

Do Vision-Language Transformers Exhibit Visual Commonsense? An Empirical Study of VCR

Zhenyang Li
Shandong University
zhenyanglidz@gmail.com

Xiaolin Chen
Shandong University
cxlicd@gmail.com

Yangyang Guo
National University of Singapore
guoyang.eric@gmail.com

Liqiang Nie*
Harbin Institute of Technology,
Shenzhen
nieliqiang@gmail.com

Kejie Wang
Shandong University
kjiwang.henry@gmail.com

Mohan Kankanhalli
National University of Singapore
mohan@comp.nus.edu.sg

ABSTRACT

Visual Commonsense Reasoning (VCR) calls for explanatory reasoning behind question answering over visual scenes. To achieve this goal, a model is required to provide an acceptable rationale as the reason for the predicted answers. Progress on the benchmark dataset stems largely from the recent advancement of Vision-Language Transformers (VL Transformers). These models are first pre-trained on some generic large-scale vision-text datasets, and then the learned representations are transferred to the downstream VCR task. Despite their attractive performance, this paper posits that the VL Transformers do not exhibit visual commonsense, which is the key to VCR. In particular, our empirical results pinpoint several shortcomings of existing VL Transformers: small gains from pre-training, unexpected language bias, limited model architecture for the two inseparable sub-tasks, and neglect of the important object-tag correlation. With these findings, we tentatively suggest some future directions from the aspect of dataset, evaluation metric, and training tricks. We believe this work could make researchers revisit the intuition and goals of VCR, and thus help tackle the remaining challenges in visual reasoning.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**; *Natural language processing*; Neural networks.

KEYWORDS

Visual Commonsense Reasoning, Visual Question Answering, Vision-Language Transformer

ACM Reference Format:

Zhenyang Li, Yangyang Guo, Kejie Wang, Xiaolin Chen, Liqiang Nie, and Mohan Kankanhalli. 2023. Do Vision-Language Transformers Exhibit Visual Commonsense? An Empirical Study of VCR. In *Proceedings of the 31st ACM*

*Corresponding Author: Liqiang Nie.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MM '23, October 29–November 3, 2023, Ottawa, ON, Canada

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0108-5/23/10...\$15.00
<https://doi.org/10.1145/3581783.3612395>

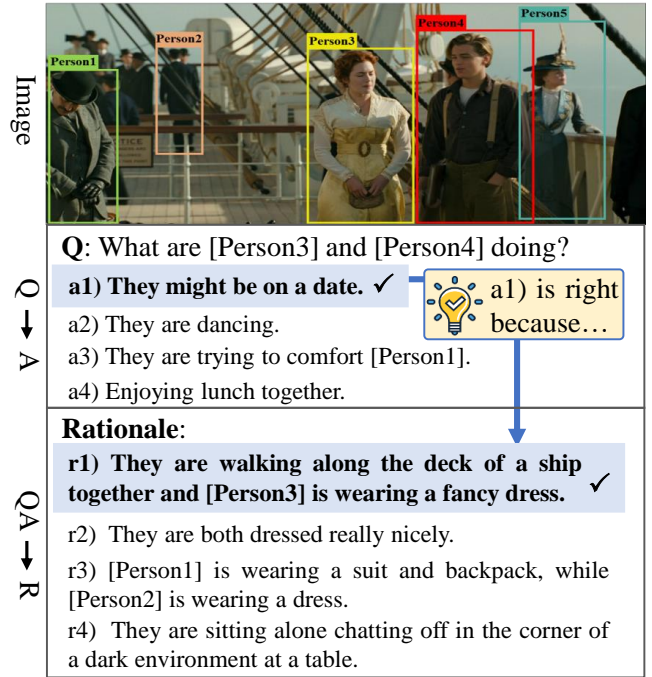


Figure 1: An exemplar of VCR. The task is composed of two sub-tasks: $Q \rightarrow A$ and $QA \rightarrow R$, where the challenge mainly lies in the cross-modal reasoning from the latter.

International Conference on Multimedia (MM '23), October 29–November 3, 2023, Ottawa, ON, Canada. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3581783.3612395>

1 INTRODUCTION

The intersection of vision and language areas has spawned numerous multi-modal tasks, such as image captioning [11, 15, 44] and Visual Question Answering (VQA) [1, 2, 8] over the past few years. Among these, Visual Commonsense Reasoning (VCR) [46] has recently drawn increasing attention from researchers due to its challenging nature. Beyond answering visual questions as conventional VQA does ($Q \rightarrow A$), VCR further requires the model to pick the rationale for the $Q \rightarrow A$ process (where the visual commonsense is), namely $QA \rightarrow R$ (see Figure 1).

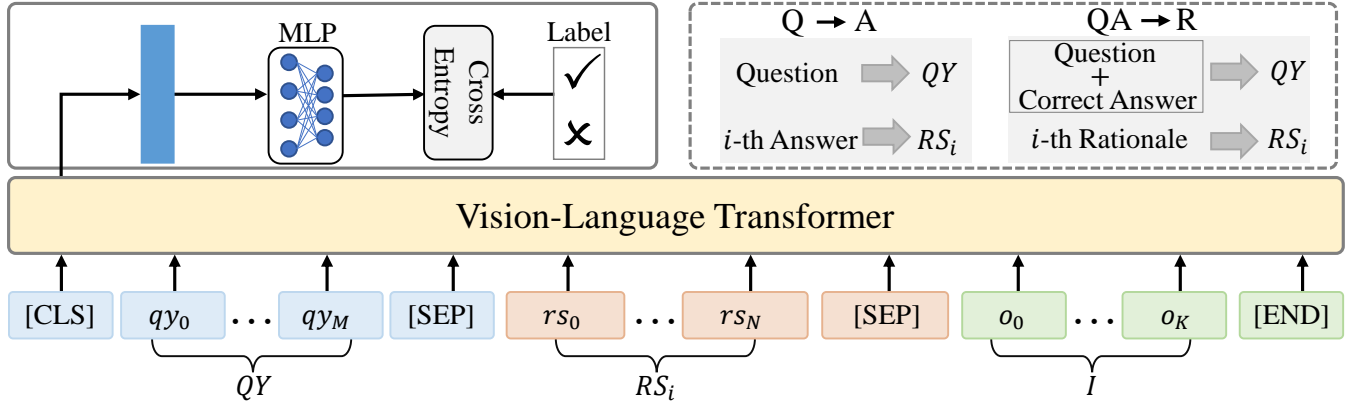


Figure 2: Pipeline of Vision-Language Transformers for VCR. $Q \rightarrow A$ and $QA \rightarrow R$ share the same pipeline where only the input query (QY) and response (RS) are slightly different.

VCR is taken as an important proxy for visual commonsense understanding. To deal with this difficult task, some initial efforts have been devoted to designing task-specific model architectures [42, 46, 49]. These models utilize the contextualized query-region affinity captured by well-designed attention mechanisms as evidence, to reason the plausibility between responses and images. Subsequently, Vision-Language Transformers (VL Transformers) [28, 34, 35] swept the multi-modal vision-language domain and rapidly prevailed over the competing counterparts on the VCR Leaderboard¹. VL Transformers first pre-train BERT style models on generic vision-language datasets (such as Conceptual Captions [32]) for task-agnostic representation learning, which is then transferred to the downstream VCR for both $Q \rightarrow A$ and $QA \rightarrow R$.

