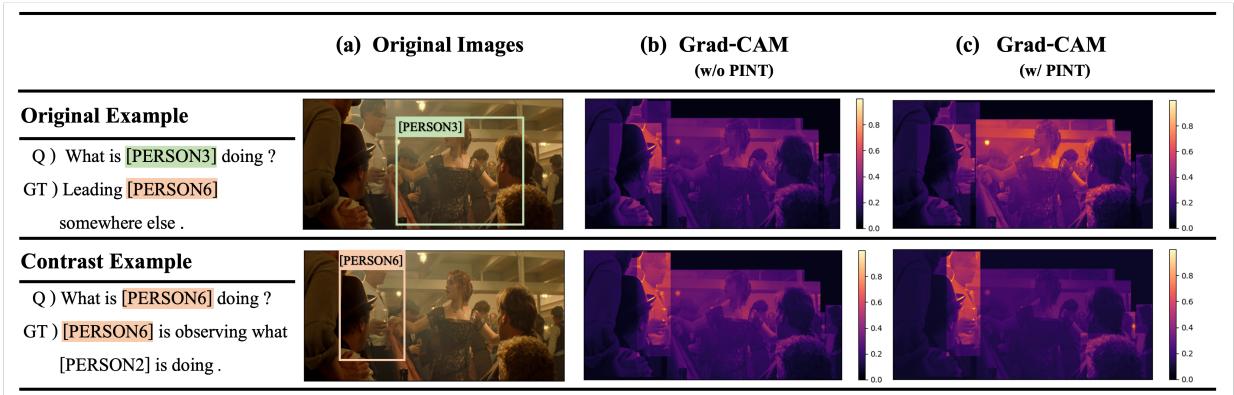


Figure 4: (a) depicts the image in original (top) and contrast examples (bottom); (b) and (c) visualize the person’s grounding without and with applying PINT by applying Grad-CAM for each original and contrast example.

Sensitivity is the average percentage when the examples of the contrast set pair were correctly predicted, in cases where the model predictions are different between the examples. It can be written as:

$$Sensitivity = \sum_{i=1}^N \frac{\mathbb{1}_{[(y_o^i = g_o^i) \wedge (y_p^i = g_p^i)]}}{\mathbb{1}_{[(y_o^i \neq y_p^i)]}}$$

Furthermore, we report the model accuracy of the original and contrast examples as an additional evaluation metric for VCR-CS.

Challenges faced by VCR models on VCR-CS
We evaluated the state-of-the-art models on VCR benchmarks on VCR-CS. Table 2 presents the evaluation results for the model. The baseline models exhibited a low performance. *Consistency* changes the ranking between existing VCR models, as the models with significant accuracy gap on VCR validation set (e.g., VILLA and MERLOT-Reserve) have small performance gap in *Consistency* and *Sensitivity*. The accuracy of MERLOT-Reserve,

which achieved the highest performance in VCR test and validation sets, decreased significantly from 69.18% to 39.62% after [PERSON#] perturbation, thereby showing a tendency similar to that of the other baseline models.

Effortless Success of Humans on VCR-CS To ensure that the contrast examples are not more difficult or noisy than the original examples, we evaluate whether humans will fail at them. We selected a set of 40 pairs from VCR-CS and measured human performance for the examples with AMT workers. Table 2 shows that the human performance on the contrast examples was similar to that on the original examples. Furthermore, unlike the baseline models, the human performance was 1.3% higher in the contrast examples and achieved 93.1% at *Sensitivity*. This indicates that humans usually make decisions based on the person links mentioned in a question.

