MIRTT: Learning Multimodal Interaction Representations from Trilinear Transformers for Visual Question Answering

Junjie Wang[1]*, **Yatai Ji**[2]*, **Jiaqi Sun**[2], **Yujiu Yang**[2]†, **Tetsuya Sakai**[1]†
[1]Waseda University, Shinjuku, Tokyo, Japan
[2] Graduate School at Shenzhen, Tsinghua University, ShenZhen, GuangDong, China
wjj1020181822@toki.waseda.jp
{jyt21, sunjq20}@mails.tsinghua.edu.cn
yang.yujiu@sz.tsinghua.edu.cn tetsuyasakai@acm.org

Abstract

In Visual Question Answering (VQA), existing bilinear methods focus on the interaction between images and questions. As a result, the answers are either spliced into the questions or utilized as labels only for classification. On the other hand, trilinear models such as the CTI model of Do et al. (2019) efficiently utilize the inter-modality information between answers, questions, and images, while ignoring intra-modality information. Inspired by these observations, we propose a new trilinear interaction framework called MIRTT (Learning Multimodal Interaction Representations from Trilinear Transformers), incorporating the attention mechanisms for capturing inter-modality and intra-modality relationships. Moreover, we design a two-stage workflow where a bilinear model reduces the free-form, open-ended VQA problem into a multiple-choice VQA problem. Furthermore, to obtain accurate and generic multimodal representations, we pre-train MIRTT with masked language prediction. Our method achieves state-of-the-art performance on the Visual7W Telling task and VQA-1.0 Multiple Choice task and outperforms bilinear baselines on the VQA-2.0, TDIUC and GQA datasets.

1 Introduction

One key challenge for building robust artificial intelligence systems is to handle information that lies across multimedia data. Visual Question Answering (VQA) (Wu et al., 2017) is a specific example of the challenge, where, given a natural language question about an accompanying image, the system is required to produce a correct answer. This is a typical multimodal problem since the intelligence system needs to understand images and texts simultaneously.

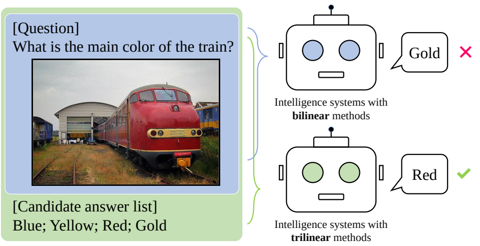
[Question]
What is the main color of the train?
[Candidate answer list]
Blue; Yellow; Red; Gold
Gold ✗
Intelligence systems with bilinear methods
Red ✓
Intelligence systems with trilinear methods

Figure 1: Visual question answering task[1]

From the perspective of a single modality, there have been plenty of backbone methods for learning better representations of either language or vision. For learning language representations, researchers have developed several pre-trained models, such as GPT-2 (Radford et al., 2019), BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019). These approaches can learn the universal language representations on the large-scale corpus, which are beneficial for downstream tasks (Qiu et al., 2020). Concerning visual representations, He et al. (2016); Ren et al. (2017); Simonyan and Zisserman (2015) have been widely applied to extract image features. Despite the success of these single-modality works, learning the relationships between different modalities is still an unsolved problem.

Existing VQA approaches focus on modeling the relationship between visual and language features represented by bilinear models. For example, through applying bilinear feature fusion methods, the image and text representations are projected into a uniformed higher-dimensional space. Multimodal Compact Bilinear pooling (MCB) (Fukui et al., 2016) processes the vectors in Fast Fourier Transform (FFT) space. For

* These authors contributed equally to this work and should be considered co-first authors.

† Corresponding author.

[1] An example of image-question-answer pair from Visual7W dataset (Zhu et al., 2016)

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 2280–2292
November 7–11, 2021. ©2021 Association for Computational Linguistics

better cross-modality information exchange, (Yu et al., 2019) and (Tan and Bansal, 2019) utilize co-attention/cross-attention networks to capture the high-level fusion features. Generally, these bilinear approaches only consider how to learn the joint representations between the questions and the images while the answers are processed as labels, making VQA task a multi-class classification task (Tan and Bansal, 2019; Zhu et al., 2016; Lu et al., 2019; Li et al., 2020).

However, the answers contain semantic information, related to the question and the visual context. For considering the answer information, trilinear models represented by Compact Trilinear Interaction (CTI) (Do et al., 2019), are designed to learn the alignment relationships between the visual context, the answers, and the questions. Unfortunately, the trilinear interaction in CTI only considers the inter-modality relationships but ignores the intra-modality information, leading to unsatisfactory inference results.

To tackle the above problems in the context of VQA, we propose a new trilinear modalities interaction framework called MIRTT (Learning Multimodal Interaction Representations from Trilinear Transformers). Specifically, MIRTT can extract more refined high-level feature information from the inter-modality and intra-modality relationships by introducing interactive attention networks across three modalities and three self-attention networks within a single modality. In general, MIRTT can accommodate requirements for processing three different modal features and efficiently utilize the information from the answers.

The contributions of our work are as follows:

- By considering the inter-modality and intra-modality relationships, we introduce a new end-to-end trilinear interaction model MIRTT, that enhances each single modality representation by proposed attention networks, resulting in better inference ability in VQA.
- We propose a two-stage workflow to simplify the harder Free-Form Opened-Ended (FFOE) VQA into simpler Multiple Choice (MC) VQA, which provides a method to solve difficult VQA tasks.
- Our proposed MIRTT achieves state-of-the-art performance on Visual7W telling task (Zhu et al., 2016) and VQA-1.0 for MC VQA and outperforms the bilinear methods on the VQA-2.0 (Goyal et al., 2017), (Kafle and Kanan, 2017a) and GQA (Hudson and Manning, 2019) datasets for FFOE VQA. Moreover, we take advantage of the pre-training task on our model, improving multi-modality understanding.

