electronics

Article

MSGeN: Multimodal Selective Generation Network for Grounded Explanations

Dingbang Li [1], Wenzhou Chen [2] and Xin Lin [1,*]

[1] School of Computer Science and Technology, East China Normal University, Shanghai 200241, China; dbli@stu.ecnu.edu.cn
[2] College of Computer Science, Hangzhou Dianzi University, Hangzhou 310018, China; wenzhouchen@hdu.edu.cn
* Correspondence: xlin@cs.ecnu.edu.cn

Abstract: Modern models have shown impressive capabilities in visual reasoning tasks. However, the interpretability of their decision-making processes remains a challenge, causing uncertainty in their reliability. In response, we present the Multimodal Selective Generation Network (MSGeN), a novel approach to enhancing interpretability and transparency in visual reasoning. MSGeN can generate explanations that seamlessly integrate diverse modal information, providing a comprehensive and intuitive understanding of its decisions. The model consists of five collaborative components: (1) the Multimodal Encoder, which encodes and fuses input data; (2) the Reasoner, which is responsible for generating stepwise inference states; (3) the Selector, which is utilized for selecting the modality for each step's explanation; (4) the Speaker, which generates natural language descriptions; and (5) the Pointer, which produces visual cues. These components work harmoniously to generate explanations enriched with natural language context and visual cues. Our extensive experimentation demonstrates that MSGeN surpasses existing multimodal explanation generation models across various metrics, including BLEU, METEOR, ROUGE, CIDEr, SPICE, and Grounding. We also show detailed visual examples highlighting MSGeN's ability to generate comprehensive and coherent explanations, showcasing its effectiveness through practical case studies.

Keywords: visual question answering; explanation generation; multimodal; vision and language

check for updates

Citation: Li, D.; Chen, W.; Lin, X. MSGeN: Multimodal Selective Generation Network for Grounded Explanations. *Electronics* **2024**, *13*, 152. https://doi.org/10.3390/electronics13010152

Academic Editor: Fabio Grandi

Received: 8 November 2023
Revised: 13 December 2023
Accepted: 24 December 2023
Published: 29 December 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Artificial intelligence (AI) seeks to develop systems capable of processing intricate real-world data and mimicking human cognitive processes to make decisions influenced by environmental cues. In this pursuit, multimodal learning has emerged as a cornerstone of AI research that seamlessly integrates natural language and vision to fortify agent reasoning. Despite the proficiency of numerous models across diverse multimodal tasks [1–4], the need for interpretability remains a persistent challenge that hinders confident real-world deployment. This becomes evident in domains like visual question answering (VQA) [5–10], where a deficiency in explaining decisions exists. This limitation casts doubt on whether these models genuinely comprehend the causal relationships between multimodal inputs and answers or if they rely on data biases [11,12]. Consequently, enhancing the interpretability and grasp of visual reasoning models has become a pressing direction within the current research landscape.

In prior research, researchers have explored various methodologies to decipher the workings of visual reasoning models. These techniques encompass visualizing attention distributions or gradients to facilitate the analysis of decision-making processes [13–15]. Additionally, efforts have been made to generate natural language explanations that supplement answers [16–19]. Despite these endeavors, these approaches still fail to comprehensively explain the rationale underlying responses and to offer intuitive portrayals of the decision-making process itself. As illustrated in Figure 1, an attention mechanism's

visualization underscores localized image areas that capture the model's focus but do not delve into specific concepts or logical constructs. This limitation results in people having a vague and speculative understanding of the model. Natural language explanations can inadvertently introduce ambiguity at the image level. For instance, phrases like "yellow dog" in isolation fail to pinpoint the precise dog being referenced in the image. The model might erroneously associate "yellow dog" with the left dog and misconstrue the spatial relationship, potentially yielding an accurate but indistinct answer.

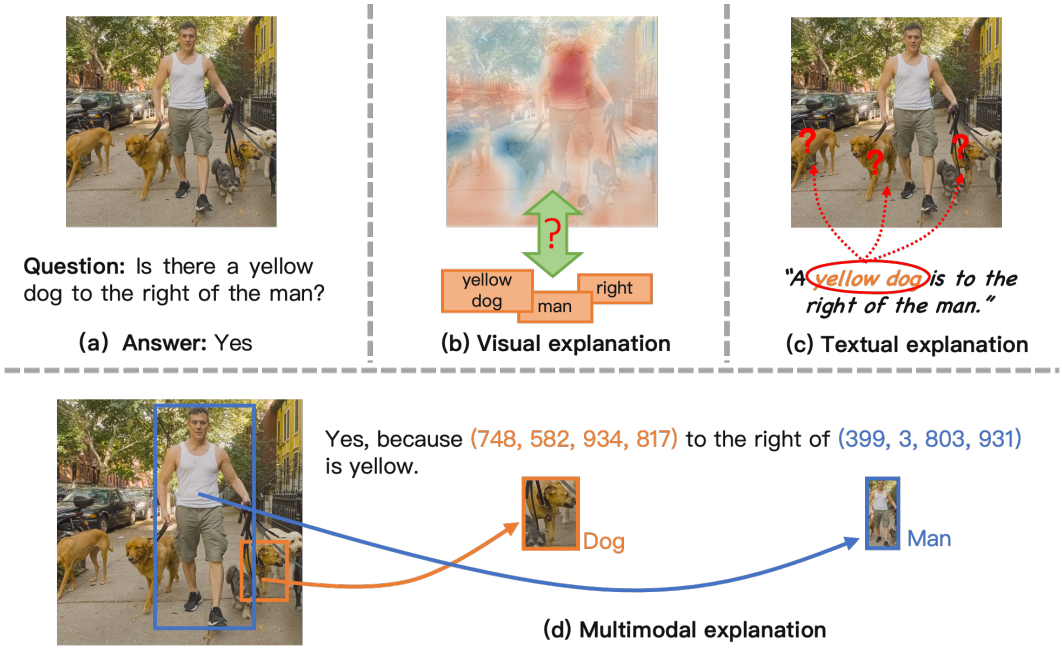


Figure 1. The figure shows our novel approach in contrast to previous methods. (**a**) Basic models predict answers without explanations. (**b**) Some methods visualize gradients or attention distributions from the visual module, but these explanations can lack clarity. (**c**) Language-generation-based methods may result in ambiguous textual explanations. (**d**) Our innovative approach enhances interpretability by generating cohesive natural language descriptions and visual regions.

Recognizing the limitations inherent in relying on a single modality to capture the intricacies of the model's reasoning fully, we propose an innovative approach: the **M**ultimodal **S**elective **Ge**neration **N**etwork (**MSGeN**). This network is designed to generate explanations that seamlessly amalgamate insights from diverse modalities, effectively eradicating ambiguity and enhancing specificity, as vividly depicted in Figure 1d. The MSGeN comprises five pivotal components, visually delineated in Figure 2: (1) Multimodal Encoder, (2) Reasoner, (3) Selector, (4) Speaker, and (5) Pointer. The Multimodal Encoder takes on the initial role of encoding the image and question into a shared embedding space, integrating information across modalities. The fused data are then channeled into the Reasoner, which embarks on sequential reasoning to generate latent state embeddings for each step. Dynamic modality selection is orchestrated by the Selector, which determines the appropriate generator (Speaker or Pointer) based on the current state embedding. The Speaker, functioning as a linguistic generator, crafts the natural language facet of the explanation, while the Pointer, a visual generator, forecasts bounding box coordinates by leveraging contextual cues from the explanation. By harnessing the prowess of MSGeN, we seamlessly orchestrate the generation of data across diverse modalities within a coherent sequence, culminating in an enriched and holistic understanding of the intricate decision-making process of the model.

