

NMN-BART: Generating Natural Language Explanations for Visual Question Answering

Yan Zhou¹, Baifan Zhou^{2,1} and Ingrid C. Yu¹

¹Department of Informatics, University of Oslo, Norway

²Department of Computer Science, Oslo Metropolitan University, Norway

Abstract

Visual Question Answering (VQA) is a challenging task that requires reasoning over both visual and textual information. Recently, there has been growing interest in enhancing VQA with Natural Language Explanations (NLEs) to improve transparency and trust. While existing methods leverage powerful language models for explanation generation, many score high on lexical-level text similarity rather than capturing the underlying reasoning process. In this work, we propose NMN-BART, a novel architecture that combines Neural Module Networks (NMNs) with the pretrained BART language model, using cross-modal fusion to bridge visual semantics and textual reasoning. We evaluate NMN-BART on the VQA-X dataset, where it significantly outperforms baselines on semantic-based metrics, despite lower scores on lexical similarity metrics. This suggests that our method excels in capturing the meaningful content of the explanations, rather than matching the references in wording. The case study with human evaluation further verifies our finding that our method produces semantically rich and persuasive explanations.

Keywords

visual question answering with explanations, natural language explanation, neural module network

1. Introduction

Background. Visual Question Answering (VQA) [1] is a challenging task at the intersection of computer vision and natural language processing, where an accurate answer needs to be generated by reasoning over both visual and textual information given an image and a corresponding question. This task is crucial as it mirrors real-world scenarios where machines must integrate multimodal data, enabling more natural human-computer interactions.

Recently, there has been growing interest in enhancing VQA with Natural Language Explanations (NLEs) [2], leading to the task of Visual Question Answering with Explanation (VQA-E), which aims to produce both an accurate answer and a human-understandable explanation for that answer (Figure 1). While the full VQA-E task involves predicting both answers and explanations, in this work we focus on the generation of NLEs, assuming access to the correct answer, following prior work [3]. This targeted setup allows us to isolate and evaluate the model’s ability to produce semantically meaningful and interpretable justifications, an essential component for enhancing transparency and user trust in real-world applications such as medical diagnostics and autonomous driving. We further discuss the rationales in Section 3.

Challenges. Generating natural language explanations for VQA presents several challenges. A key difficulty is *aligning visual and textual modalities*: the system must interpret image content and link it to linguistic constructs to produce coherent explanations. Common approaches use vision-language models that separate answer prediction and explanation generation [4, 5], while end-to-end models [6] attempt joint prediction of both. These models typically combine vision encoders like CLIP [7, 8, 6] with transformer-based language models [6, 3, 9]. However, prior studies suggest that such models may rely on dataset biases or visual “shortcuts” [10], enabling them to achieve high answer accuracy

NAIS 2025: Symposium of the Norwegian AI Society, June 17–18, 2025, Tromsø, Norway

✉ yanzho@uio.no (Y. Zhou); baifan.zhou@oslomet.no (B. Zhou); ingridcy@ifi.uio.no (I. C. Yu)

id 0009-0003-5092-0059 (Y. Zhou); 0000-0003-3698-0541 (B. Zhou); 0009-0003-9764-3319 (I. C. Yu)



© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Question: "What is the animal doing?"

Answer: "Eating."

Explanation: "he is biting a vine of leaves with his mouth",
"he is very hungry at the moment",
"the giraffe is pulling food from the basket on the pole."

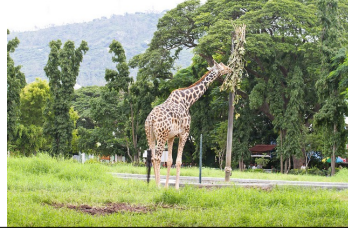


Figure 1: Example of visual question answering with explanation: the explanation should provide rationales that justify the answer.

without fully capturing the semantic content of the visual scene. This can lead to shallow or unfaithful explanations that fail to reflect the actual reasoning process.

Another limitation lies in generating *explanations that faithfully reflect the reasoning process*. Many multimodal models fail to explicitly connect reasoning steps with the generated explanation. Neural Module Networks (NMNs) [11, 10, 12] offer a structured reasoning framework by decomposing questions into interpretable modules. While effective for answer prediction in VQA tasks, NMNs remain under-explored for explanation generation, leaving their potential to produce reasoning-aligned outputs largely untapped.