Though the state-of-the-art keeps advancing, the reasoning capability of these VL Transformers still remains debatable. As an improvement of VQA, VCR is promising not simply because the performance of traditional VQA benchmarks has saturated, but it is expected to uncover the complex reasoning behind answer prediction, *i.e.*, rationale prediction [46]. On the flip side, the key to the success of VL Transformers, namely pretext training objectives (*e.g.*, masked language modeling), deviates substantially from the reasoning goal. In particular, typical pretext tasks usually focus either on the reconstruction from partially masked elements, or the coherence between the two given modalities. However, why these modality matching-driven objectives aid visual reasoning on VCR remains less persuasive.

Given the above concern, in this work, we empirically find that VL Transformers perform well mostly in those instances requiring less reasoning while failing on difficult ones (refer to Figure 3). We then conduct an in-depth investigation into this problem and obtain the following findings:

- Limited benefits are transferred from pre-training to VCR. Pre-training on large-scale vision-language datasets enhances some downstream tasks like image retrieval with significant performance margin [4, 28]. In contrast, VCR improves little from these carefully designed pre-training steps. We attribute this finding to two reasons: 1) domain shift between

pre-training and VCR fine-tuning, and 2) weak reasoning of these pretext objectives.

- Language bias prevents the model from cross-modal reasoning. The language shortcut between textual queries and responses leads the model to make decisions based on the text modality only [45], especially for $QA \rightarrow R$. When it comes to cases that require visual reasoning, the model is misled to leverage the bias between text due to their overwhelming co-occurrence than that of image and text.
- The architecture does not lend itself to a holistic solution for both $Q \rightarrow A$ and $QA \rightarrow R$. Based on the definition and intuition of VCR, the $Q \rightarrow A$ and $QA \rightarrow R$ should be made consistent rather than being treated separately [22]. Unfortunately, existing VL Transformer models are limited in handling these two sub-tasks with consistency.
- The unique tag labels are somehow under-utilized by current VL Transformers. As demonstrated in Figure 1, the tag (such as ‘person3’) defines an exclusive link between an object label and a certain bounding box. It is essential to consider such relationships for proper reasoning rather than treating them as being independent.

This paper shows that the above drawbacks are common among some representative VL Transformers². Nevertheless, it is not our goal to propose a novel method to close the gap. Instead, the key contribution of this work lies in its insights for developing new methods, which helps bypass certain limitations. Towards this, we also outline several potential research directions given the analysis of these problems.

2 PRELIMINARY

2.1 Problem Formulation

Given a natural image and a textual question, Visual Commonsense Reasoning (VCR) aims to predict the answer to this question as well as the explanation of the answering process. Compared to VQA, the questions in VCR are made more challenging and the models are expected to provide the rationale behind the question answering.

¹<https://visualcommonsense.com/leaderboard/>.

²<https://github.com/SDLZY/VCR-VLTransformer>.

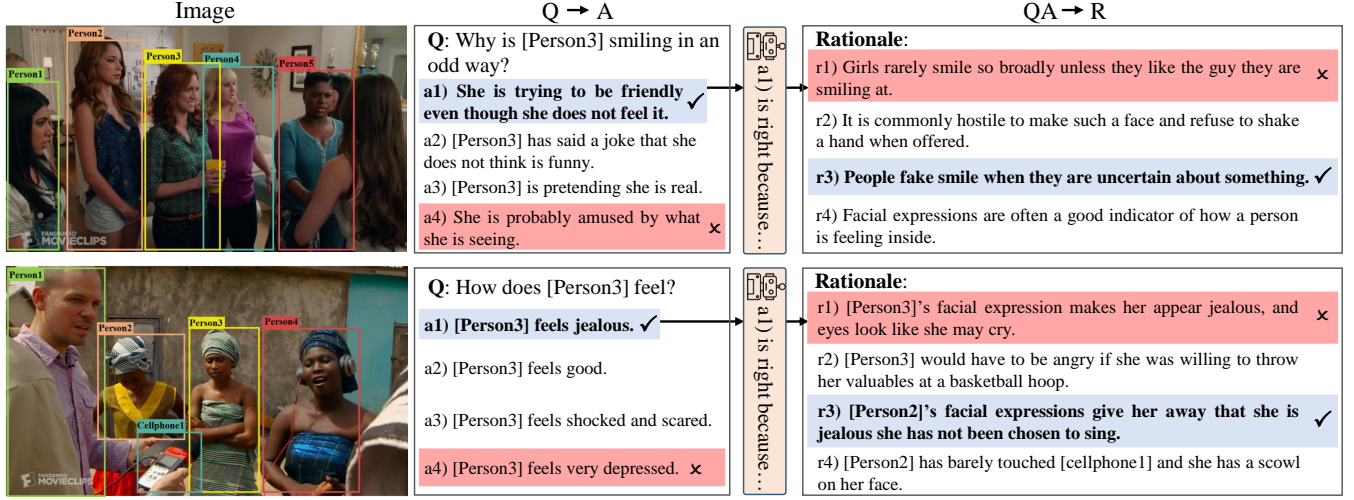


Figure 3: Failure cases from VILLA. The input to QA→R consists of the correct answer (blue one from Q→A), rather than the predicted answer (red one from Q→A) following the default setting. It can be seen that the model makes mistakes on cases calling for fine-grained reasoning.

Accordingly, VCR focuses mainly on higher-order cognition and commonsense understanding of images. Specifically, a typical VCR model can be formulated as follows,

$$RS = \operatorname{argmax}_{RS_i \in RS} f(I, QY, RS_i | \Theta), \quad (1)$$

where I and QY are the given image and query, respectively; RS denotes the response set where RS_i is the i -th element, Θ denotes the involved optimized parameters, and the function f predicts a compatible score based on the given inputs. In practice, VCR is often decomposed into the following two multiple-choice sub-tasks:

Question answering (Q→A) – For the given image I and a corresponding question $Q \leftarrow QY$, the model is required to choose the right answer $A \leftarrow RS$ from a set of answer choices $\mathcal{A} \leftarrow RS$.

Answer justification (QA→R) – Similar to the inputs of Q→A, the model in this stage takes the correct answer A as the additional input ($Q + A \leftarrow QY$), and is expected to select the right rationale $R \leftarrow RS$ from a set of rationale choices $R \leftarrow RS$.

2.2 VL Transformer Pre-training

The past few years have witnessed the rapid development of VL Transformers. In addition to the large-scale datasets, the pretext pre-training tasks or objectives are the key to the success of these models. We summarize three typical widely applied tasks, *i.e.*, Cross-modal Masked Language Modeling (MLM), Masked Region Classification (MRC), and Image-Text Matching (ITM).

MLM originates from the MLM task in BERT [7]. The key difference is that the visual clues are incorporated in VL Transformers for capturing the dependencies between linguistic and visual contents,

$$\mathcal{L}_{MLM} = -\mathbb{E}_{(T,I) \in D} \log P_{\theta}(t_m | T_m, I), \quad (2)$$

where θ represents the parameters of the VL Transformer, t_m and T_m denote the masked and the remaining tokens, respectively. Each pair $(T, I) \in D$ is composed of a text T and an image I sampled from a vision-language dataset D .