PINT improves person-grounded reasoning
The central part of Table 2 presents the effective-

| Tasks | VCR-CS | | VCR | | |
|---------------------------|--------------|--------------|--------------|--------------|--------------|
| | Consistency | Sensitivity | Q→A | QA→R | Q→AR |
| VCR + ITM + MLM | 19.08 | 46.03 | 73.72 | 76.64 | 56.68 |
| VCR + IPTM + PMLM | 24.34 | 57.81 | 73.53 | 76.51 | 56.49 |
| VCR + IPTM + MLM | 18.42 | 45.9 | 73.87 | 76.54 | 56.94 |
| VCR + IPTM + PMLM (ours.) | 25.66 | 58.21 | 73.97 | 76.71 | 56.82 |

Table 3: Effect of the suggested person-centric pre-training tasks, IPTM and PMLM, compared to the original tasks, ITM and MLM. Performance in consistency and sensitivity metrics is enhanced by both IPTM and PMLM.

| Methods | VCR-CS | |
|---------------------------|--------------|--------------|
| | Consistency | Sensitivity |
| VILLA | 20.39 | 47.69 |
| w / Add person embeddings | 21.71 | 54.1 |
| w / Draw person boxes | 22.37 | 51.52 |
| w / PINT (ours.) | 25.66 | 58.21 |

Table 4: Effect of the different grounding methods on VCR-CS dataset. While the two methods that inject grounding information into the model input show a slight performance improvement, PINT that learns to focus on the grounding information achieved the highest performance gain.

ness of the proposed scheme (PINT). PINT shows the lowest performance degradation from the original to contrast examples in VCR-CS. It gains the performance by up to 5.27% and 10.52% on the two metrics: *Consistency* and *Sensitivity*. These results indicate that PINT, as a training strategy, improves the model’s reasoning ability based on person grounding. However, compared to human performance, there is still room for improvement.

Effectiveness of the person-centric tasks IPTM and PMLM are more challenging and enhanced versions of ITM and MLM, respectively. They generate training examples by focusing on person references, thereby allowing the model to learn person-grounded reasoning. In contrast, ITM and MLM randomly generate training instances. Specifically, we train the model by replacing IPTM with ITM or PMLM with MLM. In Table 3, we show that IPTM-ITM and PMLM-MLM variants performed inferiorly in terms of both *Consistency* and *Sensitivity* than PINT. This suggests that training on both ITM and PMLM tasks is effective for reasoning tasks that rely on person grounding. The results reveal the importance of training sophisticated person-centric tasks to improve a model’s ability in consistent person-grounded commonsense rea-

soning.

Attribution Visualization of PINT training

Training using PINT enhanced the models’ ability to reason about different individuals on a single image. Figure 4 presents a visualization of the Grad-CAM (Selvaraju et al., 2017) weights for VILLA trained without PINT (w/o PINT), and VILLA trained using PINT (w/ PINT). The brighter the area of the image, the more the model referred to the region when reasoning. For the given original question in the first row, “What is [PERSON3] doing?,” a model trained without PINT focuses more on irrelevant visual reasons, such as the [PERSON6]. In contrast, given the person-perturbed question in the second row, “What is [PERSON6] doing?,” a model trained with PINT focuses highly on the [PERSON6]-relevant regions and less on other objects. The visualization shows that PINT, when applied as a training strategy, effectively allows the model to focus more on the person described in the question.

Comparison with other relevant methods

In Table 4, we present an experiment comparing the person grounding method used for the person-centric visual commonsense task with our method. The first method follows the method in (Park et al., 2020), adds text embedding to the visual embedding corresponding to the person, and uses it as the input value for the model. We call this “add person embeddings.” The second method follows a previous study (Zellers et al., 2021, 2022) and displays the corresponding image area in color related to each person designation. We refer to this “Draw person boxes.” Both methods insert person-grounding information at the input stage, and our experiments confirmed that these methods failed to maximize the grounding-based reasoning ability in VCR-CS. Moreover, PINT was shown to be more effective than the existing ground-based methods.