Finally, *evaluating* NLEs remains a challenge. Many works [6, 3, 9, 8] try to achieve high scores on n-gram-based metrics such as BLEU and ROUGE. These metrics focus on lexical comparison, often insufficient in reflecting semantic alignment. Even semantically oriented metrics such as METEOR [13] or SPICE [14] still compare generated explanations against reference texts, and may overlook extra relevant information introduced in the generated explanation. Thus, new evaluation strategies are needed to assess explanation quality beyond similarity to references.

Contributions. This work presents our ongoing research on a novel architecture, NMN-BART, which integrates NMN with the pretrained language model BART using a cross-modal fusion module. This cross-modal fusion module integrates reasoning over Scene Graph (SG) representations with textual information from the language model, facilitating the generation of explanations grounded in both visual and textual reasoning. The compositional and explainable mechanism of NMN enables NMN-BART to model deeper reasoning process over the semantic relationships among the question, answer, and visual content, thereby producing explanations that are both semantically richer and more interpretable.

We evaluate NMN-BART on the VQA-X dataset, where our model achieves performance comparable to state-of-the-art methods. Notably, NMN-BART scores significantly higher on semantic-based metrics such as METEOR and SPICE, while exhibiting lower performance on n-gram based metrics such as BLEU and ROUGE-L. We interpret these findings as evidence that our approach is capable of generating explanations with enhanced semantic understanding, even when the generated text diverges from the reference in terms of exact phrasing. Case studies with human evaluation further confirm that NMN-BART generates rich and meaningful explanations, sometimes providing more information than the reference explanations. Our main contributions are summarised as follows:

- We propose NMN-BART, a novel architecture that combines the reasoning capabilities of NMN with the text generation of BART, for generating natural language explanations in VQA .
- We demonstrate through extensive experiments and evaluations on the VQA-X dataset that our approach yields explanations with improved semantic quality, as evidenced by semantic-based metrics. We adopt and adapt metrics for representative case studies and human evaluation.

2. Related Work

Visual question answering with explanation (VQA-E). Recent advances in VQA emphasise the importance of generating natural language explanations (NLEs) to justify answers and improve interpretability. Existing methods can be broadly grouped into two types: (i) generating explanations without conditioning on the answers, and (ii) generating explanations conditioned on the answers.

Category (i) approaches jointly predict answers and explanations. For example, NLX-GPT [6] frames the task as unified text generation. It integrates a CLIP image encoder with a distilled GPT-2 decoder, allowing the explanation to be generated as part of the reasoning process.

Category (ii) methods typically decouple the VQA and explanation generation stages. They first predict an answer, then condition the explanation on the answer, question, and image features [5, 4, 3, 9]. For instance, the Rational Transformer [3], combines GPT-2 with outputs from object detection, situation recognition, and commonsense inference to generate rationales for complex visual reasoning tasks. e-UG [9] similarly use GPT-2, conditioning explanations on various combinations of visual and textual features. S³C [8] improves explanation by incorporating answer scores as rewards in a self-critical learning framework, using CLIP-based encoders and prompt-based templates to guide the generation.

Reasoning-enhanced Vision-Language (VL) models. Recent work has shown that incorporating explicit reasoning into language models improves performance on complex tasks [15, 16]. However, due to the complexity of aligning and integrating cross-modal information, VL models still struggle to capture visual reasoning effectively.

One line of research adopts Neural Module Networks (NMNs) [11], which dynamically compose neural modules based on the input question. To mitigate the vision-to-reasoning shortcut in NMNs, XNM [10] employs scene graphs for visual reasoning, instead of using “low-level” visual perception, especially in datasets like CLEVR [17] with ground-truth scene graph annotations. For datasets without scene graphs, such as VQA [1], XNM constructs scene graph representations from visual features to enable dynamic reasoning [18]. We adopt this strategy with the VQA-X [5] dataset.