2 Related Work

Visual question answering (VQA) task. Following Antol et al. (2015) who defined the VQA task (i.e., obtaining answers from a given image-question pair), has received significant attention from the entire artificial intelligence community (Wu et al., 2017). There are two major types of VQA tasks, Multiple Choice (MC) VQA and Free-Form Opened-Ended (FFOE) VQA (Do et al., 2019). In MC VQA (Zhu et al., 2016; Kafle and Kanan, 2017b), the answer is chosen from a candidate answer list for a given image-question pair accessible in both training and test scenarios. FFOE VQA is more complicated since the answers are only available in the training phase, and there is no candidate answer list for choosing answers. However, FFOE VQA is the most common VQA task and almost all models are aimed at this problem. The general solution is to extract the visual features and linguist features first and then fuse them with a multi-modality fusion model, followed by a classifier or a generator to obtain the answer (Wu et al., 2017). Among them, exploring different fusion approaches is the mainstream research direction. On the one hand, the interactive relationships between the query image and the question have been fully modeled, such as element-wise operations (Antol et al., 2015) and bilinear methods (Fukui et al., 2016; Kim et al., 2018; Ben-Younes et al., 2017, 2019). On the other hand, some works have improved the VQA performance by considering the answer information (Hu et al., 2018; Do et al., 2019). For example, Jabri et al. (2016) combines the three input representations through a simple Multilayer Perceptron (MLP), and Wang et al. (2018) introduces a layered fusion operation by merging the image-question bilinear embeddings and the image-answer bilinear embeddings in joint embedding space. In order to solve VQA in a targeted manner, we make full use of the answer information and propose a two-stage workflow, which converts FFOE VQA to MC VQA.

Attention-based networks. Inspired by human's natural mechanism, Yang et al. (2016) introduce the attention mechanism to VQA and achieve success.

For bilinear feature fusion, some attention mechanisms have been proposed, such as co-attention (Lu et al., 2016) and dual attention (Nam et al., 2017). In terms of trilinear feature fusion, the attention map for trilinear inputs is computed by PARALING decomposition (Do et al., 2019). However, the output is only a joint vector for classification. In order to enhance each single modality representation by fusing the other modalities, we propose Trilinear Interaction Attention (TrI-Att). Moreover, although self-attention can not fuse different modalities, it can enhance the interaction within each modality (Yu et al., 2019). Therefore, we design Self-Attention (Self-Att) unit for capturing the intra-modality information.

Multimodal contextual representations. The transformer-based models can achieve good performance in the vision-language tasks. These models normally employ multi-layer transformers to learn multimodal contextual representations. There are two basic types of their architectures: single-stream and two-stream. The single-stream models concatenate image and language features first, and then they get the cross-modality representations with a single multi-layer transformer, such as VL-BERT (Su et al., 2020) and UNITER (Chen et al., 2020). The two-stream models take advantage of self-attention transformers to encode language and image features respectively, and then build joint representations with cross-attention transformers, such as LXMERT (Tan and Bansal, 2019) and ViLVERT (Lu et al., 2019). To better align vision-language semantic, some works try to pre-train transformer-based structures on a large corpus of image-text pairs. The pre-training tasks usually include masked language prediction, RoI-feature regression, detected-label classification and cross-modality matching (Tan and Bansal, 2019). In this paper, our proposed trilinear transformers deal with the three input embeddings different from former transformer-based methods.

3 MIRTT: Learning Multimodal Interaction Representations from Trilinear Transformers

As shown in Figure 3, our model considers three modality forms of input (e.g., images, questions and answers). The backbone of MIRTT is two transformers with multiple layers, which are based on TrI-Att and Self-Att mechanisms. Finally, in the output layer, we adopt MLP for specific downstream tasks.

3.1 Single-modality Embedding Extraction

Image embeddings. The image embeddings are extracted from a Faster R-CNN model (Anderson et al., 2018), a regional visual feature extractor. In terms of specification, for each object, it extracts a vector with d_v dimensions. Therefore, an image with v objects is represented as an embedding matrix $V \in \mathbb{R}^{v \times d_v}$.

Question and answer embeddings. We adopt BERT (Devlin et al., 2019) to fine-tune as our text extractor in the experiments. Specifically, the text is converted to WordPiece embeddings first (Wu et al., 2016). Then through fine-tuning, each embedding will be projected into $\mathbb{R}^{d_q}$ or $\mathbb{R}^{d_a}$, for question and answer, respectively. Finally, the question with a max length of q is represented as $Q \in \mathbb{R}^{q \times d_q}$, and the same for the answer that $A \in \mathbb{R}^{a \times d_a}$.