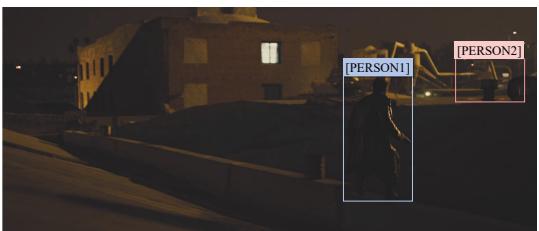
| Overall Loss | | | VCR-CS | | | | VCR | | |
|--------------|------------|------------|---------------|----------------|-------------|-------------|-------------------|--------------------|--------------------|
| L_{VCR} | L_{IPTM} | L_{PMLM} | Original Acc. | Contrast Acc. | Consistency | Sensitivity | $Q \rightarrow A$ | $QA \rightarrow R$ | $Q \rightarrow AR$ |
| ✓ | | | 61.18 | 42.76 (-18.42) | 20.39 | 47.69 | 74.84 | 77.55 | 58.25 |
| ✓ | ✓ | | 59.21 | 40.13 (-19.08) | 19.14 | 54.55 | 73.50 | 76.03 | 56.15 |
| ✓ | | ✓ | 61.84 | 44.08 (-17.76) | 22.37 | 53.97 | 74.07 | 76.03 | 56.89 |
| ✓ | ✓ | ✓ | 61.84 | 46.71 (-15.13) | 25.66 | 58.21 | 73.97 | 76.71 | 56.82 |

Table 5: Main ablative experiments of PINT on VCR-CS and original VCR validation sets. The performances in all VCR-CS metrics are improved by both IPTM and PMLM, with a slight decrease in VCR validation set.



Question: Why is [PERSON9] looking at [PERSON8]?

- A1. [PERSON9] is waiting for [PERSON8] to hand him some money
- A2. He is curious about what he's writing down
- A3. [PERSON8] has said something that has caught his interest **X**
- A4. [PERSON1] is waiting on [PERSON8] to replace the tire on his truck



Question: Why is a light on in the warehouse near [PERSON1]?

- A1. They are waiting for it to light up so [PERSON2] can play the game.
- A2. It is dark outside. **X**
- A3. The candle they had went out.
- A4. Criminals are hiding in the warehouse, and [PERSON1] is going to confront them.

Figure 5: Qualitative case analysis of VCR validation example where the base model (VILLA) succeeded, but PINT failed. Incorrect PINT predictions are marked with a red “X” and correct answers are in bold.

Ablation Study Reasoning based on person grounding is integrally learned from our suggested pre-training tasks and VCR task, which is essential for improving person-centric visual commonsense. To verify this, ablation studies were conducted using three objective function variants of PINT. Table 5 shows the performance improvement for training VILLA for each objective function that comprises PINT. L_{IPTM} and L_{PMLM} trained together with L_{VCR} help maximize both consistency and sensitivity, complementing each other to improve the grounding-based reasoning ability.

In addition, we noticed a trade-off between consistency, sensitivity, and accuracy while applying PINT during VCR training. Although there was a significant increase in the consistency and sensitivity on VCR-CS, we observed a slight decrease in the accuracy on VCR validation set when applying PINT strategy to VILLA. We suspect that this effect occurs because the model relies on spurious correlations (Ye and Kovashka, 2021) to achieve a high performance. Enhancing person-centric reasoning leads to improved consistency but a slight decline in accuracy. This suggests that our consistency and sensitivity metrics can effectively measure the ability of the model to reason about multiple people described in the images.

Limitations of PINT We performed a detailed qualitative analysis of the limits of PINT on VCR validation set in Figure 5. We marked the incorrect PINT prediction with a red “X” and the correct answer with a bold. We observed that PINT suffers in some examples in which the correct answer can be predicted by word overlap between the question and the answer. For example, in the example at the bottom in Figure 5, given an image depicting the outside of a faintly lit warehouse, PINT replies, “It’s dark outside”. In night scenarios, such a prediction may seem plausible, but it may fail if the overlap between the words used in the questions and answers, “in the warehouse” and “[PERSON1]”, can lead to the correct label.

7 Related Work

Person-Centric Vision-Language Task The person-centric vision-language task (Zellers et al., 2019; Dong et al., 2022; Cui et al., 2021; You et al., 2022), is mainly based on grounding references to a person; therefore, person-centric visual grounding ability is a crucial component. VCR (Zellers et al., 2019) is a task that answers commonsensical questions about the people depicted in an image.

The person-centric visual grounding task (Cui et al., 2021) aims to predict a mentioned person, given an image and a contextual textual description. The person-centric commonsense grounding task (You et al., 2022), which extends a person-centric visual grounding task to a commonsense domain, is designed to identify the person mentioned in the commonsense description in the image. However, they consider grounding and high-level reasoning as separate tasks and focus on each single task. Our study differs from the above-mentioned studies in that we address commonsense reasoning based on person grounding, which can fill the gap between the two tasks.