MRC is a dual task of MLM. It learns to predict the semantic class of each masked object based on the corresponding text and its surrounding visual objects. To pre-train this, the cross-entropy loss (CE) between the output distribution normalized by a softmax function $s(i'_n)$ and class label $c(i_n)$ for the masked region i_n is employed,

$$\mathcal{L}_{MRC} = \mathbb{E}_{(T,I) \in D} \sum_{n=1}^N CE(s(i'_n), c(i_n)), \quad (3)$$

where $s(i'_n)$ denotes the VL Transformer output of i_n and N is the number of the objects detected from image I .

ITM is similar to the Next Sentence Prediction task utilized in BERT [7]. Given an image-text pair as input, the Transformer must predict whether the image and text are aligned, *e.g.*, whether the text describes the image,

$$\mathcal{L}_{ITM} = -\mathbb{E}_{(T,I) \in D} [y \log s_{\theta}(T, I) + (1 - y) \log(1 - s_{\theta}(T, I))], \quad (4)$$

where y is the ground truth and s_{θ} is the score function to measure the alignment probability of (T, I) .

2.3 Fine-tuning on VCR

Input Formats – Figure 2 illustrates the input format of typical VL Transformers in VCR. Like in other VL tasks, the inputs are composed of a special classification token [CLS], the corresponding text, a separation token [SEP] between the two modalities, the given image, and an end token [END]. Specifically, pertaining to the textual input, the concatenation of Q and A_i with a [SEP] is applied for the Q→A sub-task; while for QA→R, the usual way is to concatenate Q , ground-truth answer A and the candidate rationale R_i .

Optimization – For fine-tuning on VCR, the final output of the [CLS] token is utilized to predict whether the given answer or rationale is the correct choice. The VL Transformer is trained in

an end-to-end fashion by minimizing the multi-class cross-entropy loss between the prediction for each response and the ground truth label. During inference, models are also comprehensively evaluated with the classification accuracy on the two sub-tasks, namely the holistic $Q \rightarrow AR$ task (the accuracy set intersection of $Q \rightarrow A$ and $QA \rightarrow R$).

Note that there are often two separable models for the two sub-tasks. The logical connection between these two is surprisingly ignored in the existing literature.

3 EXPERIMENTAL RESULTS AND FINDINGS

Our findings are based on four popular SOTA VL-Transformers - UNITER [4], ViLBERT [28], VL-BERT [35] and VILLA [12]. We employed these four representative models for two reasons: 1) they cover a wide variety of different pre-training datasets, objectives, and architectures, and 2) the models are evaluated on VCR and the released codes can be rerun for reproduction. After examining several failure cases in Figure 3, we found that these models often made mistakes on challenging queries. As the pretext tasks used by VL Transformers are mostly matching goal-driven rather than reasoning-oriented, we argue that they fail to do visual reasoning over commonsense scenes, which is the key to VCR. As a result, the superior performance is somewhat misleading. To explore this problem, we conducted extensive experiments and summarize some key findings below (Note that some experimental results have been moved to supplementary material due to space limitation.). This section elaborates on the following research questions:

- **RQ1** – How much does the pre-training help the downstream VCR task?
- **RQ2** – Is the image modality really helpful for ‘visual reasoning’ in VCR?
- **RQ3** – Are there any correlations between the two separately trained models for the two sub-tasks?
- **RQ4** – Is it desirable to ignore the distinct connection between tags and objects?

3.1 Limited Gain from Pre-training

Pre-training is of vital importance to downstream VL tasks, wherein the pretext training objectives play an important part. These pretext objectives, as detailed in Section 2.2, focus mainly on the conventional recognition ability. Nevertheless, a good VCR model should be made cognition-aware and capable of visual reasoning. In light of this, the contribution of pre-training to VCR remains somewhat less trustworthy. As shown in Table 1, 2 and 3, the gains from pre-training for these matching-oriented tasks, *e.g.*, cross-modal retrieval [48], are quite substantial (more than 10 points in Table 2). By contrast, when it comes to VCR, the performance gain is limited to 0 ~ 2 points. It illustrates that current VL Transformers benefit much more from the architecture itself than the pre-training objectives.

There are two possible explanations for this. First, these pretext objectives are specially designed for simple modality reconstruction and cross-modal matching. However, VCR demands fine-grained visual reasoning which cannot be achieved by the above means. Second, the domain shift between pre-training and VCR fine-tuning makes the transfer difficult. The images in the VCR dataset are

about movie plots which are distinctive from the pre-training image captioning datasets [32].

We also investigated the convergence of three VL Transformers and display the results in Figure 4. It shows that pre-training boosts models with a good initialization point. With more training steps, the advantage from pre-training decays until the models with and without pre-training reach a similar level of performance.

3.2 Language Bias

It is well-known that VQA has been long affected by the language bias problem [13, 17]. It refers to the correlation shortcut between textual questions and answers. To study whether VCR is afflicted with this problem, we tested two settings: VL Transformers respectively removing query and image (see Figure 2), and observed the results. Table 4 shows that for $Q \rightarrow A$, the model variants without query and without image lead to similar performance degradation. However, for the sub-task of $QA \rightarrow R$, there exists a significant performance gap between these two variants, *i.e.*, the difference between the former variant and the full model is over 2× more than that of the latter. It is mainly because of the language shortcut between textual inputs (One such example can be seen in Figure 1, that only the correct r1 contains the [Person3] tag.). VL Transformers tend to utilize such bias for prediction rather than performing visual reasoning.

To further validate this hypothesis, we then explored the attention weight distribution of these models. In general, the attention weights express how much other elements contribute to the learning of the current element. We first estimated the learned attention weights from UNITER [4] and VILLA [12] based upon the [CLS] token, as the output from [CLS] is leveraged for predicting the correct response. It can be observed in Figure 5 that the two models pay too much attention to the textual rationale and answer elements (except for the [CLS] token itself in the second layer) while focusing less on the visual objects. Since the goal of VCR is to pursue reasoning with images, models making decisions without the involvement of vision seem less desirable.

Thereafter, we further evaluated the attention weights from one modality to the other, and illustrate the results in Figure 6. We can observe that the language modality sees almost only itself, especially for the $QA \rightarrow R$ sub-task. It further validates that the language bias dominates the prediction of both sub-tasks, and $QA \rightarrow R$ is affected more. By contrast, the vision modality looks more balanced with respect to its attention distribution. This modality bias issue is further reflected by two examples in Figure 7, where it can be seen that each modality mainly focuses on its own encoded tokens.