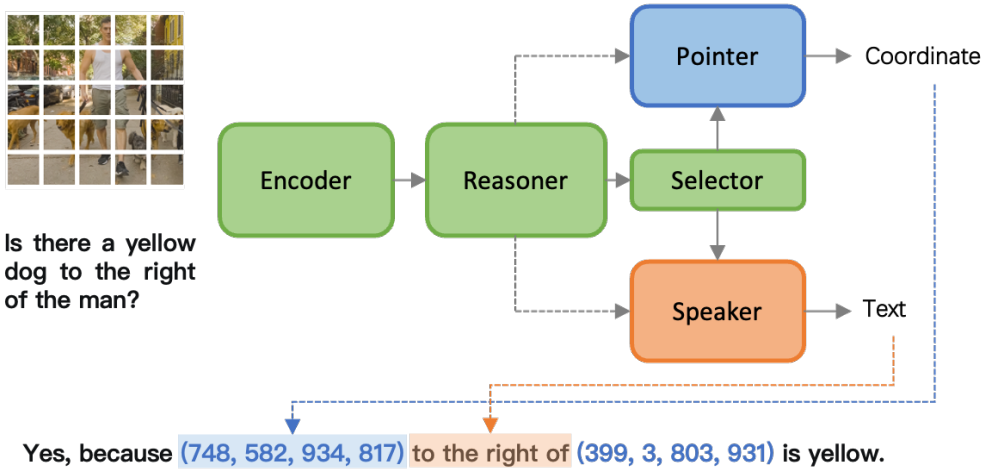


Figure 2. The schematic diagram of MSGeN.

We undertook a series of experiments to evaluate the efficacy of our novel approach. The results unequivocally showcase our method's capability to generate intuitive and coherent interpretations by seamlessly integrating visual and linguistic elements. This enriched fusion of modalities affords a more holistic grasp of the model's decision-making process, promoting both comprehensiveness and consistency. Furthermore, our proposed method surpasses the performance of established multimodal interpretation generation models and attains optimal outcomes.

From a practical application perspective, our MSGeN holds significant potential for widespread impact across various domains. For instance, within autonomous systems, MSGeN can enhance decision-making processes by providing action explanations that incorporate visual information, thus raising safety and reliability standards. In the healthcare sector, its interpretative capabilities can prove pivotal in disease diagnosis by offering clinicians intuitive rationales to establish trust and enhance diagnostic accuracy. Furthermore, in education, MSGeN can provide clear and comprehensible explanations based on visual content, thereby improving the model's expressiveness within intelligent educational processes. These applications underscore the importance of multimodal explanations for bridging the gap between AI decision-making and human understanding, emphasizing MSGeN's pivotal role in advancing these critical fields. To sum up, our paper contributes significantly to the following areas:

- We introduce an innovative multimodal explanation paradigm to enhance the interpretability of visual reasoning models.
- We propose MSGeN, a fresh methodology that allows for dynamic shifts in modality during the sequence generation process.
- Experiments showcase the superiority of our MSGeN over existing multimodal explanation generation models. Additionally, we conducted experimental analyses on various variants of MSGeN.

2. Related Works

2.1. Visual Question Answering

Visual Question Answering has progressed from rudimentary models to sophisticated architectures over time. Initially, VQA models represented images and questions as global features and employed cross-modal fusion to predict answers [20]. While image features, often derived from pre-trained Convolutional Neural Networks (CNNs) [21], served as a cornerstone, the field of question feature extraction has evolved from simplistic bag-of-words representations to intricate language models, including Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) [22]. Some advancements have introduced practical cross-modal fusion techniques, encompassing multimodal compact bilinear, pooling [23], low-rank bilinear pooling [24], and multimodal Tucker fusion [25]. Incorporating

bottom-up features and top-down attention, Anderson et al. [5] pioneered the understanding of image content by learning attention mechanisms over objects. Dense co-attention models further enhance the interaction between image regions and question words, facilitating a deeper comprehension of the image–question relationship and improved accuracy in question answering. Notably, the bilinear attention network [26], a sophisticated deep co-attention model, employs multiple bilinear attention layers to significantly enhance VQA performance. MCAN [6] extends these developments by learning intra-modal relevance through dense self-attention mechanisms.

2.2. Visual Reasoning Explanations

Prior research has explored diverse techniques to enhance model reliability by generating explanations for decisions. Methods include attention-based visualization modules [15], gradient analysis [27], and Shapley values [28], to highlight focus areas. Wang et al. [29] introduced explanatory scene graphs capturing object relationships. Zellers et al. [30] employed a multiple-choice setup for model-selected explanations. Krojer et al. [31] used diagnostic experiments, while Amizadeh et al. [32] decoupled question answering and perception with differentiable logic. Gokhale et al. [33] integrated logic with neural networks. Our focus is on generating more intuitive explanations. Park et al. [15] used LSTM for textual explanations and attention maps for visual evidence. Li et al. [16] multitasked answer and explanation generation. Dua et al. [19] employed generative models to overcome limited vocabularies. Sammani et al. [17] compactly combined explanations and reasoning using GPT. Vaideeswaran et al. [34] enhanced VQA models by integrating LSTM and Transformer decoders for generating interpretable and intuitive textual explanations. Chen et al. [35] proposed explanation generation with ROI indexes from object detection. In contrast, our approach generates multimodal explanations and autonomously searches for image evidence when needed.

2.3. Sequence Models

Sequence models, originating with Recurrent Neural Networks (RNNs) [36,37] and evolving into Transformer architectures [38,39], have excelled in natural language tasks like translation [40,41], question answering [42,43], and summarization [44,45]. They have also been adapted to multimodal tasks such as image captioning [46,47], video summarization [48], and more. Recent advancements demonstrate their capacity to handle non-sequential data by converting objects into token sequences—as seen in language models [49]—enabling applications like object detection [50,51] and image generation [52]. As multimodal models have evolved, versatile pre-training models have emerged [53–55]. Our work introduces a novel approach that fuses natural language generation and object detection via sequence modeling to generate grounded explanations for visual reasoning.

3. Approach

3.1. Preliminary

Our approach generates multimodal explanations for image question answering. Figure 3 presents the architecture of MSGeN, which takes images and related questions as inputs. Visual features are extracted from raw images, and questions are transformed into embedding features for multimodal fusion. The Reasoner, a decision-making component, adopts an attention-based decoder for superior handling of historical dependencies. The Selector controls modal information generation, allowing selective activation of the Speaker or Pointer module based on context and generating word sequences or region coordinates, respectively. This approach yields more comprehensive, coherent explanations than models relying on pre-extracted visual region features. Subsequent sections detail each module and illustrate how MSGeN generates interpretable multimodal explanations.

3.2. Multimodal Encoder

The multimodal encoder is responsible for processing the inputs and performing cross-modal feature fusion.

3.2.1. Feature Extraction

Most existing VQA models need offline preprocessed object detection networks to obtain visual region features. However, these approaches can misalign visual and textual information due to a lack of coupling. We adopt a more resource-efficient approach using a pre-trained ResNet [56] as the image feature extractor. We extract a feature map $V \in \mathbb{R}^{\frac{W}{P} \times \frac{H}{P} \times d}$ from the input image $I \in \mathbb{R}^{W \times H \times C}$ with a given patch size P. V is then flattened to obtain a sequence of patch features $V' \in \mathbb{R}^{\frac{WH}{P^2} \times d}$, which is used as a partial input for multimodal fusion.

For the textual modality to encode natural language questions, we transform the questions into sub-token sequences using byte-pair encoding [57] and learn embedding vectors for a fixed vocabulary, which allows us to encode the input question Q into $T \in \mathbb{R}^{N \times d}$, where N is the length of the sub-token sequence.