Consistency on Contrast Sets Language-based adversarial examples were generated to investigate the robustness of the models in the natural language processing and vision-language fields (Zhou et al., 2020; Wang et al., 2021; Jin et al., 2020; Akula et al., 2020; Gardner et al., 2020; Jimenez et al., 2022). In the natural language processing fields, the language model’s commonsense reasoning ability is investigated by generating and evaluating dual test samples (Zhou et al., 2020). In vision-language fields, the grounding abilities of the visual referring expression models were measured by manipulating the word order of text descriptions and verifying whether the grounding was performed correctly (Akula et al., 2020). It is found that the model’s performance on various tasks is significantly lower on the contrast sets, which are created by manually changing words in a manner that changes the gold labels (Gardner et al., 2020). Although they focused on analyzing a model’s poor performance using contrast sets, they did not consider strategies for improving model performance. Our approach suggests a novel training method, PINT, to improve the model’s reasoning ability, even though we use contrast sets that are similar to the previous methods.

Task-Specific Transfer Learning Although large-scale pre-trained language models (PLM) (Devlin et al., 2019; Liu et al., 2019; Clark et al., 2020) are fine-tuned for various NLP tasks to achieve competitive performance, this fine-tuning approach has limitations in capturing the important patterns of downstream tasks (Yuan et al., 2023; Dodge et al., 2020; Gu et al., 2020). To mitigate such a problem, continuous pre-training and fine-tuning regularization techniques (Hua et al., 2021;

Qu et al., 2021; Gururangan et al., 2020; Gu et al., 2020) are proposed. Continual pre-training of PLM on given downstream domain data has proven effective for final target task performance (Gururangan et al., 2020). Task-specific pre-training with a selective masking strategy is suggested for learning task-specific expressions based on domain data (Gu et al., 2020). We adopted target-specific tasks, such as a person-centric masking strategy; however, in contrast to the above studies, our framework additionally focuses on person-centric image-text matching in a self-supervised manner.

8 Conclusion

In this study, we examined the consistency of visual commonsense reasoning systems based on person grounding. We proposed a novel test dataset called VCR-CS to evaluate whether the models can reason about individuals depicted in an image. We demonstrated that the models trained on VCR dataset exhibited a limited capacity for consistent reasoning regarding different individuals depicted in a given image. To mitigate this problem, we designed a multitask learning framework, PINT, which learns from two person-centric pre-training tasks: IPTM and PMLM. Our experiments show that PINT enhances a model’s ability in person-grounded commonsense reasoning, as indicated by the minimal performance decline in VCR-CS and improvements in both consistency and sensitivity metrics.

Acknowledgement

This work was supported by the Basic Research Program through the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (2021R1A2C3010430).

Limitations

In this work, we proposed a test dataset and a training methodology to examine and improve the models’ ability for person-grounded commonsense reasoning. However, the performance gap with humans shown in experimental results suggests the need for a training set to learn more granular and person-grounded commonsense reasoning. In future studies, we plan to construct a larger dataset for training and extensive evaluation by incorporating an automatic process into VCR-CS construction pipeline.