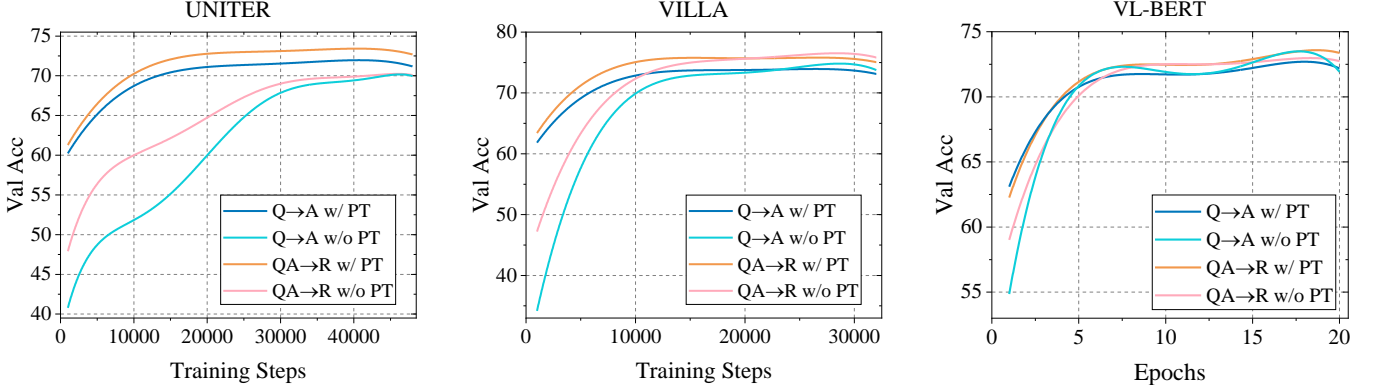
3.3 Sparse Correlation between Two Models

Figure 2 shows that VL Transformers take the $Q \rightarrow A$ and $QA \rightarrow R$ as two independent processes. In other words, there are two separate models with similar architectures and training protocols. However, separately treating $Q \rightarrow A$ and $QA \rightarrow R$ deteriorates the visual scene understanding, considering that these two processes share a common goal [21, 22]. In the following, we investigated the correlation between these two models from two angles.

One intuitive idea is to study the overlapped instances between $Q \rightarrow A$ and $QA \rightarrow R$, see Figure 9. Note that the query to $QA \rightarrow R$

Table 1: Pre-training gains from VILLA on five cross-modality tasks.

Pre-train	NLVR ²		Retrieval		VQA	VCR Validation		
	dev	test-P	Text	Image		Q→A	QA→R	Q→AR
×	50.9	51.2	80.5	65.4	68.4	72.6	75.1	54.7
✓	78.4	79.3	86.6	74.7	73.6	73.9	76.1	56.5
Δ	27.5	28.1	6.1	9.3	5.2	+1.3	+1.0	+1.8

**Figure 4: Convergence analysis of three VL Transformers with and without pre-training.****Table 2: Pre-training gains from UNITER on three tasks.**

Pre-train	Retrieval		VCR Validation		
	Text	Image	Q→A	QA→R	Q→AR
×	83.3	73.9	71.5	72.9	52.2
✓	94.3	85.8	72.7	74.5	54.4
Δ	11.0	11.9	+1.2	+1.6	+2.2

Table 3: Pre-training gain from Vil-BERT.

Pre-train	Image Retrieval		VCR Validation		
	R@1	R@5	Q→A	QA→R	Q→AR
×	45.5	76.8	69.3	71.0	49.5
✓	58.2	84.9	72.4	74.5	54.0
Δ	12.7	8.1	+3.1	+3.5	+4.5

is the given question concatenated with the right answer. Given the same question, on the one hand, we observed that around 3/4 of correctly predicted answers lead to the right rationale (from round box A to round box R). This shows that some answers are at least not predicted based on the same reasoning as humans. On the other hand, we found that a large proportion of wrongly predicted answers (3/4) correspond to the right rationale (from the square box A to round box R). We suspect that one possible reason is due to the shortcut learning between correct answers and rationales, as

Table 4: Performance of three VL Transformers within three variants: full model, without query and without image. Note that Q→AR tells the interaction between correctly predicted Q→A and QA→R instances.

Model	Q→A	QA→R	Q→AR
UNITER	74.4	76.9	57.5
w/o query	59.3 (-15.1)	57.2 (-19.7)	34.5 (-23.0)
w/o image	59.6 (-14.8)	68.6 (-8.3)	41.0 (-16.5)
VL-BERT	72.6	74.0	54.0
w/o query	56.4 (-16.2)	53.5 (-20.5)	30.8 (-23.2)
w/o image	58.8 (-13.8)	66.0 (-8.0)	38.9 (-15.1)
VILLA	75.4	78.7	59.5
w/o query	60.6 (-14.8)	58.8 (-19.9)	36.2 (-23.3)
w/o image	60.5 (-14.9)	71.0 (-7.7)	43.1 (-16.4)

discussed in Section 3.2 because the input to QA→R contains the ground-truth answer instead of the predicted one.

As discussed before, existing methods all employ two models to separately tackle Q→A and QA→R. This makes us wonder, are there any differences between these two models? Or does the separation of the two sub-tasks really allow the two models to better deal with answering and reasoning? To answer this question, we conducted tests with the same model for both sub-tasks. Figure 8 illustrates that when employing the same model for these two, the model performance is slightly increased. We speculate that the cause of this phenomenon is the doubling of the dataset employed for training. This experiment, on the other hand, shows that VL

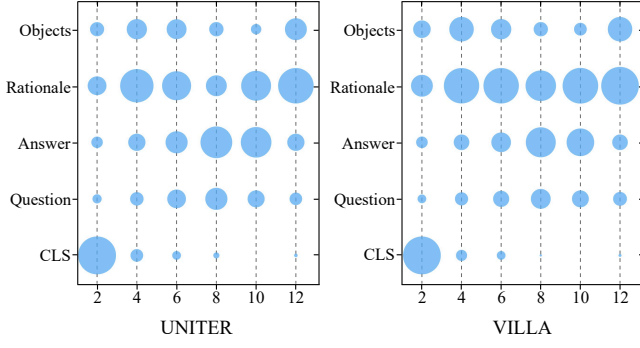


Figure 5: Attention distribution from the token of [CLS]. We empirically selected even layers for demonstration.

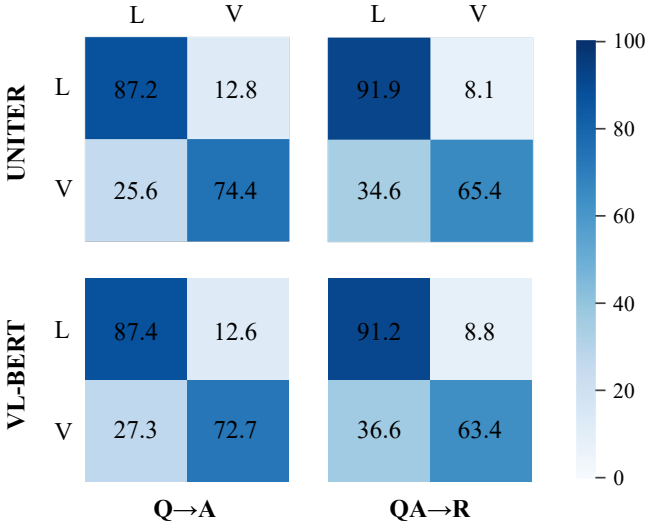
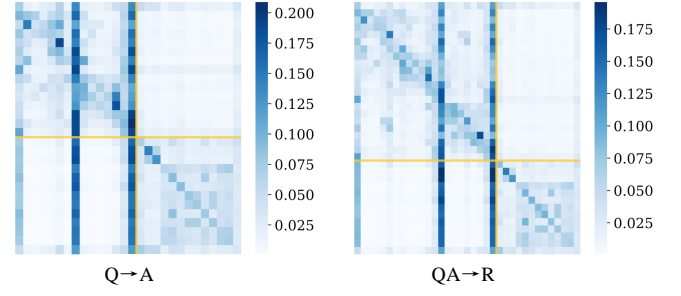


Figure 6: The attention weight distribution from the modality of each row to the modalities of columns.