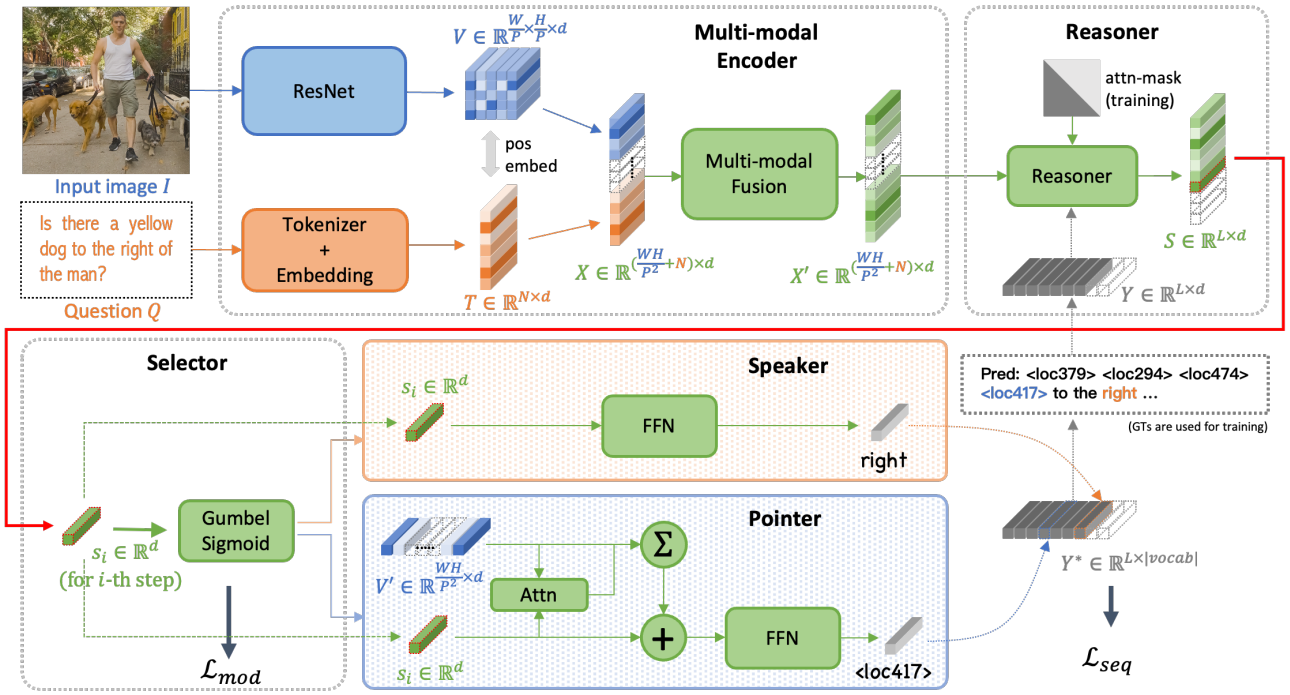


Figure 3. The overview of our Multimodal Selective Generation Network. The MSGeN model processes images and text inputs through its Multimodal Encoder using tools like ResNet for visual encoding and Tokenizer + Embedding for text. The encoded data are fused and passed to the Reasoner to generate reasoning states. The Selector then determines the use of the Speaker or Pointer modules based on these states. The Speaker generates natural language explanations, while the Pointer identifies visual target coordinates. This synergy results in comprehensive multimodal explanations that blend visual and textual elements.

3.2.2. Multimodal Fusion

The multimodal fusion module begins by concatenating the visual feature V' and textual feature T to obtain the input $X \in \mathbb{R}^{(\frac{WH}{P^2}+N) \times d}$. As shown in Figure 4, we stack multiple multi-head attention (MHA) layers as the central computational unit to perform multimodal fusion in order to capture cross-modal interactions effectively. Each MHA layer consists of a self-attention unit and a feed-forward network.

We add positional encodings to retain the position information of both the image and text during computation, which involves adding a trainable absolute positional embedding to the text and image features. Additionally, to enable the model to acquire more specific position information within the image, we add 2D relative attention [58,59] to the visual features. For the entries $x_i \in X$ that require the addition of 2D relative bias, the calculation process undergoes the following transformation, where w_{i-j} is the relative bias between i and j:

$$\sum_{x_j \in X} \frac{\exp\left(x_i^\top x_j\right)}{\sum_{x_k \in X} \exp\left(x_i^\top x_k\right)} x_j \Downarrow \sum_{x_j \in X} \frac{\exp\left(x_i^\top x_j + w_{i-j}\right)}{\sum_{x_k \in X} \exp\left(x_i^\top x_k + w_{i-k}\right)} x_j \tag{1}$$

After multiple self-attention layers, we get the multimodal fused feature $X' \in \mathbb{R}^{(\frac{WH}{P^2}+N)\times d}$.

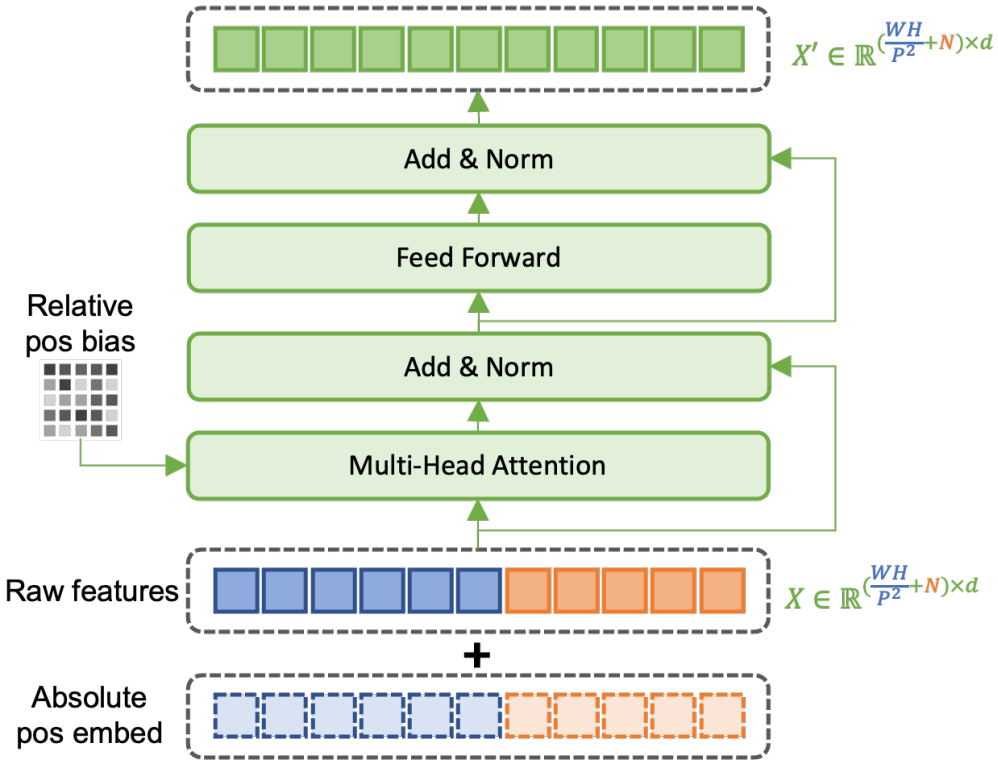


Figure 4. An MHA-based layer for multimodal fusion. We added absolute positional embedding at the beginning and relative positional bias at each layer.

3.3. Reasoner

The Reasoner serves as an intermediary, facilitating the generation of final explanations. While traditional transformer models generate natural language text sequences directly, our objective is to seamlessly produce information in different modalities in a unified sequence. To achieve this, we introduce hidden reasoning states as intermediate information carriers and perform sequence modeling. In visual reasoning tasks, the deduction process often unfolds step-by-step, necessitating inference based on prior information and updates to the ongoing reasoning based on the current state. The Reasoner module is set to effectively accommodate this stepwise deduction process and thus assists the model's reasoning capabilities. The Reasoner layer establishes essential connections between the historical

outputs and the fused multimodal features by combining a self-attention unit, a feed-forward neural network, and a cross-attention unit.

$$\begin{aligned} S &= \text{cross-attn}(Y', X') \\ &= \text{softmax}\left(\frac{Y'X'^{T}}{\sqrt{d}}\right)X' \\ Y' &= \text{self-attn}(Y) \end{aligned} \tag{2}$$

where Y is the embedding feature of the previously generated outputs. S is the sequence of hidden states, and the i-th entry denotes the reasoning state embedding s_i at step i. In practice, we input previously generated explanations into the Reasoner for inference but input ground truth during training and employ a triangular matrix mask to prevent ground-truth leakage. We stack multiple layers to construct the Reasoner. We also leverage multi-head attention, residual connections, and layer normalization techniques to enhance the model's learning capacity and stability.