3.2 Trilinear Transformers

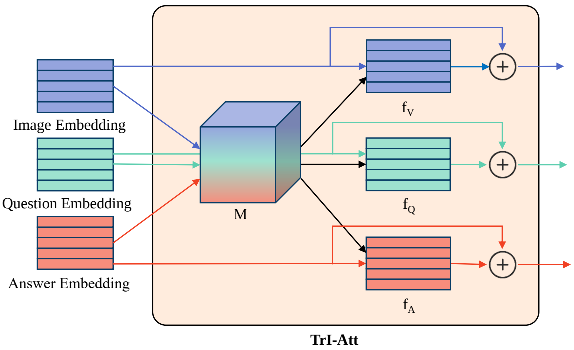

Figure 2: Trilinear interaction attention

TrI-Att for inter-modality representations. For better cross-modality information fusion, we design TrI-Att to project single-modality embedding into inter-modality enhanced space (Figure 2). From section 3.1, let $S = \{V, Q, A\}$ be the multimodal input collection. Firstly, we introduce the attention map $M \in \mathbb{R}^{v \times q \times a}$, which is mainly computed by matrix multiplication and sum-based dimension reduction. The detailed calculation process is as follows:

$$M = \mathrm{softmax}\left(\frac{\sum_{d_v} \sum_{d_q} \sum_{d_a} V \otimes Q \otimes A}{\sqrt{d}} \right) \tag{1}$$

where $\mathrm{softmax}$ is a normalization operation of all elements in M, and d is the arithmetic mean of d_v,

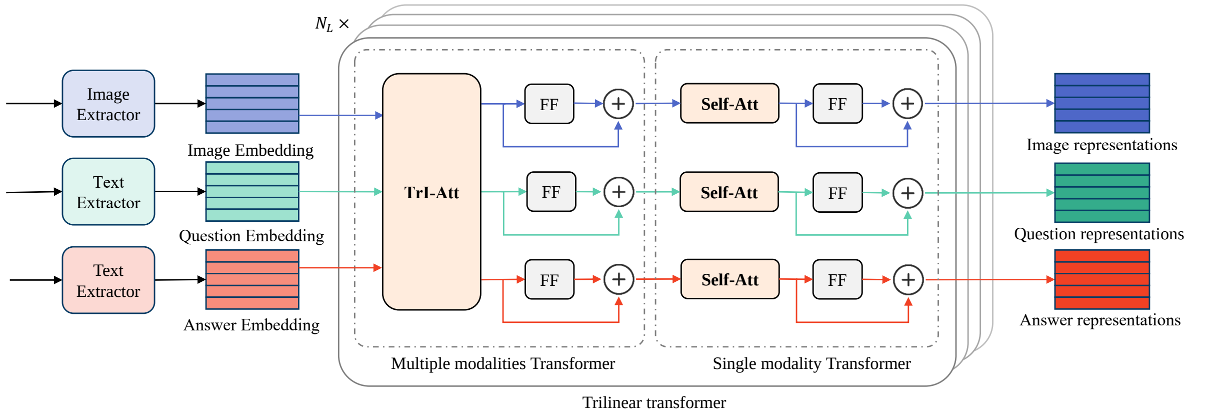

Figure 3: Model architecture of MIRTT

d_q and d_a. Secondly, the fusion of initial single-modality representation and the attention map f is conducted as follows:

$$f_V = \sum_q \sum_a MV = \text{TrI-Att}_\text{V}(V, Q, A) \quad (2)$$

here, we take image representation V for example (questions and answers are as the same), and the fusion operation is similar to Eq. 1.

We further utilize multi-head attention (Vaswani et al., 2017) to improve the robustness by introducing a linear mapping for each single-modality representation. In general, the complete calculation of inter-modality fusion is as follows:

$$f_V = \overset{N_h}{\underset{i}{||}} \text{TrI-Att}_\text{V}{}^i \left(VW_V^i, QW_Q^i, AW_V^i\right) \quad (3)$$

where W_V^i, W_Q^i and W_A^i are multi-head linear mappings, which are shared across the three forms of representations. N_h is the number of heads. $||$ indicates the concatenation of all multi-heads. Similarly, the fusion representations of questions and answers are:

$$f_Q = \overset{N_h}{\underset{i}{||}} \text{TrI-Att}_\text{Q}{}^i \left(VW_V^i, QW_Q^i, AW_V^i\right) \quad (4)$$

$$f_A = \overset{N_h}{\underset{i}{||}} \text{TrI-Att}_\text{A}{}^i \left(VW_V^i, QW_Q^i, AW_V^i\right) \quad (5)$$

After that, a fully connected feed-forward network with residual connection follows.

Self-Att for intra-modality representations. We apply the encoder of Transformer (Vaswani et al., 2017) to capture the intra-modality relationships. We deploy a multi-head self-attention mechanism, followed by a feed-forward network with the residual connection. With input feature $X \in \mathbb{R}^{n\times d}$, the multi-head self-attention is working as:

$$\overset{N_h}{\underset{i}{||}} \text{Self-Att}_\text{M}\left(X\right) = \overset{N_h}{\underset{i}{||}} \text{softmax}\left(\frac{XX^T}{\sqrt{d}}\right) XW_M^i \quad (6)$$

where $W_M^i \in \mathbb{R}^{d\times d_h}$ is the projection matrix for a certain modality M in ith head. This structure can enhance the long-distance dependency among the multi-modality features, while weaken negative impact on the result to a certain degree.

Trilinear transformers stacks. In total, the trilinear transformer stacks N_L layers, where each layer efficiently combines two transformer modules. The multiple modalities transformer has a trilinear interaction attention module and a fully connected feed-forward (FF) network. And the single modality transformer has three self-attention modules, following the same structure of the encoder in Transformer (Vaswani et al., 2017). Our essential motivation is to take advantage of the answer information, so a trilinear model is deployed first to fuse the three modality information. However, this leads to the loss of information in each modality to some extent, so a single-modality transformer is followed to reinforce the information of each own. For the MC VQA task, we put the pooled answer representations of the final layer into a binary classifier. Pick the answer of the highest binary score as the right one.