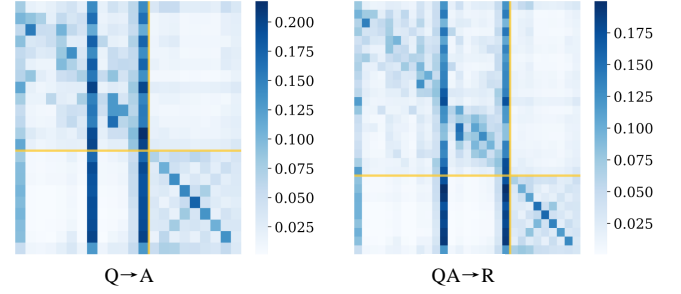
Table 5: Performance of three VL Transformers within three variants: full model, without tags, and with the random replacement of tags.

Model	Q→A	QA→R	Q→AR
UNITER	74.4	76.9	57.5
w/o tag	68.6 (-5.8)	72.8 (-4.1)	50.5 (-7.0)
repl. tag	73.5 (-0.9)	76.3 (-0.6)	56.4 (-1.1)
VL-BERT	72.6	74.0	54.0
w/o tag	65.7 (-6.9)	69.5 (-4.5)	46.1 (-7.9)
repl. tag	72.5 (-0.1)	73.9 (-0.1)	53.9 (-0.1)
VILLA	75.4	78.7	59.5
w/o tag	69.7 (-5.7)	74.8 (-3.9)	52.5 (-7.0)
repl. tag	74.5 (-0.9)	78.2 (-0.5)	58.5 (-1.0)

Transformers do not differentiate these two sub-tasks, despite the fact that the latter one requires more visual commonsense.



(a) Sample obtained from the validation set with an instance ID of 5431.



(b) Sample obtained from the validation set with an instance ID of 9686.

Figure 7: Qualitative results of the attention map from the last self-attention layer in VL-BERT. Each row represents the attention weight of a given input token with respect to all tokens. We use the yellow lines to separate textual tokens from visual tokens.

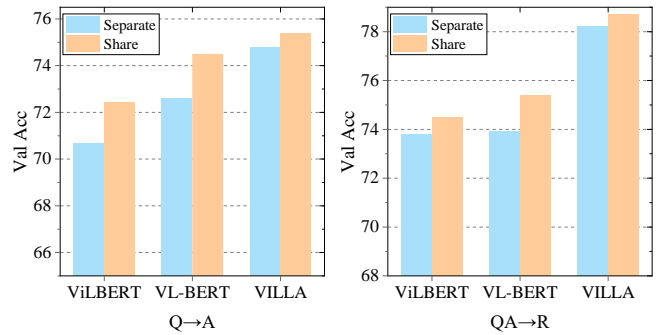



Figure 8: Performance comparison of three models with sharing and separate parameters.

3.4 Incompleteness of Tag Handling

In VCR, the questions, answers and rationales are written in a mixture of rich natural language as well as detection tags, like ‘[person1 

To investigate the importance of these tag labels, we first removed the tag input and observed the results in Table 5. It can be

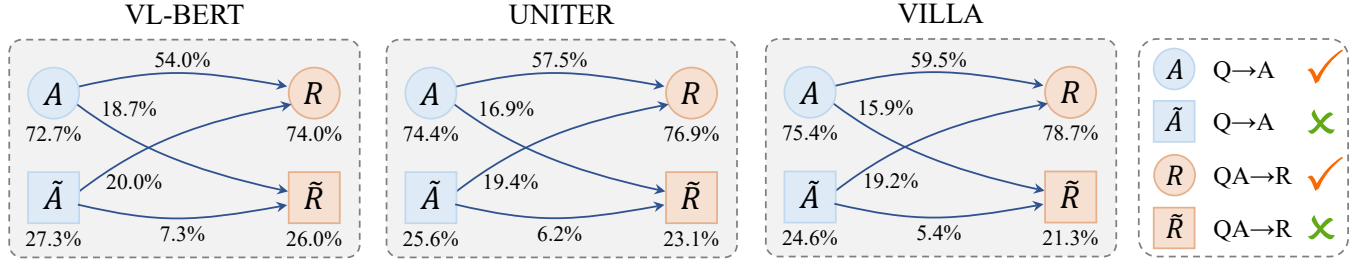


Figure 9: Proportion of correctly predicted instances from three VL Transformers and their intersection of four states in VCR.

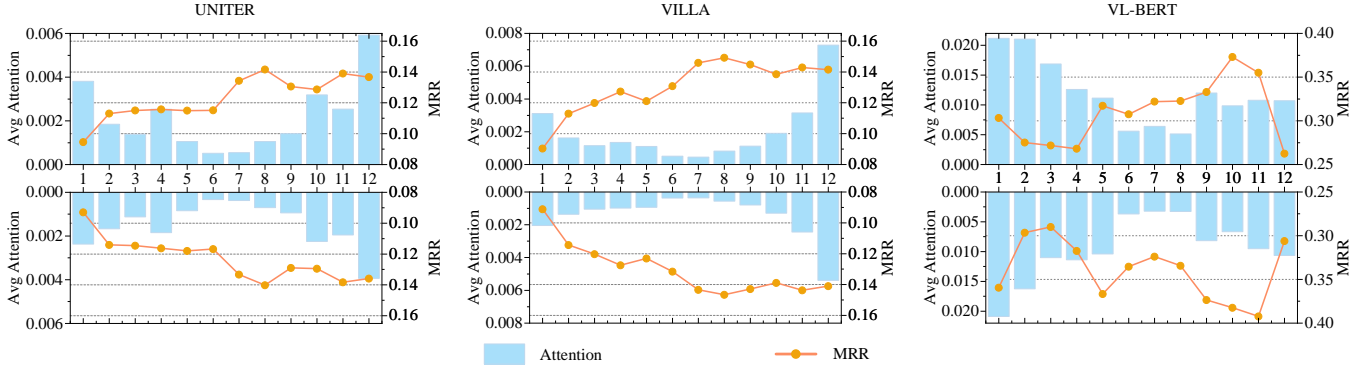


Figure 10: Average attention weights and MRR value with respect to each tag for different VL Transformer layers (1 to 12). We use three typical VL Transformers for demonstration. Top: Q→A; Bottom: QA→R.

seen that the model performance degrades to some extent. The key reason is that the input sentences lack important subjects, and therefore makes VL Transformers confusing. We then randomly replaced each tag with another one, *e.g.*, ‘[table]’ ⇒ ‘[person]’. The results in Table 5 demonstrate that the models show only minor deterioration. This illustrates that existing handling for tags is largely limited. That is, randomly replacing tags barely impacts the performance though the link between each tag and object is deliberately broken.

Like in Section 3.2, we also studied the attention distribution, especially the attention weights attached to each given tag. In particular, we used two metrics to quantify this effect:

Avg Attention is adopted to count the averaged attention values between the given tag and the visual object to which it referred (upper bound is 1.0),

$$att_l = \frac{1}{\sum_{i=1}^n m_i} \sum_{i=1}^n \sum_{j=1}^{m_i} score(t_{ij}, o_{ij}), \quad (5)$$

where $score(t_{ij}, o_{ij})$ denotes the attention score from the tag token t_{ij} to its corresponding visual object o_{ij} in layer l , n and m_i are the number of samples in the validation set and the number of tags contained in each sample, respectively.