Inspired by NMN’s compositional reasoning, recent methods explore code generation [19, 20] or large language models (LLMs) [21] for step-by-step visual reasoning. Chain-of-Thought (CoT) prompting has also been applied in VL reasoning. For instance, CCoT [22] generates scene graphs via LLMs and uses them in prompts to extract compositional knowledge. These approaches offer flexibility and strong generalisation, with their zero-shot performance avoiding task-specific training or fine-tuning.

In this work, we integrate a NMN [18, 23] with BART, a transformer-based language model [24], leveraging NMN for compositional reasoning over scene graph representations and BART for natural language generation. This guides explanation generation and helps the model capture richer semantic relations among the image, question, and answer.

3. Task Formulation and Rationale

In Visual Question Answering with Explanation (VQA-E), the objective is to develop a model that can answer questions about images, and generate textual explanations of the answers, by understanding and reasoning over both visual and textual information. We follow the general task formulation in [9], which denotes a visual information as V (e.g., image), a textual information as Q (e.g., question), and the objective of VQA-E is to learn a function \mathcal{F} to predict the answer A to the question, and the explanation E that justifies the answer A : $A, E = \mathcal{F}(V, Q)$.

There are generally two paradigms for achieving this. One class of approaches generates the answer and explanation simultaneously, without conditioning the explanation on the answer. Another class adopts a *post-hoc* (after the fact) strategy, where the answer is first determined or given, and the explanation is then generated conditioned on that answer. The task is then changed to $A = \mathcal{F}_A(V, Q), E = \mathcal{F}_E(V, Q, A)$. Some works generate both answers and explanations (conditioned or unconditioned on the answer) and then filter out explanations where the answers are incorrect during evaluation, a setting referred to as the *filtered setting* [8]. In this work, we adopt the post-hoc strategy and condition the explanation generation on the answer. We follow a design choice similar to [3], in which a given answer is provided to the explanation generation as additional input to V and Q . This design offers several advantages:

- *No dataset filtering*: It allows evaluation over the full dataset without filtering for only correct predictions, preserving the diversity and difficulty of the original examples.

- *Bias mitigation*: It avoids biases introduced by the filtered setting that remove cases where the predicted answer is incorrect, which are often difficult or ambiguous cases, and potentially informative and important for assessing explanation quality.
- *Focused generation*: By receiving the answer as input, the model can concentrate on elaborating, contextualising, and justifying the answer, leading to more relevant and detailed explanations.

Based on these rationales, we formulate our task as learning a function \mathcal{F}_E that generates a natural language explanation E from the image V , the question Q , and the given answer A (Eq. 1). Here, Q and E are natural language sentences, and A is typically a word or a short phrase. The explanation E consists of one or more sentences that provide a human-understandable rationale for the given answer.

$$E = \mathcal{F}_E(V, Q, A) \quad (1)$$

4. Our Approach: NMN-BART

NMN-BART consists of (1) visual preprocessing, (2) an NMN-BART encoder, and (3) a BART decoder (Figure 2). Visual preprocessing transforms the image to scene graph representations G , the NMN-BART encoder takes the G and the text (question Q and answer A) as input, and produces fused encoding, which is then processed by the BART decoder to generate the explanation E .

Visual preprocessing. Here the image V is transformed into scene graph representations G . Visual features are extracted from images using the Bottom-Up Attention model [25], a pretrained *Visual Model* that employs a Faster R-CNN detector trained on the Visual Genome dataset. These visual features, termed visual Region-of-Interest (RoI) features, serve as a visual foundation for the construction of G .

To structure visual information for compositional reasoning, we convert these region-level features into scene graph representations following [10]. A scene graph is a structured representation of an image that encodes objects (nodes), their attributes, and pairwise relations (edges) between them. In our approach, each node \mathbf{v} in the graph corresponds to a detected object and is constructed from the visual RoI features. Each edge \mathbf{e} represents a spatial or semantic relation between two objects and is constructed by concatenating the visual RoI features of the two connected nodes: $\mathbf{e}_{ij} = [\mathbf{v}_i; \mathbf{v}_j]$.

This graph-based representation provides a structured abstraction over raw pixel data, which enables NMN to reason over entities and relations in a compositional manner, facilitating interpretable reasoning for the explanation generation.