3.4. Selector

The reasoning state s_i obtained through the Reasoner serves as a guide for the model to generate the final data at each step. However, the final data in our setting combine information from different modalities, introducing complexities in subsequent calculations. To address this, we propose a Selector module that selects the appropriate generation module based on the reasoning state.

Formally, the Selector parameterizes the module selection $m_{gen} \in \{Speaker, Pointer\}$ for each step's reasoning state, where *Speaker* represents the selection of the module for generating natural language text, and *Pointer* represents the generation of coordinates of image regions. We can view this step as a binary classification problem of different modalities. Typically, we can use *sigmoid* as the activation function to calculate the probability π of selecting the Speaker module or the probability $1 - \pi$ of selecting the Pointer module. The output module selection sequence $M \in \mathbb{R}^L$ is calculated as:

$$\begin{aligned} \pi_i &= \text{sigmoid}(s_i') \\ s_i' &= s_i W_s \end{aligned} \tag{3}$$

where $W_s \in \mathbb{R}^{d\times 1}$ is a learnable parameter matrix. The process of selecting different modules in the multimodal selection module introduces discrete variables, which causes the objective function to become non-differentiable. To overcome this issue, Jang et al. [60] introduced the Gumbel–Softmax technique, which provides a continuous approximation of the categorical distribution sampling process. Building on this approach, Geng et al. [61] introduced the Gumbel–Sigmoid method for selecting a subset of elements from a more extensive set, which allows the model to pay more attention to content words that contribute to the meaning of a sentence. Our model uses Gumbel–Sigmoid for module selection by adding Gumbel noise to the sigmoid function. This allows us to update the model parameters through back-propagation, resulting in a more efficient and effective selection process.

$$\begin{aligned} &\text{Gumbel-Sigmoid}\ (s') \\ &\quad = \text{sigmoid}\left((s' + g' - g'')/\tau\right) \\ &\quad = \frac{\exp((s' + g')/\tau)}{\exp((s' + g')/\tau) + \exp(g''/\tau)} \end{aligned} \tag{4}$$

where g' and g'' are samples drawn from two different Gumbel noise distributions. The term $\tau \in (0, \infty)$ is a temperature parameter controlling the distribution tendency of sampling results. As τ approaches zero, the samples drawn from the Gumbel–Sigmoid distribution become "cold" and resemble one-hot samples. During training, Gumbel–Sigmoid can

be used to obtain a differentiable probability π_i, i.e., Gumbel-Sigmoid(s'). In the inference stage, we select the module with the highest probability as the final generation module.

The proposed Selector module empowers the model to dynamically choose the generation module suitable for the given reasoning state. By leveraging the Reasoner's provided reasoning state and collaborating with the Selector module, our model achieves effective multimodal reasoning. This synergy ensures the generation of explanations encompassing different modalities, such as linguistic and visual components.

3.5. Generation

3.5.1. Speaker

We employ a two-layer feed-forward network to process the reasoning state s_i further and to map it into the output embedding space.

$$f_{s,i} = \sigma(s_i W_{s1} + b_{s1}) W_{s2} + b_{s2} \tag{5}$$

where W_{s1} and W_{s2} are learnable parameter matrices, b_{s1} and b_{s2} are corresponding biases, and $f_{s,i}$ denotes the Speaker output at step i.

3.5.2. Pointer

The Pointer module assumes a vital role in our model by generating precise coordinates corresponding to relevant visual regions within the image. In contrast to conventional visual detection models that yield direct coordinate values, our model uniquely generates both natural language and coordinates. The previous step's outcomes, integral to ongoing reasoning, are input to the Reasoner for predicting subsequent sequences, accentuating the need to unify representations across diverse modalities. A promising solution involves discretizing and representing image coordinates using tokens within a shared vocabulary. Recent progress [50] in computer vision research has highlighted the viability of tokenized coordinate representations for object detection. To represent the coordinate box of each visual region in a $W \times H$ image, we map the coordinates of the top–left (x_1, y_1) and bottom–right (x_2, y_2) corners of the detection box to a discrete coordinate system uniformly. The mapped discrete coordinates are then transformed into a sequence of $\{\lfloor \frac{1000x_1}{W} \rfloor, \lfloor \frac{1000y_1}{H} \rfloor, \lfloor \frac{1000x_2}{W} \rfloor, \lfloor \frac{1000y_2}{H} \rfloor\}$ tokens.

The Pointer model enables further access to visual information when generating the coordinate sequence. The reasoning state s_i interacts with visual features $V' \in \mathbb{R}^{(\frac{WH}{P^2}) \times d}$ that contains position information through the attention mechanism, resulting in more accurate coordinate values.

$$\begin{aligned} a_i &= \text{softmax}(V' s_i^T) \\ z_i &= V'^T a_i \end{aligned} \tag{6}$$

To generate the sequence of coordinates, we incorporate attention-weighted features $z_i \in \mathbb{R}^d$ with the original reasoning state h_i and feed them into a feed-forward network.

$$f_{p,i} = \sigma([s_i; z_i] W_{p1} + b_{p1}) W_{p2} + b_{p2} \tag{7}$$

where $[;]$ represents the concatenation of vectors, and $f_{p,i}$ denotes the Pointer output at step i.

3.5.3. Output

Using a linear network, we convert $z_{\cdot,i}$ into z_o. During inference, $z_{\cdot,i}$ is determined through argmax, while we substitute $\pi_i z_{s,i} + (1 - \pi_i) z_{p,i}$ for $z_{\cdot,i}$ during training. Subsequently, z_o is passed through the softmax function to compute a probability distribution, which ultimately governs the determination of the outcome for step-i. The token with the highest probability is chosen as the selection.

$$\begin{aligned} z_o &= z_{\cdot,i} W_o + b_o \\ P(\hat{y}) &= \text{softmax}(z_o) \\ y^* &= \text{argmax}\, P(\hat{y}) \end{aligned} \tag{8}$$

where $P(\hat{y})$ is the probability distribution of the current word estimated by the model, and y^* is the most likely word or coordinate.

3.6. Training and Inference

MSGeN uses maximum likelihood loss to optimize parameters. Given input image V, question Q, and multimodal explanation Y, the loss function is:

$$\mathcal{L}_{seq} = -\sum_{i=1}^{L} w_i \log P_\theta(y_i \mid y_{1:j-1}, V, Q) \tag{9}$$

where θ refers to the model parameters, L is the length of the target sequence, and w_i is the pre-assigned weight of the i-th token in the sequence. We set $w_i = 1, \forall i$ in our training process. The weight can be adjusted according to the actual situation for different modalities of tokens.

In addition, we can obtain the Selector's supervision signal through the ground-truth tokenization results and optimize the model parameters through binary cross-entropy loss.

$$\mathcal{L}_{mod} = -\sum_{i=1}^{L} (\gamma_i \log \pi_i + (1 - \gamma_i) \log (1 - \pi_i)) \tag{10}$$

where γ_i is the true probability of a modality at the i-th step. In summary, we combine the two loss functions to train our model jointly:

$$\mathcal{L} = \mathcal{L}_{seq} + \mathcal{L}_{mod} \tag{11}$$

During inference, we sample the tokens from the probability distribution P output by the model. This can be achieved by directly selecting the highest probability token or utilizing other random sampling techniques. Recent research has shown that utilizing nucleus sampling results in higher recall than argmax sampling [62]. In addition, we employ the beam search to enhance the quality of the results. The sequence is terminated when the model generates the "EOS" token.

4. Experiments

This section encompasses a comprehensive introduction to the datasets employed in our experiments, detailed information on the implementation of the model, and an analysis of the performance of the proposed model. Experimental results evince that our approach markedly enhances the accuracy and interpretability of the model, culminating in lucid and insightful multimodal explanations. Furthermore, we evaluated the model's proficiency in visual attribute extraction.