MRR is employed to estimate the rank of the true object based on the predicted attention values for each tag (upper bound is 1.0),

$$mrr_l = \frac{1}{\sum_{i=1}^n m_i} \sum_{i=1}^n \sum_{j=1}^{m_i} \frac{1}{rank(t_{ij}, o_{ij})}, \quad (6)$$

where $rank(t_{ij}, o_{ij})$ is the ranking of the $score(t_{ij}, o_{ij})$ among the attention scores between t_{ij} and all objects (5 to 13 per image). As shown in Figure 10, for all three VL Transformers, both values are pretty low in relation to their upper bound (1.0). This result demonstrates that the link between the tag and its attached object is almost nonexistent. As a result, whether the VL Transformers perform reasoning remains doubtful since such an important correlation is ignored.

4 WHAT TO DO NEXT?

The above findings tell us that existing VL Transformers do not offer a good solution for visual commonsense understanding. As good as the recognition capability is, a VCR model is expected to be endowed with more reasoning strengths. In what follows, we outline several possible directions worth exploring in the next:

Dataset – Curating more challenging datasets which can circumvent the shortcut modeling problem encountered by existing VL Transformers. In addition, probing other means of visual commonsense understanding instead of giving explanations to question answering is also of great potential, *e.g.*, the collection of spatial or attribute commonsense.

Evaluation Metric – Designing specific metrics for quantifying the reasoning capability of models, for which the current vanilla accuracy metric is limited in its form. Besides time-intensive subjective evaluations, we can also gain insights from the text generation tasks like machine translation [6] and image captioning [39].

Pre-training Task – Making the pre-training objectives focus more on cognition and understanding. In particular, how to enhance VL Transformers with reasoning strengths is an interesting path for improving downstream VCR. One possible attempt is to incorporate large-scale knowledge into pre-training pretext tasks.

De-biasing – Guiding models with suitable de-biasing tricks can help overcome the language bias problem. It is worth noting that the bias problem is more challenging than VQA because of the spurious correlation between a single query and response. In contrast to VCR, the bias in the sister VQA domain results from the statistical shortcut between question type and answers³, whereby some trivial loss re-balancing tricks can be employed [14].

Model Architecture – Developing suitable model architectures to take advantage of the unique tag-object label and the two-step visual commonsense understanding. For instance, we can approach $Q \rightarrow A$ and $QA \rightarrow R$ simultaneously with collaboration, where the common image information provides an essential proxy to achieve this goal.

Prompting on Large Language Models – One last promising future direction is to prompt the pre-trained large language models, such as ChatGPT⁴. A typical approach is to decode the image into text, and then the generalization capability of these large models can be leveraged to provide the correct explanation for visual questions. Nevertheless, the captioning quality of another proxy model is still under questioning.

5 RELATED WORK

5.1 Vision-Language Transformers

The success of Transformers in the field of Natural Language Processing (NLP) [7] and Computer Vision (CV) [3, 9, 18, 43] brings a lot of progress for multi-modal vision and language tasks [27, 29, 33, 38, 51]. Based on how the vision and language branches are fused, current VL Transformers can be roughly categorized into single-stream (e.g., UNIMO [20] and SOHO [16]) and dual-stream cross-modal Transformers (e.g., LXMERT [37] and ALBEF [19]). A typical VL Transformer model often employs the *pretrain-then-finetune* learning schema: the model is first pre-trained on large vision-text datasets and then fine-tuned on downstream tasks by transferring their rich representations from pre-training. Specifically, the pretext tasks play a vital role in pre-training, where masked language modeling, masked region prediction, and image-text matching are extensively studied. The fine-tuning step mirrors that of the BERT model [7], which includes a task-specific input, output, and objective. The pre-trained model is thereafter optimized to maximize the performance on the corresponding vision-and-language task. The VL Transformers mainly help in the following three groups of downstream tasks: cross-modal matching, cross-modal reasoning, and vision language generation [10]. The first group focuses on learning cross-modal correspondences between vision and language, such as image text retrieval and visual referring expression. Reasoning ones require VL Transformers to perform language reasoning based on visual scenes, such as VQA. The last group aims to generate the targets of one modality given the other as input [5, 37]. The

desired visual or textual tokens are decoded in an auto-regressive generation manner.

5.2 Visual Commonsense Reasoning

Recently, multimodal tasks have garnered increasing research attention [23–26, 30, 31, 36, 40, 41], with a notable example being Visual Question Answering (VQA). However, conventional VQA models often face limitations due to their black-box reasoning capabilities. To move one step further, VCR was presented to supplement VQA by inferring the rationale behind the question-answering process. Some initial work designs specific architectures to address VCR for the purpose of finer-grained visual understanding. For instance, R2C [46] performs three inference steps - grounding, contextualization, and reasoning, to move towards cognition-level understanding step by step. Inspired by the neuron connectivity of the human brain, CCN [42] designs a connective cognition network to globally and dynamically integrate the local visual neuron connectivity. For explicit cross-modal representation learning, syntactic information is incorporated into the visual reasoning and natural language understanding [50]. Recent progress on the VCR leaderboard is mostly derived from VL Transformers. For example, UNITER [4] utilizes a single-stream encoder and four elaborate pre-training tasks to learn universal image-text representations for various downstream multi-modal tasks. ViLBERT [28] employs a dual-stream fusion encoder with co-attention layers to model the inter-modality interactions. MERLOT [47] first learns commonsense representations of multi-modal events by pre-training over millions of videos and then transfers them to the target images in the VCR dataset. Despite the impressive performance achieved by these VL Transformer models, whether they truly possess visual commonsense remains an intriguing yet under-explored question.

6 CONCLUSION

This paper recognizes several limitations of utilizing existing VL Transformers to VCR, which takes an essential step towards visual commonsense understanding. Though the results on the benchmark dataset are impressive, we argue that these numbers are less trustworthy in terms of visual reasoning – the key ingredient of VCR. On the flip side, this ‘all in Transformer’ wave may mislead the VCR research in the wrong direction, preventing the community from pursuing models that are reasoning-aware. To the best of our knowledge, we are the first to comprehensively study this problem in literature though we strongly believe we are not the only one afflicted with it. With the discovery of this problem and the tentative proposal for several future directions, we hope this work can help motivate more interesting and plausible ideas for visual commonsense reasoning in the future.

ACKNOWLEDGMENTS

This research is supported by the National Research Foundation, Singapore under its Strategic Capability Research Centres Funding Initiative, and the National Natural Science Foundation of China, No.:U1936203. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

³The question type is often referred to the first few words of the given question, e.g., how many. Answers in VQA datasets are mostly composed of a few keywords.