References

- Arjun Akula, Spandana Gella, Yaser Al-Onaizan, Song-Chun Zhu, and Siva Reddy. 2020. **Words aren’t enough, their order matters: On the robustness of grounding visual referring expressions.** In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6555–6565.
- Sanyuan Chen, Yutai Hou, Yiming Cui, Wanxiang Che, Ting Liu, and Xiangzhan Yu. 2020a. **Recall and learn: Fine-tuning deep pretrained language models with less forgetting.** In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 7870–7881.
- Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020b. **UNITER: Universal image-text representation learning.** In *European Conference on Computer Vision*, pages 104–120.
- Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. **Unifying vision-and-language tasks via text generation.** In *International Conference on Machine Learning*, pages 1931–1942.
- Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. **ELECTRA: Pre-training text encoders as discriminators rather than generators.** In *International Conference on Learning Representations*.
- Yuqing Cui, Apoorv Khandelwal, Yoav Artzi, Noah Snavely, and Hadar Averbuch-Elor. 2021. **Who’s waldo? linking people across text and images.** In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1374–1384.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. **Bert: Pre-training of deep bidirectional transformers for language understanding.** In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186.
- Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. **Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping.** *arXiv preprint arXiv:2002.06305*.
- Qingxiu Dong, Ziwei Qin, Heming Xia, Tian Feng, Shoujie Tong, Haoran Meng, Lin Xu, Zhongyu Wei, Weidong Zhan, Baobao Chang, Sujian Li, Tianyu Liu, and Zhifang Sui. 2022. **Premise-based multimodal reasoning: Conditional inference on joint textual and visual clues.** In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 932–946.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. **An Image is Worth 16x16 Words: Transformers for image recognition at Scale.** In *International Conference on Learning Representations*.
- Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–782.
- Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. 2020. **Large-scale adversarial training for vision-and-language representation learning.** *Advances in Neural Information Processing Systems*, 33:6616–6628.
- Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihaio Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. **Evaluating models’ local decision boundaries via contrast sets.** In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323.
- Yuxian Gu, Zhengyan Zhang, Xiaozhi Wang, Zhiyuan Liu, and Maosong Sun. 2020. **Train no evil: Selective masking for task-guided pre-training.** In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 6966–6974.
- Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. **Don’t stop pretraining: Adapt language models to domains and tasks.** In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360.
- Hang Hua, Xingjian Li, Dejing Dou, Cheng-Zhong Xu, and Jiebo Luo. 2021. **Noise stability regularization for improving bert fine-tuning.** In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3229–3241.
- Carlos E. Jimenez, Olga Russakovsky, and Karthik Narasimhan. 2022. **CARETS: A consistency and robustness evaluative test suite for VQA.** In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 6392–6405.
- Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. **Is bert really robust? a strong baseline for natural language attack on text classification and entailment.** In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8018–8025.
- Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. 2020. **What is more likely to happen next? video-and-language future event prediction.** In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 8769–8784.

- Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. VisualBERT: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in Neural Information Processing Systems*, 32:13–23.
- Jae Sung Park, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, and Yejin Choi. 2020. Visual-COMET: Reasoning about the dynamic context of a still image. In *European Conference on Computer Vision*, pages 508–524.
- Yanru Qu, Dinghan Shen, Yelong Shen, Sandra Sajeev, Jiawei Han, and Weizhu Chen. 2021. CoDA: Contrast-enhanced and diversity-promoting data augmentation for natural language understanding. *International Conference on Learning Representations*.
- Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 28:91–99.
- Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30:5998–6008.
- Wenjie Wang, Pengfei Tang, Jian Lou, and Li Xiong. 2021. Certified robustness to word substitution attack with differential privacy. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1102–1112.
- Keren Ye and Adriana Kovashka. 2021. A case study of the shortcut effects in visual commonsense reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3181–3189.
- Da Yin, Liunian Harold Li, Ziniu Hu, Nanyun Peng, and Kai-Wei Chang. 2021. Broaden the vision: Geodiverse visual commonsense reasoning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2115–2129.
- Haoxuan You, Rui Sun, Zhecan Wang, Kai-Wei Chang, and Shih-Fu Chang. 2022. Find someone who: Visual commonsense understanding in human-centric grounding. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5444–5454.
- Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. ERNIE-ViL: Knowledge enhanced vision-language representations through scene graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3208–3216.
- Hongyi Yuan, Zheng Yuan, Chuanqi Tan, Fei Huang, and Songfang Huang. 2023. HyPe: Better pre-trained language model fine-tuning with hidden representation perturbation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 3246–3264.
- Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From Recognition to Cognition: Visual Commonsense Reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6720–6731.
- Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. 2022. MERLOT reserve: Neural script knowledge through vision and language and Sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16375–16387.
- Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. 2021. MERLOT: Multimodal neural script knowledge models. *Advances in Neural Information Processing Systems*, 34:23634–23651.
- Xuhui Zhou, Yue Zhang, Leyang Cui, and Dandan Huang. 2020. Evaluating commonsense in pre-trained language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9733–9740.