3.3 Two-stage Workflow

In FFOE VQA, previous models usually do not take the answer as input for keeping the same input dimensions in the training and test phases because

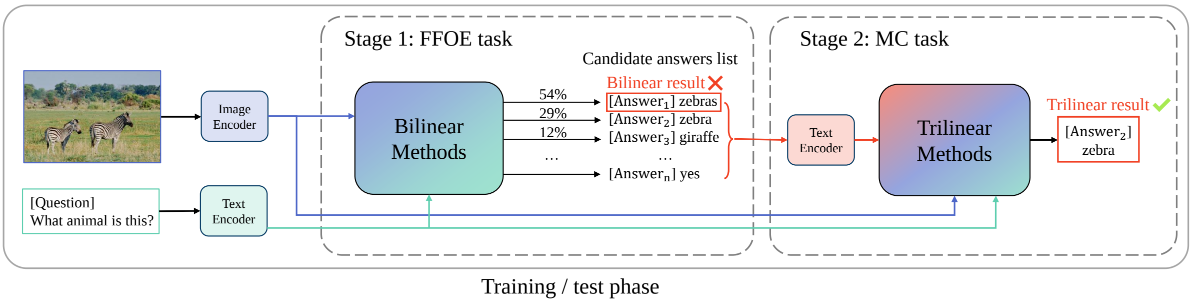

Figure 4: The overview of two-stage workflow[2]

the answer is not available in the test set. Knowledge distillation is a solution that trilinear methods could run by teaching a bilinear model in the training phase, and the bilinear model is evaluated in the test phase. However, the answer information is still inaccessible in the test set.

Therefore, as shown in Figure 4, we introduce a two-stage workflow to make full use of the dataset. Our proposal is a universal simplification process for the FFOE VQA task, which gives full play to the advantages of the bilinear and trilinear models.

In the first stage, we train a bilinear model for the questions and images first, and then the top four candidate answers are provided for each question based on the output logits. Since the bilinear model performs very high accuracy on the training set, the candidate answers basically contain the correct answer. In the test phase, the trilinear model is fully dependent on the candidates from the bilinear model.

In stage two, the candidates are first restructured into several image-question-answer pairs by reuse the input image and question; therefore, the number of the pairs is equal to that of the candidates. Then the trilinear model utilizes the image-question-answer pairs to choose a confident answer under the MC VQA task setting (illustrated in Section 3.2), where the answer-question and answer-image alignment information is learned.

4 Experiments

4.1 The Pre-training Strategy

In the hope of initializing our model effectively, we pre-train our model with the masked language modeling task, which is in a way similar to BERT (Devlin et al., 2019). Since our model is trilinear, the pre-training data format is triple of the question, image, and correct answer. We utilize Visual7W, VQA-2.0, and TDIUC datasets (the training set) to pre-train MIRTT. In detail, We mask the tokens of questions and answers with a probability of 15%. In these masked tokens, 80% of them are replaced by sign [MASK], 10% of them are kept, and the other 10% are replaced with random tokens.

Dataset	Model	Acc-MC
Visual7W	MCB	62.2
	CTI	72.3
	MIRTT (Ours)	**74.4**
VQA-1.0 MC	Dual-MFA	70.0
	MCB	70.1
	MFH	73.4
	MIRTT (Ours)	**77.0**

Table 1: Comparison with the state-of-the-art results on Visual7W and VQA-1.0. Our pre-trained MIRTT model outperforms previous methods.

4.2 Datasets and Evaluation Metrics

4.2.1 MC VQA Tasks

Dataset. Visual7W is a subset of Visual Genome (Krishna et al., 2017). For each question-image pair, there are four candidate answers, where only one choice is correct. There are two tasks for Visual7W: pointing and telling, and we conduct our method on telling task. VQA-1.0 MC (Antol et al., 2015) is similar to Visual7W, while there are 18 candidate answers for each question.

Metrics. Each question only has one correct answer. Accuracy (Acc-MC) is used to measure the performance (Zhu et al., 2016; Antol et al., 2015). We evaluate our methods on "test" split of Visual7W and "test-std" split of VQA-1.0 MC.

[2] An example image-question pair from VQA-2.0 dataset (Goyal et al., 2017)

Dataset	Text extractor	Bilinear method	Bilinear result	Trilinear result	Ensemble result
VQA-2.0	GRU	BAN2	66.5	65.5	**68.9**
	GRU	SAN	63.0	65.0	**66.8**
	BERT	MLP	59.5	64.3	**65.5**
	BERT	ViLBERT	69.2[3]	68.0	**70.3**
TDIUC	GRU	BAN2	85.5	87.4	**87.6**
	GRU	SAN	82.3	**86.0**	85.6
	BERT	MLP	80.0	**84.5**	83.6
GQA	BERT	BAN2	55.0	52.8	**55.7**
	BERT	SAN	54.8	52.4	**55.9**

Table 2: The results from test sets of VQA-2.0 and TDIUC. Comparisons between different bilinear methods and text encoders on stage one. The trilinear results are from MIRTT models on stage two.

4.2.2 FFOE VQA Tasks

Dataset. VQA-2.0 is built from MSCOCO dataset (Lin et al., 2014). VQA-2.0 minimizes answer biases so that a language-only "blind" model can not guess the right answers. TDIUC is a large VQA deadset of real images, which has over 1.6M questions of 12 categories. GQA consists of 22M questions, and each image corresponds to a scene graph. The questions focus on visual reasoning and compositional question answering.

Metrics. In VQA-2.0, each question has ten human-generated answers. To present the inter-human variability, we define the accuracy-based evaluation metric (ACC) as follows (Wu et al., 2017):

$$\mathrm{ACC} = \min\{\frac{n}{3}, 1\} \tag{7}$$

where n is the frequency of the answer given by the model in the answer set of the corresponding image-question pair. In TDIUC and GQA, there is only one right answer for each question. Therefore, normal accuracy is used. For details, we evaluate our methods on "test-dev" split of VQA-2.0, "Valid" split of TDIUC, and "test-std" split of GQA.