4.1. Datasets

We conducted experiments primarily on the GQA-REX dataset [35], a multimodal explanation dataset based on GQA [63]. The dataset comprises a whopping 1,040,830 pairs of question–answer explanations that are automatically parsed from the logical program in GQA. The explanations are structured as "…#1 *is to the left of* #2 …" and are accompanied by a list of pre-detected objects, where "#*i*" signifies the object index. To enhance model coupling and generalization as well as to bolster its object detection capacity, we replaced candidate indices with direct coordinates. Consequently, the adapted explanations adopt the form "… $(x_1^{(1)}, y_1^{(1)}, x_2^{(1)}, y_2^{(1)})$ *is to the left of* $(x_1^{(2)}, y_1^{(2)}, x_2^{(2)}, y_2^{(2)})$…". We optimized the model on the training set and evaluated it on the validation set. Furthermore, we

tested the model's answer prediction accuracy on the GQA dataset based on its dataset split. To delve deeper into the impact of data bias on the model's performance, we also conducted experiments on the GQA-OOD dataset [64].

4.2. Implementation Details

In this experiment, we opted for hidden sizes of 768 for both text and visual features. Notably, we expanded the dimensions of the feed-forward networks within the multimodal fusion and Reasoner modules to 3072. Multimodal fusion and Reasoner modules have 6 stacked layers each, employing 12 heads for multi-head attention. We implemented a two-phase training approach to guarantee the model's stability and performance. Initially, in the first phase, individual components were trained separately: the Speaker module was pre-trained on the MSCOCO dataset for image captioning tasks [65], and the Pointer module was trained on the RefCOCO dataset for referring expression comprehension tasks [66,67]. The second phase involved combined training, where the model was trained on the GQA-REX dataset to generate multimodal explanations and progressively advanced from single-modal to multimodal outputs and from simpler to more complex scenarios. We used the AdamW optimizer [68] with a learning rate of $1e-5$ for the first phase and $3e-5$ for the second phase. Additionally, label smoothing was set to 0.1, the warm-up ratio was 0.06, and the dropout was 0.1. We selected the checkpoint that yielded the highest CIDEr score on the training set for evaluation.

4.3. Evaluation

Our model's evaluation encompassed reasoning ability, explanation quality, and visual grounding performance. Reasoning ability was assessed using answer accuracy. To gauge explanation quality, we calculated Intersection over Union (IoU) scores between ground-truth and predicted coordinates, which were matched using the Hungarian algorithm. We employed five language evaluation metrics (BLEU4, METEOR, ROUGE-L, CIDEr, and SPICE) [14–16,35] to assess explanation quality further. We aggregated predicted grounding regions to evaluate visual grounding and calculated IoU against ground truth, following [63].

4.4. Experimental Results

In this section, we present the results of our comparative analysis, where we assessed the performance of our MSGeN model compared to several other explanation generation models. Notably, we benchmarked our model against the VQA baseline by Li et al. [69], which focuses solely on predicting answers. Additionally, we evaluated EXP [14] and VQAE [16], which were initially designed for generating natural language explanations but which have both been adapted to generate explanations containing object detection indices, akin to REX [35].

The results, as summarized in Table 1, provide compelling evidence that including explanations significantly enhances reasoning and improves answer accuracy. Notably, models originally designed for single-modality explanations exhibit limited visual grounding capabilities. In contrast, the REX model enhances visual grounding by incorporating cross-modal-interaction modules.

Our MSGeN model, designed to facilitate end-to-end multimodal explanation generation, demonstrates superior performance across various dimensions. It exhibits enhanced reasoning consistency, excels in terms of answer accuracy, and delivers high-quality explanations. Of particular note is our model's remarkable visual grounding ability. It detects objects during explanation generation and achieves visual grounding scores that surpass the adapted single-modality explanation model and are comparable to those of REX.

The exceptional visual grounding of our model, achieved without reliance on pre-object detection models, underscores its superiority in the realm of multimodal explanation and reasoning. These results are further supported by subsequent analyses presented in the discussion section.

Our experimental findings clearly demonstrate the advantages of our MSGeN model in terms of multimodal explanation generation, reasoning ability, and visual grounding, thereby highlighting its potential for a wide range of applications.

Table 1. Comparative results on explanation generation and question answering. GQA- and OOD- denote results on GQA and GQA-OOD, respectively. The best results of our model are highlighted in bold. The underscore represents the best grounding score achieved with an offline object detection model's assistance.

Method	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE	Grounding	GQA-Val	GQA-Test	OOD-Val	OOD-Test
Detection-based										
VisualBert	-	-	-	-	-	-	64.14%	56.41%	48.70%	47.03%
VQAE	42.56%	34.51%	73.59%	358.20	40.39%	31.29%	65.19%	57.24%	49.20%	46.28%
EXP	42.45%	34.46%	73.51%	357.10	40.35%	33.52%	65.17%	56.92%	49.43%	47.69%
REX	54.59%	39.22%	78.56%	464.20	46.80%	67.95%	66.16%	57.77%	50.26%	48.26%
Detection-free										
$MSGeN_{\alpha}$	74.59%	50.03%	**84.57%**	674.68	74.29%	**66.79%**	71.64%	62.31%	56.98%	54.52%
w/o cls	73.42%	**50.22%**	83.81%	675.37	74.02%	66.23%	71.13%	62.05%	56.76%	54.13%
w/o π	**75.02%**	49.65%	84.31%	671.21	74.18%	65.62%	70.94%	61.95%	56.05%	53.98%
$MSGeN_{\beta}$	73.34%	49.90%	84.45%	**684.92**	**75.41%**	66.44%	**72.24%**	**63.04%**	**57.02%**	**54.97%**
w/o cls	72.20%	49.13%	82.98%	676.12	73.89%	66.31%	71.34%	61.98%	56.83%	54.66%
w/o π	74.13%	50.03%	83.73%	676.57	74.04%	65.81%	71.22%	61.83%	56.41%	54.38%

4.5. Discussion

We conducted ablation experiments to investigate the factors that affected the performance of our model and evaluated the faith in the model's reasoning through quantitative and qualitative analyses of its explanations.

4.5.1. Effect of Generation Order

This section examines the influence of generation order on the efficacy of the MSGeN model by focusing on two variants: MSGeNα and MSGeNβ.

In the MSGeNα variant, the model adheres to a conventional sequence wherein the generation of explanations follows the prediction of the answer, formatted as: "*Answer: {A}. Because {Y}*". Conversely, the MSGeNβ variant employs a reversed order, initiating with explanation generation: "*Because {Y}. Answer: {A}*". This inversion is predicated on the hypothesis that preemptive explanation generation might substantively influence the model's reasoning trajectory and the quality of its outputs.

Table 1 presents the results of this comparison and shows nuanced differences in the performance of these variants. An observation is that $MSGeN_{\beta}$ exhibits a higher degree of answer accuracy. This enhancement can be attributed to the model's capability to refine its answer prediction based on the context established by the preceding explanation. The results suggest that the order impacts the model's reasoning efficacy in scenarios necessitating reasoning. The generation of explanations prior to answers appears to establish a more coherent and contextually informed framework for the prediction of answers.

This discovery highlights the crucial importance of explanations for enhancing the reasoning capabilities of our model, illuminating the significance of the order in which elements are generated within our model's architecture.

4.5.2. Enhancing Visual Content Recognition

Emulating human-like proficiency in evidence extraction from images is essential for developing multimodal explanation models. We train our model to concurrently predict bounding box coordinates and identify corresponding object labels, thereby enhancing its visual interpretative capabilities.

In this ablation experiment, the objective was to quantitatively evaluate the contribution of object recognition to the model's visual capacity. For this purpose, we trained a variant of our model to output only bounding box coordinates, omitting object labels.