NMN-BART Encoder. The NMN-BART encoder can be seen as BART encoder enhanced with NMN (Figure 2). The BART encoder layers transform text into contextualised feature representations. The NMN processes these text features along with the scene graph representations to learn neural modules, producing intermediate results referred to as module output. The cross-modal fusion layers then combine these module outputs with the output of the previous encoder or fusion layer to form the fused encoding \mathbf{H}^L , which is used by the BART decoder to generate explanations. The overall functionality of NMN-BART encoder can be summarised as: $\mathbf{H}^L = \text{NMN-BART-Encoder}(Q, A, G)$.

BART encoder layers. We use a pretrained BART as the foundational language model, leveraging its encoder to process textual inputs (questions and answers). We choose BART, a transformer-based sequence-to-sequence model, for it combines the strengths of bidirectional and autoregressive transformers, making it effective for natural language generation tasks [24]. The text inputs Q, A are tokenised into a sequence of tokens $\mathbf{W} = \{\mathbf{w}_0, \mathbf{w}_1, \dots, \mathbf{w}_N\} = \text{Tokeniser}(Q, A)$ and embedded into dense vector representations $\mathbf{H}^0 = \text{Embed}(\mathbf{W})$, which are then passed to the first encoder layer. The output representation of each encoder layer l is computed as:

$$\mathbf{H}^l = \{\mathbf{h}_0^l, \mathbf{h}_1^l, \dots, \mathbf{h}_N^l\} = \text{Encoder-Layer}(\{\mathbf{h}_0^{l-1}, \mathbf{h}_1^{l-1}, \dots, \mathbf{h}_N^{l-1}\}) \quad \text{for } l = 1, \dots, L \quad (2)$$

where $\text{Encoder-layer}(\cdot)$ is a single transformer encoder layer, and $\{\mathbf{h}_0^l, \mathbf{h}_1^l, \dots, \mathbf{h}_N^l\}$ denotes the hidden states \mathbf{H}^l of the l -th layer. L is the number of encoder layers of the pretrained BART model.

Neural Module Network. NMN takes two inputs: 1) textual features from Q and A , including text embedding \mathbf{H}^0 and the hidden states from the first encoder layer \mathbf{H}^1 ; 2) the scene graph representations G .