4.3 Implementation Details

Except for the referenced models and special instructions, we fine-tune BERT as our text extractor for questions and answers. And we freeze the Faster R-CNN detector (Anderson et al., 2018) without fine-tuning as the image extractor. For images, the maximum detected bounding box is set to 50. For texts, the questions and answers are trimmed to a sentence with a maximum length of 12 tokens and 6 tokens, respectively.

The hyper-parameters of MIRTT follow the default unless otherwise noted. The dimensions of input images (d_v), questions (d_q) and answers (d_a) are 2048, 768 and 768. To simplify the calculation, we reduce d_v to 768 with a linear projection. For the TrI-Att and Self-Att, the number of heads is 12, and the hidden dimension d_h is 64.

In all experiments with a two-stage workflow, we utilize six layers MIRTT with collection 2 (Table 3). Furthermore, our codes will be made publicly available with instructions https://github.com/IIGROUP/MIRTT. More experimental settings can be found in the Appendix.

4.4 MIRTT Performance on MC VQA

As shown in Table 1, we compare our methods with previous methods on Visual7W telling task and VQA-1.0 multiple-choice task.

MCB (Fukui et al., 2016): a method that considers FFT space to combine multimodal features.

CTI (Do et al., 2019): a method that learns high-level associations between three inputs by using multimodal-tensor-based decomposition.

Dual-MFA (Lu et al., 2018): a framework that fuses input embedding by selecting the free-form image regions and detection boxes most related to the input question.

MFH (Yu et al., 2018): a framework that models both the image attention and question attention simultaneously.

Our MIRTT with fine-tuning (Table 3) improves the CTI ACC-MC by 2.1% and improves the MFH ACC-MC by 3.6%.

[3] The result is not the same as in the cited paper. Regrettably, after a lot of experiments, we still cannot reach the accuracy in the cited papers. Under the fair experimental environment, the ensemble result outperforms the bilinear result.

4.5 MIRTT Performance on FFOE VQA

To evaluate the effectiveness of the two-stage workflow (Figure 4), we apply several bilinear methods as our backbones in stage one and set our MIRTT as trilinear methods in stage two. In detail, the candidate answers lists are generated by baselines; each contains four answers.

SAN (Yang et al., 2016): Stacked Attention Network utilizes multiple attention layers by querying an image multiple times to infer the answer.

BAN2 (Kim et al., 2018): Bilinear Attention Network fuses the question embeddings and image embeddings by utilizing co-attention.

MLP: in this method, we use the first output token embedding as the global representation of a question. Then, we sum up all object embeddings of an image after multiplying a learning weight for each one. The global representations of the question and the image are then added and fed into an MLP layer for classification.

ViLBERT (Lu et al., 2019) builds intra- and inter-relationship between vision and language base on a pretrained transformer structure.

Ensemble results: the predictions are calculated by considering the outputs of stage one and stage two. The ensemble method normalizes the two results separately and adds them together. The final prediction is the candidate answer with the highest probability.

As shown in Table 2, the trilinear results and ensemble results outperform the bilinear results. Our two-stage workflow solves the problem that trilinear models are not able to be deployed in FFOE VQA. Furthermore, ensemble results show that bilinear models can utilize the answers after modeling the answer information by the trilinear models.

In particular, the GQA dataset is not introduced in pre-training data. Our two-stage workflow and MIRTT present better performances than the baseline methods, which shows the generalization capability of our approaches.

4.6 Ablation Studies

4.6.1 The Components of MIRTT

Stacking layers and the size of pre-training data. As shown in Table 3, MIRTT only needs two layers to significantly outperform the others in "Random" based on accuracy.

Random: MIRTT is trained on Visual7W without pre-training.

Layers	Random	Collection 1	Collection 2
1	70.3	-	-
2	**70.9**	73.0	73.7
4	70.4	73.3	74.2
6	70.6	**73.5**	74.2
8	70.3	**73.5**	**74.4**

Table 3: The behaviors of the MIRTT with a different number of layers and different sizes of pre-training data on the Visual7W dataset.

Collection 1: MIRTT is pre-trained on the train sets of Visual7W and VQA-2.0.

Collection 2: MIRTT is pre-trained on the train sets of Visual7W, VQA-2.0, and TDIUC.

After pre-training, MIRTT outperforms the non-pre-trained one in each layer from random and collection 1. And as the number of layers increases, the accuracy of MIRTT with collection 1 is improved. However, as the number of layers increases, the capability of MIRTT with collection 1 seems to reach its limit at six layers, and growth hits a bottleneck.

Therefore, we add one more dataset to pre-train MIRTT. Comparing with collection 1, MIRTT in collection 2 can break the previously mentioned bottleneck and reaches the best score at the highest layer with more pre-train data. Perhaps similar to ViT (Dosovitskiy et al., 2021), these attention-based deep models are sensitive to dataset size. Therefore, the pre-trained MIRTT benefits from a larger number of parameters and more data, achieving an accuracy of 74.4%. Moreover, we conduct the randomized Tukey HSD p-values and effect sizes based on one-way ANOVA (Sakai, 2018) to support statistical significance of our results. Details are in the Appendix.

Attention mechanisms. Since CTI does not consider the intra-modality information, we attempt to build some structures to enhance it. In the term "CTI + Self-Att", the original output of CTI is a joint representation, then make a fusion by adding text embeddings and the joint representation. After that, we implement the Transformer's encoders (Vaswani et al., 2017) with two layers. As shown in Table 4, after adding self-attention to obtain fine-grained information within the modality, the CTI is improved by 0.5% compared to the original model.