Appendix

A Dataset Analysis

In this section, we offer dataset statistics and an analysis of question frequency, inference types, and bounding boxes of persons depicted in VCR-CS questions. The high-level dataset statistics are presented in Table 6. The average question length, answer length, and number of objects mentioned in VCR-CS and VCR validation sets are similar. The question of the over-image ratio denotes how different images are used for the questions. The results revealed that VCR-CS uses various questions from different images. The cumulative distribution

| | VCR-CS | VCR val. |
|--------------------------------|--------|----------|
| Number of questions | 159 | 26534 |
| Number of images | 143 | 9929 |
| Number of movies covered | 16 | 244 |
| Average question length | 6.38 | 6.63 |
| Average answer length | 7.08 | 7.65 |
| Average # of objects mentioned | 1.88 | 1.85 |
| Question versus Image ratio | 1.11 | 2.67 |

Table 6: Dataset statistics for VCR-CS compared to VCR validation set (VCR val.).

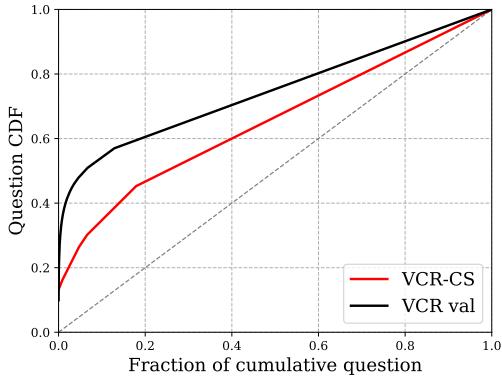


Figure 6: CDF of questions ordered by frequency. Red and black denote the questions from VCR-CS and VCR validation sets, respectively. We considered the two questions to be the same when they were equal after tokenization, lemmatization, and stop word removal.

function (CDF) of a question, ordered by frequency, is shown in Figure 6. The graph from VCR-CS is closer to $x = y$, indicating that various questions are contained compared with VCR validation set. Figure 9 shows the inference types required for the questions in VCR-CS. This indicates that VCR-CS is an underlying dataset with diverse types of reasoning. Notably, each answer requires more than one type of inference. We follow the type pattern in (Zellers et al., 2019). The center of the bounding box of the person objects obtained by normalization is shown in Figure 7. The red dots represent the original set, whereas the yellow dots represent the contrast set. The person positions are more widespread in the contrast set than in the original set. Figure 8 shows box plots of the normalized size of the bounding boxes of the person tags depicted in the questions. VCR-CS is collected to ask for various sizes of bounding boxes for persons in the images.

B Implementation Details

B.1 Baseline Details

We adopt six Transformer-based vision-language models as baselines. All the baseline models, ex-

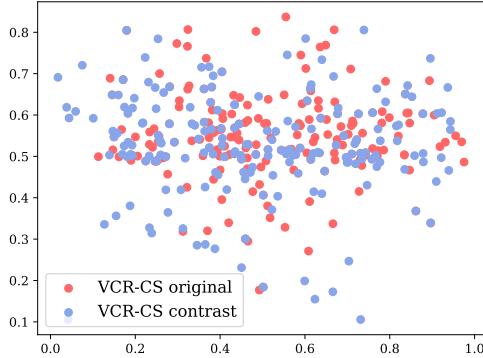


Figure 7: Normalized center position of the bounding box for the person depicted in the question. Red and yellow dots denote samples in the original and contrast examples on VCR-CS.