⁴<https://chat.openai.com/>

REFERENCES

- [1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *IEEE Conference on Computer Vision and Pattern Recognition*. 6077–6086.
- [2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *IEEE International Conference on Computer Vision*. 2425–2433.
- [3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-End Object Detection with Transformers. In *European Conference on Computer Vision*, Vol. 12346. 213–229.
- [4] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: UNiversal Image-Text Representation Learning. In *European Conference on Computer Vision*. 104–120.
- [5] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying Vision-and-Language Tasks via Text Generation. In *International Conference on Machine Learning*. 1931–1942.
- [6] Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2021. A Survey of Multilingual Neural Machine Translation. *Comput. Surveys* 53, 5 (2021), 99:1–99:38.
- [7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *North American Chapter of the Association for Computational Linguistics*. 4171–4186.
- [8] Yang Ding, Jing Yu, Bang Liu, Yue Hu, Mingxin Cui, and Qi Wu. 2022. MuKEA: Multimodal Knowledge Extraction and Accumulation for Knowledge-based Visual Question Answering. In *IEEE Conference on Computer Vision and Pattern Recognition*. 5079–5088.
- [9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiuhua 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*. 1–21.
- [10] Yifan Du, Zikang Liu, Junyi Li, and Wayne Xin Zhao. 2022. A Survey of Vision-Language Pre-Trained Models. In *International Joint Conference on Artificial Intelligence*. 5436–5443.
- [11] Ali Farhadi, Seyyed Mohammad Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David A. Forsyth. 2010. Every Picture Tells a Story: Generating Sentences from Images. In *European Conference on Computer Vision*. 15–29.
- [12] 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. In *Advances in Neural Information Processing Systems*.
- [13] Yangyang Guo, Liqiang Nie, Harry Cheng, Zhiyong Cheng, Mohan S. Kankanhalli, and Alberto Del Bimbo. 2023. On Modality Bias Recognition and Reduction. *ACM Transactions on Multimedia Computing, Communications, and Applications* 19, 3 (2023), 103:1–103:22.
- [14] Yangyang Guo, Liqiang Nie, Zhiyong Cheng, Feng Ji, Ji Zhang, and Alberto Del Bimbo. 2021. AdaVQA: Overcoming Language Priors with Adapted Margin Cosine Loss. In *International Joint Conference on Artificial Intelligence*. 708–714.
- [15] Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. 2022. Scaling Up Vision-Language Pre-training for Image Captioning. In *IEEE Conference on Computer Vision and Pattern Recognition*. 17980–17989.
- [16] Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. 2021. Seeing Out of the Box: End-to-End Pre-Training for Vision-Language Representation Learning. In *IEEE Conference on Computer Vision and Pattern Recognition*. 12976–12985.
- [17] Chenchen Jing, Yuwei Wu, Xiaoxun Zhang, Yunde Jia, and Qi Wu. 2020. Overcoming Language Priors in VQA via Decomposed Linguistic Representations. In *AAAI Conference on Artificial Intelligence*. 11181–11188.
- [18] Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. In *International Conference on Machine Learning*. 5583–5594.
- [19] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Gotmare, Shafiq R. Joty, Caiming Xiong, and Steven Chu-Hong Hoi. 2021. Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. In *Advances in Neural Information Processing Systems*. 9694–9705.
- [20] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. 2021. UNIMO: Towards Unified-Modal Understanding and Generation via Cross-Modal Contrastive Learning. In *Annual Meeting of the Association for Computational Linguistics*. 2592–2607.
- [21] Zhenyang Li, Yangyang Guo, Kejie Wang, Fan Liu, Liqiang Nie, and Mohan Kankanhalli. 2023. Learning to Agree on Vision Attention for Visual Commonsense Reasoning. *Transactions on Multimedia* (2023), 1–11.
- [22] Zhenyang Li, Yangyang Guo, Kejie Wang, Yinwei Wei, Liqiang Nie, and Mohan S. Kankanhalli. 2023. Joint Answering and Explanation for Visual Commonsense Reasoning. *IEEE Transactions on Image Processing* (2023), 3836–3846.
- [23] Fan Liu, Huilin Chen, Zhiyong Cheng, Anan Liu, Liqiang Nie, and Mohan Kankanhalli. 2022. Disentangled Multimodal Representation Learning for Recommendation. *Transactions on Multimedia* (2022), 1–11.
- [24] Fan Liu, Zhiyong Cheng, Changchang Sun, Yinglong Wang, Liqiang Nie, and Mohan Kankanhalli. 2019. User Diverse Preference Modeling by Multimodal Attentive Metric Learning. In *International Conference on Multimedia*. 1526–1534.
- [25] Fan Liu, Zhiyong Cheng, Lei Zhu, Zan Gao, and Liqiang Nie. 2021. Interest-Aware Message-Passing GCN for Recommendation. In *International World Wide Web Conferences*. 1296–1305.
- [26] Meng Liu, Liqiang Nie, Yunxiao Wang, Meng Wang, and Yong Rui. 2023. A Survey on Video Moment Localization. *Comput. Surveys* 55, 9 (2023), 188:1–188:37.
- [27] Zhenguang Liu, Runyang Feng, Haoming Chen, Shuang Wu, Yixing Gao, Yunjun Gao, and Xiang Wang. 2022. Temporal Feature Alignment and Mutual Information Maximization for Video-Based Human Pose Estimation. In *Computer Vision and Pattern Recognition*. 10996–11006.
- [28] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *Advances in Neural Information Processing Systems*. 13–23.
- [29] Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asselmann. 2021. Video Transformer Network. In *IEEE International Conference on Computer Vision Workshops*. 3156–3165.
- [30] Liqiang Nie, Leigang Qu, Dai Meng, Min Zhang, Qi Tian, and Alberto Del Bimbo. 2022. Search-oriented Micro-video Captioning. In *International Conference on Multimedia*. 3234–3243.
- [31] Leigang Qu, Meng Liu, Jianlong Wu, Zan Gao, and Liqiang Nie. 2021. Dynamic Modality Interaction Modeling for Image-Text Retrieval. In *International Conference on Research and Development in Information Retrieval*. 1104–1113.
- [32] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In *Annual Meeting of the Association for Computational Linguistics*. 2556–2565.
- [33] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. FLAVA: A Foundational Language And Vision Alignment Model. In *IEEE Conference on Computer Vision and Pattern Recognition*. 15617–15629.
- [34] Dandan Song, Siyi Ma, Zhanchen Sun, Sicheng Yang, and Lejian Liao. 2021. KVL-BERT: Knowledge Enhanced Visual-and-Linguistic BERT for Visual Commonsense Reasoning. *Knowledge-Based Systems* (2021), 107408.
- [35] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. VL-BERT: Pre-training of Generic Visual-Linguistic Representations. In *International Conference on Learning Representations*. 1–16.
- [36] Teng Sun, Wenjie Wang, Liqiang Jing, Yiran Cui, Xueming Song, and Liqiang Nie. 2022. Counterfactual Reasoning for Out-of-distribution Multimodal Sentiment Analysis. In *International Conference on Multimedia*. 15–23.
- [37] Hao Tan and Mohit Bansal. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *Empirical Methods in Natural Language Processing*. 5099–5110.
- [38] Junfeng Tu, Xueliang Liu, Zongxiang Lin, Richang Hong, and Meng Wang. 2022. Differentiable Cross-modal Hashing via Multimodal Transformers. In *International Conference on Multimedia*. 453–461.
- [39] Sijin Wang, Ziwei Yao, Ruiping Wang, Zhongqin Wu, and Xilin Chen. 2021. Faier: Fidelity and adequacy ensured image caption evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition*. 14050–14059.
- [40] Yunxiao Wang, Meng Liu, Yinwei Wei, Zhiyong Cheng, Yinglong Wang, and Liqiang Nie. 2022. Siamese Alignment Network for Weakly Supervised Video Moment Retrieval. *IEEE Transactions on Multimedia* (2022), 1–11.
- [41] Yinwei Wei, Xiang Wang, Weili Guan, Liqiang Nie, Zhouchen Lin, and Baoguan Chen. 2019. Neural multimodal cooperative learning toward micro-video understanding. *IEEE Transactions on Image Processing* (2019), 1–14.
- [42] Aming Wu, Linchao Zhu, Yahong Han, and Yi Yang. 2019. Connective Cognition Network for Directional Visual Commonsense Reasoning. In *Advances in Neural Information Processing Systems*. 5670–5680.
- [43] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. 2021. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. In *Advances in Neural Information Processing Systems*. 12077–12090.
- [44] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *International Conference on Machine Learning*. 2048–2057.
- [45] Keren Ye and Adriana Kovashka. 2021. A Case Study of the Shortcut Effects in Visual Commonsense Reasoning. In *AAAI Conference on Artificial Intelligence*. 3181–3189.
- [46] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From Recognition to Cognition: Visual Commonsense Reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition*. 6720–6731.
- [47] 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. In *Advances in Neural Information Processing Systems*. 23634–23651.