The results presented in Table 1 (denoted as *w/o cls*) clearly demonstrate the impact of weakened recognition on the grounding ability of the model. Specifically, the weakened recognition led to lower grounding scores. This outcome underscores the critical role of object recognition for enhancing the model's grounding ability and affirms its capacity to effectively identify and interpret detected visual content.

At the same time, we can also observe a decrease in the variant's scores for explanation quality and answer accuracy. These phenomena demonstrate that object label prediction enhances the model's precision in identifying visual elements and its ability to contextualize them within a coherent sequence.

Our analysis and experimental findings affirm object recognition's integral role in augmenting multimodal explanation models' visual capacity. The synergistic combination of bounding box predictions and object label identification substantially elevates the model's grounding ability, culminating in more precise and context-rich multimodal explanations.

4.5.3. Generalization

To evaluate the generalization capability of the MSGeN model, we conducted assessments by deploying the model on the VQA-v2.0 dataset. In these evaluations, we compared our model with prediction-only models and primarily focusing on answer accuracy. The summarized results of these experiments are presented in Table 2.

Table 2. Experimental results on the VQA-v2 dataset.

Model	Test-Dev				Test-Std
	All	Y/N	Num	Other	All
BUTD [5]	65.32%	81.82%	44.21%	56.05%	65.67%
MuRel [70]	68.03%	84.77%	49.84%	57.85%	68.41%
Counter [71]	68.09%	83.14%	51.62%	58.97%	-
BAN [26]	69.52%	85.31%	50.93%	60.26%	-
MCAN [6]	70.63%	86.82%	53.26%	60.72%	70.90%
GNF [72]	70.51%	86.08%	54.41%	60.52%	70.71%
MSGeN$_\beta$	**75.63%**	**91.36%**	**60.10%**	**65.46%**	**76.02%**

Table 2 clearly illustrates that our MSGeN model outperforms prediction-only models regarding answer accuracy. This compelling result indicates that including explanation generation within our model has a positive and substantial impact on its overall performance in VQA tasks. A key factor contributing to this improvement in answer accuracy and generalization is the simultaneous generation of explanations alongside answer predictions. This parallel process equips our model with a more profound comprehension of the visual content and the reasoning required to produce accurate answers. This enhancement underscores the significance of multimodal reasoning and, importantly, highlights the invaluable logical capability that explanations bring to the realm of visual question answering.

4.5.4. Enhancing Multimodal Adaptability

In our model, we employ a unified vocabulary for both coordinates and words, enabling it to direct sequence generation through a single module. To gain insights into the importance of our modular architecture, we conducted experiments involving a simplified variant of MSGF. In this variant, we excluded the Switcher and Pointer components and relied solely on the Speaker to generate coordinates.

However, this simplified setup reduced complexity but resulted in a notable decrement in the model's capacity for modality differentiation. The impairment was particularly evident in the context of integrating and aligning visual information with coordinate generation. The experiment results are shown in Table 1 and are denoted as *w/o* π; they substantiate this assertion. A salient observation from these data is the discernible decline in grounding scores, which implies a compromised ability of the model to correlate textual descriptions with their respective visual counterparts accurately.

The decline in modality differentiation capacity can be attributed primarily to the absence of the dynamic module-selection mechanism provided by the Selector. Eliminating

the Pointer in this simplified variant underscores the limitations of a purely linguistic generation approach. The Selector is the key to contextual decision-making as it discerns the most-suitable module for each distinct phase of explanation generation. This functionality is indispensable, particularly in complex explanatory scenarios that demand sophisticated textual and visual data fusion.

The conducted experiments and analyses emphatically highlight the integral roles of the Selector and Pointer components within the MSGeN framework. Their inclusion is not merely a design preference but is essential for enhancing the model's adaptability and proficiency in handling multimodal data. This adaptability is particularly critical in applications that necessitate detailed interpretations of intricate, multimodal contexts. Hence, the modular architecture of MSGeN is not just a feature of efficient design: it is crucial for the model's ability to cohesively present various types of data.

4.5.5. Visual Reasoning Abilities

Visual question answering requires a range of reasoning skills, encompassing tasks such as object attribute recognition and spatial relationships, as noted in prior research [73,74]. Effective models excel at multimodal reasoning by encompassing a broader spectrum of visual concepts.

To gauge the distinct visual reasoning capabilities of our model, we conducted an analysis of recall rates for capturing visual concepts, as detailed in Table 3. Our model exhibits superior performance in concept capture compared to REX, particularly in the "Relation" category. This can be attributed to MSGeN's end-to-end architecture, which facilitates comprehensive visual relationship modeling without reliance on offline object detection models.

In addition to quantitative analysis, we evaluated the answers and explanations generated by our model, as depicted in Figure 5. Color-coded bounding boxes correspond to explanation coordinates, and object labels are presented in accordance with the coordinate sequence. Our model's unique capability to directly illustrate its reasoning by capturing relevant image regions significantly enhances interpretability.

In summary, our method demonstrates robust performance across diverse questions, providing satisfactory results rooted in the analysis of visual content.

Table 3. Recall for capturing different types of visual concepts.

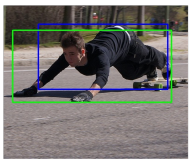
Concept	REX	MSGeN
Color	56.01%	**62.85%**
Material	49.27%	**57.31%**
Sport	72.77%	**78.53%**
Shape	40.64%	**58.94%**
Pose	74.80%	**80.71%**
Size	**65.31%**	64.87%
Activity	46.58%	**47.93%**
Relation	29.00%	**76.26%**

4.5.6. Conclusions

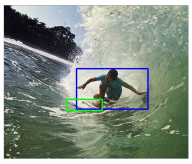
Our exploration to improve the interpretability of visual reasoning through the Multimodal Selective Generation Network (MSGeN) has yielded highly promising results. The MSGeN framework distinguishes itself by seamlessly generating explanations that intelligently integrate diverse modalities, thus providing a transparent and comprehensive understanding of the reasoning process. The collaborative components of MSGeN, including the Multimodal Encoder, Reasoner, Selector, Speaker, and Pointer, synergize to produce explanations that are both contextually rich and visually intuitive.

The comprehensive comparative analysis conducted in this study has unequivocally demonstrated that MSGeN surpasses current multimodal explanation models across a spectrum of key metrics. Qualitative case studies further substantiate these robust quantitative results and vividly illustrate MSGeN's remarkable capacity for generating coherent and detailed explanations. We also conducted several ablation studies to explore the reasons behind the superior performance of MSGeN, and further, we engaged in a discussion.

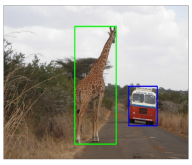
To summarize, MSGeN constitutes progress in the realm of explainable AI, specifically focusing on visual reasoning tasks. Its autonomous capability to seek and integrate pertinent image evidence as required fulfills the essential demand for models to exhibit transparency commensurate with their intelligence. This research not only introduces a multimodal explanation generation model but also sets the stage for future investigations into the complex interplay between the breadth of explanations and the dependability of models.



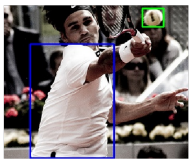
Question
What color is the shirt he is wearing?
Answer
Black
Explanation
Because (288, 123, 849, 549, shirt) that (176, 160, 867, 633, man) is wearing is black.



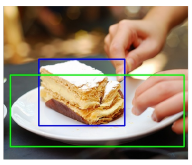
Question
What is the man on the snowboard doing?
Answer
Surfing
Explanation
Because (454, 414, 765, 677, man) on (412, 615, 567, 685, snowboard) is surfing.



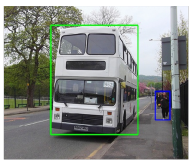
Question
Is the bus to the right of the giraffe old and blue?
Answer
No
Explanation
Because there (677, 527, 831, 783, bus) to the right of (384, 138, 609, 903, giraffe) is red and white.



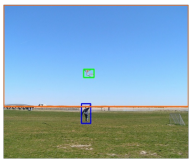
Question
Do the shirt and the ball have the same color?
Answer
Yes
Explanation
Because (248, 252, 531, 999, shirt) and (627, 21, 699, 142, ball) are all white.