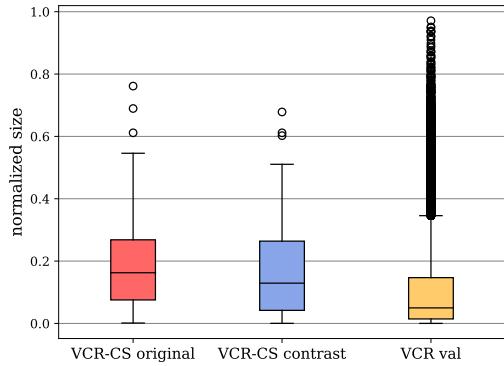


Figure 8: Comparison of the normalized bounding box sizes for the individuals depicted in the question of examples from both VCR-CS and VCR validation sets

cept for MERLOT-Reserve, use image features extracted from the Faster R-CNN (Ren et al., 2015) on images. In contrast, MERLOT-Reserve utilizes the object features extracted from ViT (Dosovitskiy et al., 2021). ViLBERT comprises two parallel Transformer (Vaswani et al., 2017) structures, whereas VILLA, UNITER, and MERLOT-Reserve adopt a single Transformer architecture to fuse cross-modal information. VL-T5 and VL-BART employ an encoder-decoder architecture.

B.2 Baseline Implementation Details

We followed the public code and hyperparameters of the original paper on the baseline models. We trained baseline models and PINT on VCR training dataset with single NVIDIA Tesla V100 GPU with 32GB of VRAM. The detailed code can be found at the following repositories:

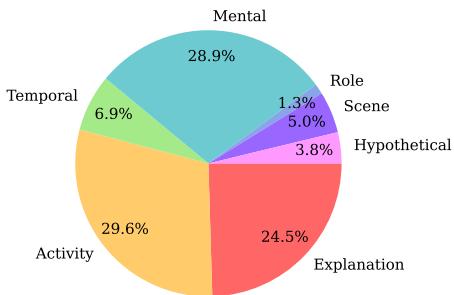


Figure 9: Overview of the inference types required by questions in VCR-CS.

| Modification Type | Masked Example |
|--------------------|----------------|
| Number of examples | 22,245 |

Table 7: VCR subset statistics used in pilot experiment.

B.2.1 ViLBERT

https://github.com/jiasenlu/vilbert_beta

B.2.2 UNITER

<https://github.com/ChenRocks/UNITER>

B.2.3 VILLA

<https://github.com/zhegan27/VILLA>

B.2.4 VL-T5 | VL-BART

<https://github.com/j-min/VL-T5>

B.2.5 MERLOT-Reserve

https://github.com/rowanz/merlot_reserve

B.3 PINT Implementation Details

We set a batch size of 32 data points, and searched learning rate between 6e-5, 5e-5 and 4e-5. We optimized the model employing the RecAdam (Chen et al., 2020a) optimizer, with a setting of $k = 0.1$ and $t_0 = 1000$. The training step was searched between 8000, 12000, 16000 and 24000.

C Performance Convergence in VCR-CS

It is only necessary that the contrast evaluation set be sufficiently large enough to verify the substantiated conclusions about the model behavior (Gardner et al., 2020). Recent studies (Gardner et al., 2020; Zhou et al., 2020) have used 70 to 646 contrast sets. Moreover, a study (Yin et al., 2021) used 108 to 282 QA pairs for evaluating VCR models. Therefore, we drew the performances of the

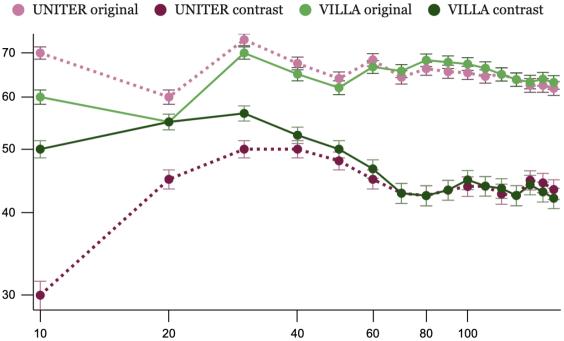


Figure 10: Accuracy on original and contrasting examples of VCR-CS for UNITER and VILLA with varying volumes of data. The x-axis represents the cumulative number of VCR-CS pairs on a logarithmic scale.

two models, UNITER and VILLA, according to the number of VCR-CS data points to validate their significance for 159 data points. Figure 10 shows that the overall performance of the models in VCR-CS gradually converged from approximately 100 data points.

D VCR subset in pilot experiments

As a motivation for VCR-CS construction, we conducted a pilot experiment. In the pilot experiment, we investigated whether the models can predict a correct answer without accurate notification of whom the question is asking. We show the statistics of the subset of VCR validation set used in our pilot experiment. It took about 83% of the overall validation dataset.