- [48] Feifei Zhang, Mingliang Xu, and Changsheng Xu. 2022. Geometry Sensitive Cross-Modal Reasoning for Composed Query Based Image Retrieval. *IEEE Transactions on Image Processing* (2022), 1000–1011.
- [49] Xi Zhang, Feifei Zhang, and Changsheng Xu. 2021. Multi-Level Counterfactual Contrast for Visual Commonsense Reasoning. In *International Conference on Multimedia*. 1793–1802.
- [50] Xi Zhang, Feifei Zhang, and Changsheng Xu. 2022. Explicit Cross-Modal Representation Learning for Visual Commonsense Reasoning. *IEEE Transactions on Multimedia* (2022), 2986–2997.
- [51] Jian Zhu and Hanli Wang. 2022. Multiscale Conditional Relationship Graph Network for Referring Relationships in Images. *IEEE Transactions on Cognitive and Development Systems* 14, 2 (2022), 752–760.

A SUPPLEMENTARY MATERIAL

This supplementary material contains the following two aspects:

- More details regarding applying the VL Transformers to the VCR task.
- More experimental results from other VL Transformers.

A.1 Further Details of VL Transformer Application in VCR

A.1.1 More Pre-training Tasks. In Section 2.2 of the main manuscript, we have introduced three mostly used pre-training tasks (*i.e.*, MLM, MRC and ITM). In what follows, we provide more pretext tasks for better understanding.

Masked Region Feature Regression (MRFR) is similar to MRC in that it is also trained to reconstruct the masked image region given the remaining regions and all texts. The difference is that MRFR learns to regress the VL Transformer output i'_n of each masked region to its visual features i_n :

$$\mathcal{L}_{MRFR} = \mathbb{E}_{(T,I) \in D} \sum_{i=1}^N \|FC(i'_n) - i_n\|^2, \quad (7)$$

where FC denotes the fully connected layer. Each pair $(T, I) \in D$ is composed of a text T and an image I sampled from a vision-language dataset D .

Cross-Model Contrastive Learning (CMCL) aims to pull matched image and text closer while pushing unmatched pairs far away. The image-to-text CMCL loss is given as:

$$\mathcal{L}_{CMCL_{i2t}} = -\mathbb{E}_{(T,I) \in D} \log \frac{\exp(i_n^T t_n / \sigma)}{\sum_{m=1}^M \exp(i_n^T t_m / \sigma)}, \quad (8)$$

where i_n denotes the normalized image embedding, t_n is the text embedding matched with i_n while t_m is a non-matched text, and σ represents the temperature parameter. The loss of text-to-image is in a similar formulation.

Word-Region Alignment (WRA) uses Optimal Transport (OT) to optimize the alignment between input words and image regions [4].

There are also many other pretext objectives like **Scene Graph Generation (SGG)**, **Contrastive frame-transcript matching (CFTM)** and **Temporal Reordering (TF)**.

A.1.2 Two-Stage Pre-training on VCR. For VCR, some VL Transformers utilize a two-stage pre-training approach (*e.g.*, UNITER, VILLA): the model is first pre-trained on standard vision-language pre-training datasets; and then pre-trained on the downstream VCR dataset. Compared to the one-stage pre-training, the two-stage one leads to a slight improvement. One reason for this is that the images in the VCR dataset are from different domains of the pre-training.

Table 6: Performance of ViLBERT within three variants: full model, without tags, and with the random replacement of tags.

Model	Q→A	QA→R	Q→AR
ViLBERT	72.4	74.5	54.0
w/o tag	66.1 (-6.3)	70.1 (-4.4)	46.8 (-7.2)
repl. tag	70.7 (-1.7)	73.3 (-1.2)	52.1 (-1.8)

Specifically, the images in VCR mainly describe the movie plots

while in pre-training captioning datasets are often about natural scenes. For a fair comparison among all the VL Transformers, we only considered the first stage of pre-training for these two-stage pre-training methods.

A.2 Additional Results

We provide additional results of other VL Transformers which are not discussed in the main manuscript due to space limitations.

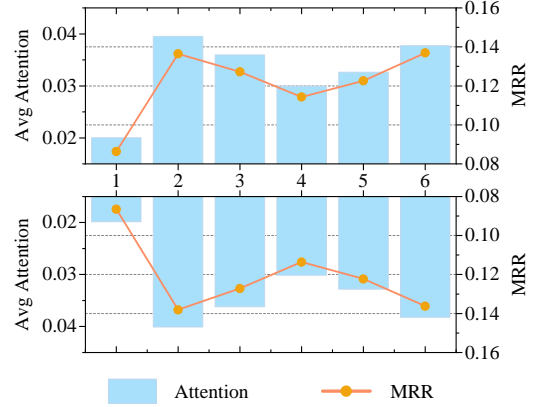


Figure 11: Average attention weights and MRR value with respect to each tag for different layers of ViL-BERT. Top: Q→A; Bottom: QA→R.

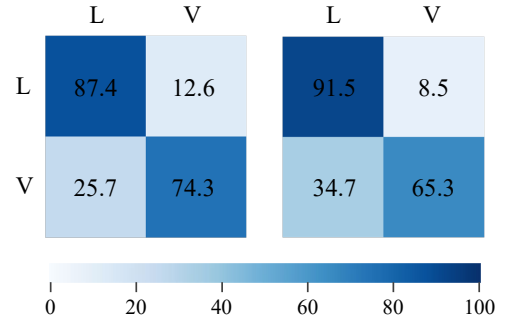


Figure 12: The attention weight distribution from the modality of each row to the modalities of columns of VILLA.

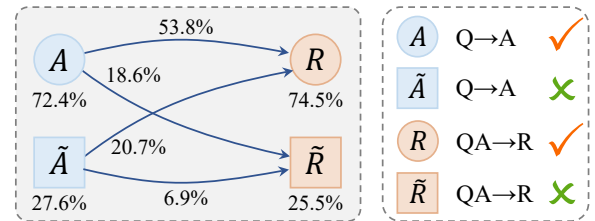


Figure 13: Proportion of correctly predicted instances from ViLBERT and its intersection of four states in VCR.