BERT*: We fine-tune BERT on input questions and answers and fuse the extracted image embed-

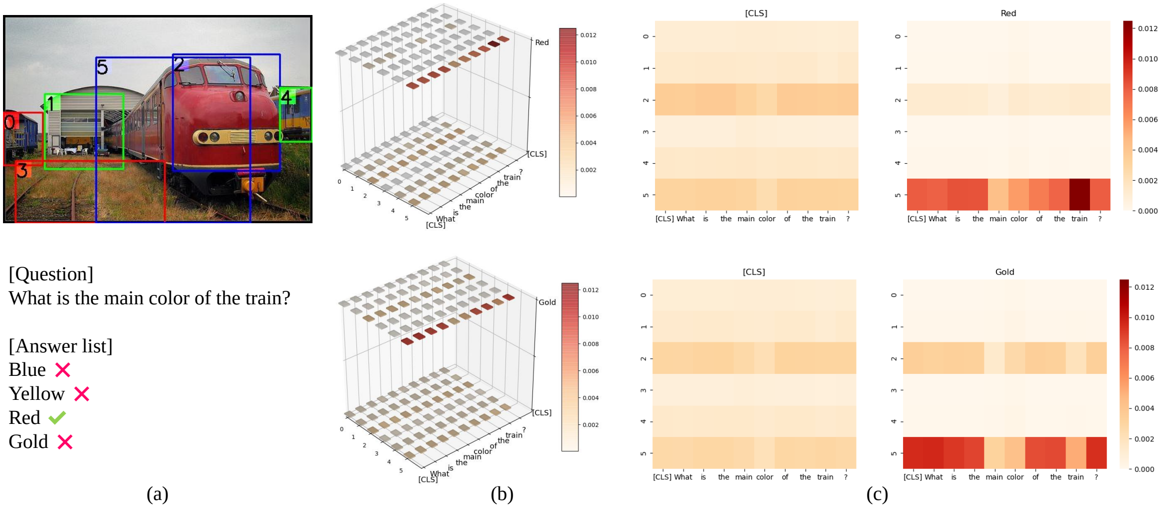

Figure 5: The visualization of the attention map M from Eq. 1. The attention map is extracted from the last layer of our best model with the best result on Visual7W. (a) is an example image-question-answer pair from the test set of Visual7W (Zhu et al., 2016). The input image is attached with bounding boxes. (b) includes the related attention maps for answers ("Red" and "Gold"). The details of each answer tokens are presented on (c).

Method	Acc-MC
CTI	72.3
CTI + Self-Att	**72.8**
BERT*	65.4
BERT + TrI-Att	70.5
BERT + TrI-Att + Self-Att (MIRTT)	**70.9**

Table 4: Ablation experiments for attention mechanism, evaluated on the test set of Visual7W.

dings. In detail, we utilize the same operation of Bottom-Up and Top-Down (BUTD) (Anderson et al., 2018) to fuse all representations.

To discuss two key components ("TrI-Att" and "Self-Att"), we utilize two layers of MIRTT without pre-training as our basic structure. By replacing the simple fusion methods like adding, we enhance the input embeddings by considering the inter-modality information in TrI-Att. 5.1% improves the accuracy as a result. Considering CTI can benefit from self-attention mechanism, we implement the Self-Att in our trilinear transformers. From the relative 0.4% improvement, our MIRTT can also learn the intra-modality information like "CTI + Self-Att".

Visualization for TrI-Att. Figure 5 visualizes the behavior of MIRTT by showing detailed attention values of TrI-Att. The detected objects are presented with their numerical labels. The special tokens in questions and answers are provided by BERT (Devlin et al., 2019). For the image-question-(answer "Red") pair, the correlation of object "5" (the train) and token "Red" has a great attention value. Moreover, the model focuses on the pair "5"-"train"-"Red", which is helpful in reasoning that the train in the image is red. In terms of the answer "Gold", the locomotive (object "2") gains more attention than the object "2" in "Red". Therefore, the answers could assist MIRTT in predicting the correct choices.

4.6.2 Cases for Two-stage Workflow

Figure 6 describes some examples with applying our two-stage workflow (Figure 4). In detail, the text extractors are all GRU, and the trilinear methods are MIRTT. The results show that our trilinear method is able to retrieve the most proper answer by utilizing the abundant information of the answers. Whether the problem requires stronger reasoning skills in (a), or the ability to find correspondences (images, questions, and answers) as in (b) and (c), MIRTT can handle it with a two-stage workflow. Following different bilinear methods as backbones, the trilinear method might predict different answers, such as (d).

5 Conclusions

We introduced a trilinear interaction framework called MIRTT, which captures inter-modality and intra-modality information of images, questions, and answers. Our method is based on TrI-Att and Self-Att mechanisms. The pre-trained model shows the effectiveness among the baselines on several datasets. Meanwhile, a two-stage workflow is introduced to apply the trilinear methods to FFOE VQA,

	[Images]				
Bilinear Backbone	[Questions]	(a) Q: What room are they located in?	(b) Q: Are people cooking?	(c) Q: What color is the glass covering the pilot?	(d) Q: How many people are in this photo?
BAN2	[Bilinear Answer]	office ✗	yes ✗	white ✗	6 ✗
	[**Trilinear** Answer]	classroom ✓	no ✓	blue ✓	5 ✗
SAN	[Bilinear Answer]	living room ✗	yes ✗	white ✗	5 ✗
	[**Trilinear** Answer]	classroom ✓	no ✓	blue ✓	4 ✓

Figure 6: A collection of image-question-answer pairs by random selection from VQA-2.0 (Goyal et al., 2017). Comparisons of whether to use two-stage workflow and different bilinear methods at the stage one in the test phase.

showing improvements on VQA-2.0, TDIUC and GQA. We achieve state-of-the-art results on Visual7W and VQA-1.0 MC. Generally, with rich experimental comparisons and extensive discussion, we demonstrate the value of the answer information and provide a solution for the VQA tasks.