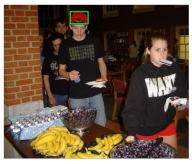
Question
Are there both cakes and plates in the picture?
Answer
Yes
Explanation
Because there are (158, 346, 545, 781, cake) and (30, 450, 813, 919, plate).



Question
Is the person on the left or right side of the bus?
Answer
Right
Explanation
Because (847, 555, 915, 741, person) is to the right of (364, 126, 767, 839, bus).



Question
Who is flying the red kite in the sky?
Answer
Man
Explanation
Because (420, 645, 470, 773, man) flying red (432, 420, 488, 472, kite) in (0, 0, 999, 663, sky) is man.



Question
What kind of clothing is not black?
Answer
Cap
Explanation
Because (358, 45, 454, 126, cap) is cap and is not black.

Figure 5. Visualization of the model output. The coordinate sequence corresponds to the bounding boxes with the same color in the image.

4.5.7. Future Directions

The MSGeN framework introduces a novel approach to explainable AI and visual reasoning and suggests several areas for pragmatic and focused future research:

1. Enhanced Generalization Capabilities: Future work should aim to extend the generalization of MSGeN across more diverse datasets and practical scenarios. This step is crucial for determining the framework's real-world efficacy and adaptability.

2. Improved Human–Model Interaction: Enhancing the interface and interaction with MSGeN is a practical goal. More-intuitive interfaces could facilitate broader use, making the technology more accessible to a wider audience.

3. Integration with Emerging Technologies: Exploring the integration of MSGeN with existing and emerging technologies, like large language models, could be beneficial. Such integration could offer new ways to present and interpret explanations, although this requires careful consideration of the added complexity and the clarity of the resulting explanations.

In summary, while MSGeN presents interesting possibilities in the realm of AI, it is important to approach its future development and application carefully considering its limitations and the challenges inherent in advancing AI technology.

Author Contributions: Methodology, D.L.; Validation, D.L.; Writing—original draft, D.L.; Writing—review and editing, D.L., W.C. and X.L.; Supervision, X.L.; Project administration, X.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key Research and Development Program of China (2021ZD0111000/2021ZD0111004) and the Science and Technology Commission of Shanghai Municipality (grant Nos. 21511100101, 22511105901, 22DZ2229004).

Data Availability Statement: Publicly available datasets were analyzed in this study. GQA dataset can be found here: https://cs.stanford.edu/people/dorarad/gqa. OOD dataset can be found here: https://github.com/gqa-ood/GQA-OOD. VQA-v2 dataset can be found here: https://visualqa.org.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Chen, S.; Jin, Q.; Wang, P.; Wu, Q. Say as you wish: Fine-grained control of image caption generation with abstract scene graphs. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 9962–9971.
2. Wang, X.; Liu, Y.; Shen, C.; Ng, C.C.; Luo, C.; Jin, L.; Chan, C.S.; van den Hengel, A.; Wang, L. On the general value of evidence, and bilingual scene-text visual question answering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10126–10135.
3. Kottur, S.; Moura, J.M.; Parikh, D.; Batra, D.; Rohrbach, M. Clevr-dialog: A diagnostic dataset for multi-round reasoning in visual dialog. *arXiv* **2019**, arXiv:1903.03166.
4. Yang, Z.; Chen, T.; Wang, L.; Luo, J. Improving one-stage visual grounding by recursive sub-query construction. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XIV 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 387–404.
5. Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; Zhang, L. Bottom-up and top-down attention for image captioning and visual question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6077–6086.
6. Yu, Z.; Yu, J.; Cui, Y.; Tao, D.; Tian, Q. Deep modular co-attention networks for visual question answering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 6281–6290.
7. Shi, J.; Zhang, H.; Li, J. Explainable and explicit visual reasoning over scene graphs. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 8376–8384.
8. Tan, H.; Bansal, M. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv* **2019**, arXiv:1908.07490.
9. Biten, A.F.; Litman, R.; Xie, Y.; Appalaraju, S.; Manmatha, R. Latr: Layout-aware transformer for scene-text vqa. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 16548–16558.
10. Ravi, S.; Chinchure, A.; Sigal, L.; Liao, R.; Shwartz, V. VLC-BERT: Visual Question Answering with Contextualized Commonsense Knowledge. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Vancouver, BC, Canada, 17–24 June 2023; pp. 1155–1165.

11. Manjunatha, V.; Saini, N.; Davis, L.S. Explicit bias discovery in visual question answering models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 9562–9571.
12. Guo, Y.; Nie, L.; Cheng, H.; Cheng, Z.; Kankanhalli, M.; Del Bimbo, A. On modality bias recognition and reduction. *ACM Trans. Multimed. Comput. Commun. Appl.* **2023**, *19*, 1–22. [CrossRef]
13. Patro, B.N.; Anupriy; Namboodiri, V.P. Explanation vs. attention: A two-player game to obtain attention for VQA and visual dialog. *Pattern Recognit.* **2022**, *132*, 108898. [CrossRef]
14. Wu, J.; Mooney, R.J. Faithful multimodal explanation for visual question answering. *arXiv* **2018**, arXiv:1809.02805.
15. Park, D.H.; Hendricks, L.A.; Akata, Z.; Rohrbach, A.; Schiele, B.; Darrell, T.; Rohrbach, M. Multimodal explanations: Justifying decisions and pointing to the evidence. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8779–8788.
16. Li, Q.; Tao, Q.; Joty, S.; Cai, J.; Luo, J. Vqa-e: Explaining, elaborating, and enhancing your answers for visual questions. In Proceedings of the European Conference on Computer Vision (ECCV), Salt Lake City, UT, USA, 18–23 June 2018; pp. 552–567.
17. Sammani, F.; Mukherjee, T.; Deligiannis, N. NLX-GPT: A model for natural language explanations in vision and vision-language tasks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 8322–8332.
18. Lu, P.; Mishra, S.; Xia, T.; Qiu, L.; Chang, K.W.; Zhu, S.C.; Tafjord, O.; Clark, P.; Kalyan, A. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 2507–2521.
19. Dua, R.; Kancheti, S.S.; Balasubramanian, V.N. Beyond vqa: Generating multi-word answers and rationales to visual questions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 1623–1632.
20. Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C.L.; Parikh, D. Vqa: Visual question answering. In Proceedings of the IEEE International Conference on Computer Vision, Boston, MA, USA, 7–12 June 2015; pp. 2425–2433.
21. Jiang, H.; Misra, I.; Rohrbach, M.; Learned-Miller, E.; Chen, X. In defense of grid features for visual question answering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10267–10276.
22. Hu, R.; Rohrbach, A.; Darrell, T.; Saenko, K. Language-conditioned graph networks for relational reasoning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Long Beach, CA, USA, 15–20 June 2019; pp. 10294–10303.
23. Fukui, A.; Park, D.H.; Yang, D.; Rohrbach, A.; Darrell, T.; Rohrbach, M. Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Las Vegas, NV, USA, 27–30 June 2016; pp. 457–468.
24. Kim, J.H.; On, K.W.; Lim, W.; Kim, J.; Ha, J.W.; Zhang, B.T. Hadamard product for low-rank bilinear pooling. *arXiv* **2016**, arXiv:1610.04325.
25. Ben-Younes, H.; Cadene, R.; Cord, M.; Thome, N. Mutan: Multimodal tucker fusion for visual question answering. In Proceedings of the IEEE International Conference on Computer Vision, Honolulu, HI, USA, 21–26 July 2017; pp. 2612–2620.
26. Kim, J.H.; Jun, J.; Zhang, B.T. Bilinear attention networks. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 1571–1581.
27. Simonyan, K.; Vedaldi, A.; Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv* **2013**, arXiv:1312.6034.
28. Merrick, L.; Taly, A. The explanation game: Explaining machine learning models using shapley values. In Proceedings of the Machine Learning and Knowledge Extraction: 4th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2020, Dublin, Ireland, 25–28 August 2020; Proceedings 4; Springer: Berlin/Heidelberg, Germany, 2020; pp. 17–38.
29. Wang, Y.; Yasunaga, M.; Ren, H.; Wada, S.; Leskovec, J. Vqa-gnn: Reasoning with multimodal semantic graph for visual question answering. *arXiv* **2022**, arXiv:2205.11501.
30. Zellers, R.; Bisk, Y.; Farhadi, A.; Choi, Y. From recognition to cognition: Visual commonsense reasoning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 6720–6731.
31. Krojer, B.; Adlakha, V.; Vineet, V.; Goyal, Y.; Ponti, E.; Reddy, S. Image retrieval from contextual descriptions. *arXiv* **2022**, arXiv:2203.15867.
32. Amizadeh, S.; Palangi, H.; Polozov, A.; Huang, Y.; Koishida, K. Neuro-symbolic visual reasoning: Disentangling. In Proceedings of the International Conference on Machine Learning (PMLR), online, 13–18 July 2020; pp. 279–290.
33. Gokhale, T.; Banerjee, P.; Baral, C.; Yang, Y. Vqa-lol: Visual question answering under the lens of logic. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XXI 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 379–396.
34. Vaideeswaran, R.; Gao, F.; Mathur, A.; Thattai, G. Towards reasoning-aware explainable vqa. *arXiv* **2022**, arXiv:2211.05190.
35. Chen, S.; Zhao, Q. Rex: Reasoning-aware and grounded explanation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 July 2022; pp. 15586–15595.
36. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
37. Dey, R.; Salem, F.M. Gate-variants of gated recurrent unit (GRU) neural networks. In Proceedings of the 2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS), Boston, MA, USA, 6–9 August 2017; pp. 1597–1600.

38. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 6000–6010.
39. Floridi, L.; Chiriatti, M. GPT-3: Its nature, scope, limits, and consequences. *Minds Mach.* **2020**, *30*, 681–694. [CrossRef]
40. Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to sequence learning with neural networks. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 3104–3112.
41. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv* **2014**, arXiv:1409.0473.
42. Akermi, I.; Heinecke, J.; Herledan, F. Transformer based natural language generation for question-answering. In Proceedings of the 13th International Conference on Natural Language Generation, Dublin, Ireland, 15–18 December 2020; pp. 349–359.
43. Duan, N.; Tang, D.; Chen, P.; Zhou, M. Question generation for question answering. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 7–11 September 2017; pp. 866–874.
44. Zhang, H.; Xu, J.; Wang, J. Pretraining-based natural language generation for text summarization. *arXiv* **2019**, arXiv:1902.09243.
45. Liu, Y.; Lapata, M. Text summarization with pretrained encoders. *arXiv* **2019**, arXiv:1908.08345.
46. Guo, L.; Liu, J.; Yao, P.; Li, J.; Lu, H. Mscap: Multi-style image captioning with unpaired stylized text. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 4204–4213.
47. Herdade, S.; Kappeler, A.; Boakye, K.; Soares, J. Image captioning: Transforming objects into words. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 11137–11147.
48. Lin, J.; Zhong, S.h.; Fares, A. Deep hierarchical LSTM networks with attention for video summarization. *Comput. Electr. Eng.* **2022**, *97*, 107618. [CrossRef]
49. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
50. Chen, T.; Saxena, S.; Li, L.; Fleet, D.J.; Hinton, G. Pix2seq: A language modeling framework for object detection. *arXiv* **2021**, arXiv:2109.10852.
51. Chen, T.; Saxena, S.; Li, L.; Lin, T.Y.; Fleet, D.J.; Hinton, G.E. A unified sequence interface for vision tasks. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 31333–31346.
52. Wang, P.; Yang, A.; Men, R.; Lin, J.; Bai, S.; Li, Z.; Ma, J.; Zhou, C.; Zhou, J.; Yang, H. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In Proceedings of the International Conference on Machine Learning (PMLR), Baltimore, MD, USA, 17–23 July 2022; pp. 23318–23340.
53. Kaiser, L.; Gomez, A.N.; Shazeer, N.; Vaswani, A.; Parmar, N.; Jones, L.; Uszkoreit, J. One model to learn them all. *arXiv* **2017**, arXiv:1706.05137.
54. Yang, Z.; Gan, Z.; Wang, J.; Hu, X.; Ahmed, F.; Liu, Z.; Lu, Y.; Wang, L. Unitab: Unifying text and box outputs for grounded vision-language modeling. In Proceedings of the Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, 23–27 October 2022; Proceedings, Part XXXVI; Springer: Berlin/Heidelberg, Germany, 2022; pp. 521–539.
55. Cho, J.; Lei, J.; Tan, H.; Bansal, M. Unifying vision-and-language tasks via text generation. In Proceedings of the International Conference on Machine Learning (PMLR), online, 18–24 July 2021; pp. 1931–1942.
56. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
57. Sennrich, R.; Haddow, B.; Birch, A. Neural Machine Translation of Rare Words with Subword Units. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin, Germany, 7–12 August 2016; pp. 1715–1725.
58. Wang, Z.; Yu, J.; Yu, A.W.; Dai, Z.; Tsvetkov, Y.; Cao, Y. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv* **2021**, arXiv:2108.10904.
59. Dai, Z.; Liu, H.; Le, Q.V.; Tan, M. Coatnet: Marrying convolution and attention for all data sizes. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 3965–3977.
60. Jang, E.; Gu, S.; Poole, B. Categorical reparameterization with gumbel-softmax. *arXiv* **2016**, arXiv:1611.01144.
61. Geng, X.; Wang, L.; Wang, X.; Qin, B.; Liu, T.; Tu, Z. How does selective mechanism improve self-attention networks? *arXiv* **2020**, arXiv:2005.00979.
62. Holtzman, A.; Buys, J.; Du, L.; Forbes, M.; Choi, Y. The curious case of neural text degeneration. *arXiv* **2019**, arXiv:1904.09751.
63. Hudson, D.A.; Manning, C.D. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 6700–6709.
64. Kervadec, C.; Antipov, G.; Baccouche, M.; Wolf, C. Roses are red, violets are blue... but should vqa expect them to? In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 2776–2785.
65. Chen, X.; Fang, H.; Lin, T.Y.; Vedantam, R.; Gupta, S.; Dollár, P.; Zitnick, C.L. Microsoft coco captions: Data collection and evaluation server. *arXiv* **2015**, arXiv:1504.00325.
66. Yu, L.; Poirson, P.; Yang, S.; Berg, A.C.; Berg, T.L. Modeling context in referring expressions. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part II 14; Springer: Berlin/Heidelberg, Germany, 2016; pp. 69–85.

67. Mao, J.; Huang, J.; Toshev, A.; Camburu, O.; Yuille, A.L.; Murphy, K. Generation and comprehension of unambiguous object descriptions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 11–20.
68. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. *arXiv* **2017**, arXiv:1711.05101.
69. Li, L.H.; Yatskar, M.; Yin, D.; Hsieh, C.J.; Chang, K.W. What does bert with vision look at? In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, online, 5–10 July 2020; pp. 5265–5275.
70. Cadene, R.; Ben-Younes, H.; Cord, M.; Thome, N. Murel: Multimodal relational reasoning for visual question answering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 5–20 June 2019; pp. 1989–1998.
71. Zhang, Y.; Hare, J.; Prügel-Bennett, A. Learning to count objects in natural images for visual question answering. *arXiv* **2018**, arXiv:1802.05766.
72. Qian, Y.; Hu, Y.; Wang, R.; Feng, F.; Wang, X. Question-Driven Graph Fusion Network For Visual Question Answering. *arXiv* **2022**, arXiv:2204.00975.
73. Mao, J.; Gan, C.; Kohli, P.; Tenenbaum, J.B.; Wu, J. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. *arXiv* **2019**, arXiv:1904.12584.
74. Whitehead, S.; Wu, H.; Ji, H.; Feris, R.; Saenko, K. Separating skills and concepts for novel visual question answering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 5632–5641.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.