E Validation and Human Evaluation on VCR-CS

Our VCR-CS dataset was validated and evaluated with the help of 14 and 28 workers from AMT, respectively. VCR-CS dataset was validated, and human performance was assessed through AMT. Our worker selection setting was inspired by that of a previous study (Jimenez et al., 2022). A total of 14 and 28 workers participated in the validation and evaluation, respectively, of HITs. The validation instructions and human evaluation instructions can be found in Figures 11 and 12, respectively. Each HIT for data validation and evaluation comprised ten contrast set pairs. The work of the annotators was validated by in-house workers, who were paid \$2 per HIT if their work was accepted.

Instructions (click to expand/collapse)

Thanks for participating in this HIT!

Notes about worker

- In this HIT, you will be presented with 10 sets of QA pairs to annotate.
- We expect this qualifier to take about 10-20 minutes.
- Upon successful completion, you will receive a reward of \$2.00 per task.
- note! A trick example is laid out to verify your work. So, poor work will not be charged.**

Your task:

Given an image and a **question-response pair**, you are expected to evaluate whether the pair is appropriate for the situation depicted in the image focusing on a person (e.g., [PERSON#]).

- Choose how likely for the **question-response pair** is true, mainly focusing on the **[PERSON#]**
 - Plausible**: I definitely agree that the stated answer is appropriate for the given question about images.
 - Implausible**: I don't think there is an appropriate answer from the choices provided for the question.
 - note!** If you choose **implausible** 1) select the **proper response** from the other choices or 2) check whether the question match with the **given image**.

Data source is the Visual Commonsense Reasoning (VCR) dataset. Three examples are given below the instruction panel. Read through

Example 1

Question A: What is [PERSON0] is doing ?

(1) [PERSON0] is helping [PERSON1].
(2) Reading a map .
(3) Trying to kill [PERSON1] .
(4) [PERSON1] is taking a break from the party to perhaps ponder a thought .

► I think the answer (1) is **plausible**

Other answer is more plausible
 don't think there is an appropriate answer from the choices

Question B: What is [PERSON1] is doing ?

(1) [PERSON0] is helping [PERSON1].
(2) Reading a map .
(3) Trying to kill [PERSON1] .
(4) [PERSON1] is taking a break from the party to perhaps ponder a thought .

► I think the answer (2) is **plausible**

Other answer is more plausible
 don't think there is an appropriate answer from the choices

Figure 11: Instructions for VCR-CS data validation HIT

Instructions (click to expand/collapse)

Thanks for participating in this HIT!

Notes about worker

- This is EVALUATION task (not validation, which you may did).
 - If you have completed our validation hit, please do not do this. We only consider the worker that did not participate in our validation HITs. We are not responding to the completion of workers who had already participated in our validation HITs. If you are not sure, then please contact us.
- In this HIT, you will be presented with 10 sets of QA pairs to annotate.
- We expect this task to take about 10-20 minutes.
- Upon successful completion, you will receive a reward of \$2.00 per task.
- note! A trick example is laid out to verify your work. So, poor work will not be charged.**
- Leave us a comment if you have any.

Your task:

In this task, we are asking you to answer by your commonsense. **Given an image and a question**, you must choose a **correct answer** from four options focusing on a person (e.g., [PERSON#]) depicted in an image .

Our data source is the Visual Commonsense Reasoning (VCR) dataset. Three examples are given below the instruction panel. Please read through the examples!

Example 1

Question A: Does [PERSON2] feel comfortable ?

(1) Yes because the person sitting next to her is smiling.
(2) No she does not .
(3) Yes , she is wearing something with thin straps .
(4) Yes , she is cold .

► I think the answer in this question is **1**

Question B: Does [PERSON1] feel comfortable ?

(1) Yes because the person sitting next to her is smiling.
(2) No she does not .
(3) Yes , she is wearing something with thin straps .
(4) Yes , she is cold .

► I think the answer in this question is **2**

Figure 12: Instructions for VCR-CS human evaluation HIT