Acknowledgements

This research was supported by the Key Program of the National Natural Science Foundation of China under Grant No. U1903213. We would like to thank members of The Real Sakai Laboratory[4], Waseda University, for giving us suggestions. Junjie Wang is especially grateful to our friend Yuxiang Zhang for his support, advice, and encouragement.

References

Peter Anderson, X. 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. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6077–6086.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. 2017. Mutan: Multimodal tucker fusion for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2612–2620.

Hedi Ben-Younes, Remi Cadene, Nicolas Thome, and Matthieu Cord. 2019. Block: Bilinear superdiagonal fusion for visual question answering and visual relationship detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8102–8109.

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*, pages 104–120. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, pages 4171–4186. Association for Computational Linguistics.

Tuong Do, Thanh-Toan Do, Huy Tran, Erman Tjiputra, and Quang D Tran. 2019. Compact trilinear interaction for visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 392–401.

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 *ICLR*. OpenReview.net.

Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *EMNLP*, pages 457–468. The Association for Computational Linguistics.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913.

[4] http://sakailab.com/english/

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*, pages 770–778. IEEE Computer Society.

Hexiang Hu, Wei-Lun Chao, and Fei Sha. 2018. Learning answer embeddings for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5428–5436.

Drew A. Hudson and Christopher D. Manning. 2019. GQA: a new dataset for compositional question answering over real-world images. *CoRR*, abs/1902.09506.

Allan Jabri, Armand Joulin, and Laurens Van Der Maaten. 2016. Revisiting visual question answering baselines. In *European conference on computer vision*, pages 727–739. Springer.

Kushal Kafle and Christopher Kanan. 2017a. An analysis of visual question answering algorithms. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1965–1973.

Kushal Kafle and Christopher Kanan. 2017b. An analysis of visual question answering algorithms. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1965–1973.

Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. 2018. Bilinear attention networks. In *NeurIPS*, pages 1571–1581.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard H. Hovy. 2017. RACE: large-scale reading comprehension dataset from examinations. In *EMNLP*, pages 785–794. Association for Computational Linguistics.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

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. *CoRR*, abs/1907.11692.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, pages 13–23.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. In *NIPS*, pages 289–297.

Pan Lu, Hongsheng Li, Wei Zhang, Jianyong Wang, and Xiaogang Wang. 2018. Co-attending free-form regions and detections with multi-modal multiplicative feature embedding for visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. 2017. Dual attention networks for multimodal reasoning and matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 299–307.

Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, pages 1–26.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2017. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6):1137–1149.

Tetsuya Sakai. 2018. *Laboratory experiments in information retrieval: Sample Sizes, Effect Sizes, and Statistical Power*. Springer. https://link.springer.com/book/10.1007/978-981-13-1199-4.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *ICLR*.

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 *ICLR*. OpenReview.net.

Hao Tan and Mohit Bansal. 2019. LXMERT: learning cross-modality encoder representations from transformers. In *EMNLP/IJCNLP (1)*, pages 5099–5110. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*, pages 5998–6008.

Zhe Wang, Xiaoyi Liu, Limin Wang, Yu Qiao, Xiaohui Xie, and Charless Fowlkes. 2018. Structured triplet learning with pos-tag guided attention for visual question answering. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1888–1896. IEEE.

Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony Dick, and Anton van den Hengel. 2017. Visual question answering: A survey of methods and datasets. *Computer Vision and Image Understanding*, 163:21–40.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alexander J. Smola. 2016. Stacked attention networks for image question answering. In *CVPR*, pages 21–29. IEEE Computer Society.

Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. 2019. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6281–6290.

Zhou Yu, Jun Yu, Chenchao Xiang, Jianping Fan, and Dacheng Tao. 2018. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. *IEEE transactions on neural networks and learning systems*, 29(12):5947–5959.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4995–5004.

A Dataset details

Dataset	Train	Valid	Test
Visual7W	69.8k	28.0k	42.0k
VQA-1.0	248.3k	121.5k	244.3k
VQA-2.0	443.8k	214.4k	447.8k
TDIUC	1115.3k	538.9k	-
GQA	15.4M	2.2M	4.2M

Table 5: The sizes of datasets associated to our paper

Pre-training Dataset. Amount of data for pre-training datasets is shown in Table 5. Pre-training dataset collection 1 includes Visual7W (train/val) and VQA-2.0 (train/val). In VQA-2.0, there is a list of human-generated answers to one question. We treat the answer with the highest score as the correct answer. The size of the image-question-answer pre-training collection is about 756k. In collection 2, we add the train set of TDIUC to pre-train MIRTT. The size of pre-training tuples grows to 1.87M.

B Implementation details

The whole details will be presented in our open-source codes on Github. The Focal loss (Lin et al., 2017) is used for training the proposed models.

Pre-training. When we pre-train MIRTT, the batch size is 128, and the initial learning rate is 1e-4. We use the model of epoch 7 for later fine-tuning.

Visual7W. Batch size is set to 32 for all models. For random initialization, the learning rate is 1e-4, and the number of the epoch is 17. For fine-tuning, the learning rate is 3e-5, and the number of the epoch is 11.

VQA-1.0. Initial learning rate is set to 1e-4, and batch size is 16. Since each question has 18 choices, there are 288 samples in one batch.

Two-stage workflow. The settings of hyper-parameters of two-stage workflow are presented in Table 6. Stage one and stage two are separate, not end-to-end structures. In stage one, the bilinear model is trained and we adopt a cross-entropy loss function to get the logits of the answers. Then, we get the top four candidates based on the logits, which is the generated answer list. In stage two, we encode the answers and put those embeddings of three modalities into MIRTT to get representations. We put the pooled answer representations into a binary classifier and apply binary cross-entropy based on labels generated from the FFOE dataset.

The number of candidate answers. To make our proposal a universal framework on both FFOE VQA and MC VQA tasks, we set the candidates to be four Visual7W on VQA and RACE (Lai et al., 2017) on QA. We will do related explorations based on this in the future. There are a few interesting problems. For example, if the candidate answers don't include the correct answer, the trilinear model won't work for this question. However, this problem is always possible unless the candidate list includes all the answers, which is impossible. A limited extension of the candidate list could help improve the coverage of correct answers while contradicting our design's universality.

C P-value based on Randomized Tukey HSD tests

Table 7, Table 8 and Table 9 show the statistical significance test results of the runs on Table 3. The name of runs are following the rules: name = D_L, where D $\in$ {Rand, Col1, Col2} is the name of the size of pre-training data and L $\in \{1, 2, 4, 6, 8\}$ is the number of layers to use in MIRTT. For example, Col2_2 stands for two layers MIRTT in collection 2.

			Stage one		Stage two	
Dataset	Text extractors	Bilinear method	BS	LR	BS	LR
VQA-2.0 (test-dev)	GRU	SAN	256	1.00E-03	64	1.00E-04
	GRU	BAN2	256	1.00E-03	64	1.00E-04
	BERT	MLP	256	1.00E-04	64	1.00E-04
	BERT	ViLBERT	256	1.00E-04	64	1.00E-04
TDIUC (valid)	GRU	BAN2	256	1.00E-03	64	1.00E-04
	GRU	SAN	256	1.00E-03	64	1.00E-04
	BERT	MLP	256	1.00E-04	64	1.00E-04
GQA (test-dev,test-std)	BERT	BAN2	256	1.00E-04	64	1.00E-04
	BERT	SAN	256	1.00E-04	64	1.00E-04

Table 6: The settings of hyper-parameters of two-stage workflow.

	Rand_1	Rand_2	Rand_4	Rand_6
Rand_2	p <0.001 (-0.835)	-	-	-
Rand_4	p <0.001 (-0.472)	p <0.001 (0.363)	-	-
Rand_6	p <0.001 (-0.480)	p <0.001 (0.355)	p = 0.706 (-0.008)	-
Rand_8	p <0.001 (-0.556)	p <0.001 (0.279)	p <0.001 (-0.084)	p <0.001 (-0.076)
Col1_2	p <0.001 (-0.659)	p <0.001 (0.176)	p <0.001 (-0.187)	p <0.001 (-0.179)
Col1_4	p <0.001 (-0.793)	p <0.001 (0.043)	p <0.001 (-0.320)	p <0.001 (-0.313)
Col1_6	p <0.001 (-0.706)	p <0.001 (0.128)	p <0.001 (-0.234)	p <0.001 (-0.227)
Col1_8	p <0.001 (-0.651)	p <0.001 (0.184)	p <0.001 (-0.179)	p <0.001 (-0.172)
Col2_2	p <0.001 (-0.549)	p <0.001 (0.286)	p <0.001 (-0.076)	p <0.001 (-0.069)
Col2_4	p <0.001 (-0.363)	p <0.001 (0.472)	p <0.001 (0.109)	p <0.001 (0.116)
Col2_6	p <0.001 (-0.280)	p <0.001 (0.555)	p <0.001 (0.192)	p <0.001 (0.200)
Col2_8	p <0.001 (-1.701)	p <0.001 (-0.866)	p <0.001 (-1.230)	p <0.001 (-1.222)

Table 7: Statistical significance calculated by Randomized Tukey HSD tests after 1,000 simulations. P-value and effect size. (Part 1)

	Rand_8	Col1_2	Col1_4	Col1_6
Col1_2	p <0.001 (-0.103)	-	-	-
Col1_4	p <0.001 (-0.236)	p <0.001 (-0.134)	-	-
Col1_6	p <0.001 (-0.150)	p <0.001 (-0.048)	p <0.001 (0.086)	-
Col1_8	p <0.001 (-0.095)	p = 0.759 (0.007)	p <0.001 (0.141)	p <0.001 (0.055)
Col2_2	p = 0.716 (0.008)	p <0.001 (0.110)	p <0.001 (0.244)	p <0.001 (0.158)
Col2_4	p <0.001 (0.193)	p <0.001 (0.295)	p <0.001 (0.429)	p <0.001 (0.342)
Col2_6	p <0.001 (0.276)	p <0.001 (0.379)	p <0.001 (0.513)	p <0.001 (0.427)
Col2_8	p <0.001 (-1.146)	p <0.001 (-1.043)	p <0.001 (-0.909)	p <0.001 (-0.995)

Table 8: Statistical significance calculated by Randomized Tukey HSD tests after 1,000 simulations. P-value and effect size. (Part 2)

	Col2_2	Col2_4	Col2_6
Col2_4	p <0.001 (0.185)	-	-
Col2_6	p <0.001 (0.269)	p <0.001 (0.084)	-
Col2_8	p <0.001 (-1.153)	p <0.001 (-1.338)	p <0.001 (-1.422)

Table 9: Statistical significance calculated by Randomized Tukey HSD tests after 1,000 simulations. P-value and effect size. (Part 3)