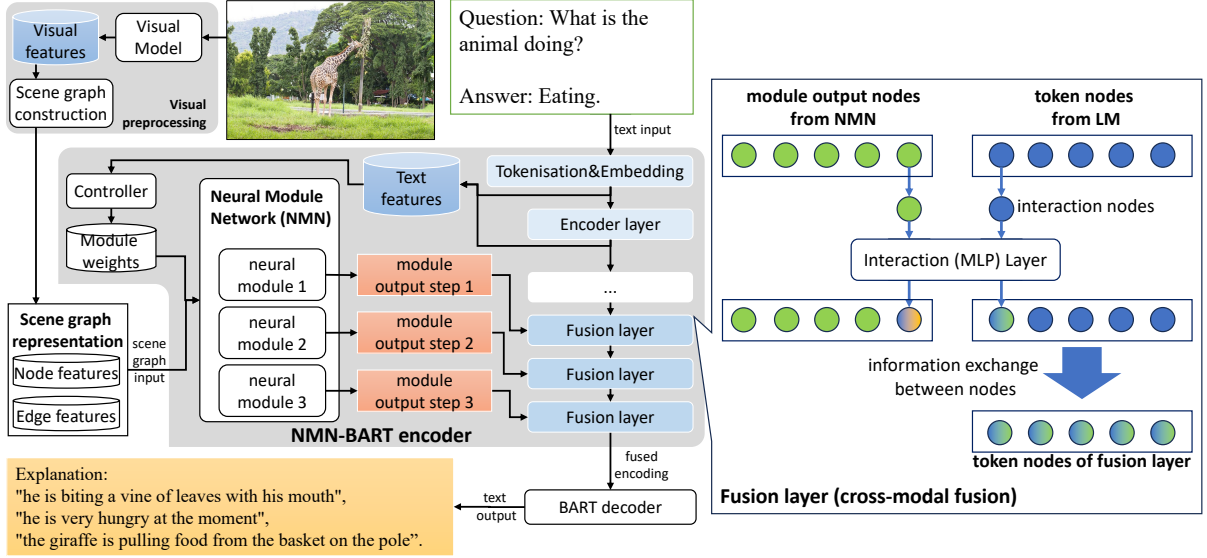


Figure 2: Overview of NMN-BART, which consists of (1) visual preprocessing, (2) an NMN-BART encoder, and (3) a BART decoder. The visual preprocessing converts the image into a scene graph representation. The NMN-BART encoder takes the scene graph along with the question and answer as input, producing a fused encoding through an NMN and cross-modal fusion, which is then processed by the BART decoder to generate the explanation.

NMN processes these inputs to produce intermediate outputs \mathbf{O}^t from the reasoning steps on the scene graph, where t denotes the reasoning step, expressed as:

$$\{\mathbf{O}^1, \mathbf{O}^2, \mathbf{O}^3\} = \text{NMN}(\mathbf{H}^0, \mathbf{H}^1, G) \quad (3)$$

where $\mathbf{O}^t = \{\mathbf{o}_0^t, \mathbf{o}_1^t, \dots, \mathbf{o}_K^t\}$, $t = 1, 2, 3$ represents the intermediate module output at steps 1, 2, and 3, with K being the dimension of the module output.

We choose NMN because it enables fully differentiable training via back-propagation without expert supervision of reasoning steps. Following StackNMN [18] for the modular reasoning process and [10] for reasoning over scene graphs, we summarise the key components of StackNMN here, and refer technical details to [18]. StackNMN consists of three components: 1) The Layout Controller converts text information into a temporal distribution over module weights, segmenting reasoning into steps. 2) Module weights are assigned to a sequence of neural modules, each designed to perform a reasoning step. These modules include *Find*, *Transform*, *And*, *Or*, *Filter*, *Scene*, *Answer*, *Compare*, and *NoOp*, and operate on scene graph representations. The module outputs are visual attention maps or score vectors over possible answers. 3) A differentiable memory stack stores and accesses intermediate module outputs during execution.

Cross-modal fusion. The fusion layers integrate visual and textual information from the NMN (module outputs) with the hidden states from the previous encoder or fusion layer. Each fusion layer corresponds to one neural module reasoning step. For the l -th fusion layer (right side of Figure 2), we concatenate the first token node from the previous layer \mathbf{h}_0^{l-1} with the last token node of the module output \mathbf{o}_K^{t-1} , and feed it into an MLP (Eq. 4). The resulting fused token $\tilde{\mathbf{h}}_0^l$ is then propagated to all nodes in the fusion layer output through the self-attention mechanism of the transformer (Eq. 5, Eq. 6). One fusion layer can be summarised in Eq. 7. The final output of the fusion layers is the fused encoding \mathbf{H}^L .

$$[\tilde{\mathbf{h}}_0^{l-1}; \tilde{\mathbf{o}}_K^{t-1}] = \text{MLP}([\mathbf{h}_0^{l-1}; \mathbf{o}_K^{t-1}]) \quad (4)$$

$$\tilde{\mathbf{H}}^{l-1} = [\tilde{\mathbf{h}}_0^{l-1}, \mathbf{h}_1^{l-1}, \dots, \mathbf{h}_N^{l-1}] \quad (5)$$

$$\mathbf{H}^l = \text{SelfAttention}(\tilde{\mathbf{H}}^{l-1}) \quad (6)$$

$$\mathbf{H}^l = \text{Fusion-Layer}([\mathbf{H}^{l-1}; \mathbf{O}^{t-1}]) \quad (7)$$

BART decoder. The fused encoding \mathbf{H}^L from the last fusion layer is passed to the BART decoder to generate explanations: $E = \text{BART-Decoder}(\mathbf{H}^L)$.

Training scheme. We initialise the encoder, fusion, and decoder layers with a pretrained BART model. Visual preprocessing (scene graph generation) is computed in advance. The entire NMN-BART encoder and the BART decoder are trained end-to-end using cross-entropy loss [24], with inputs Q , A , precomputed G , and reference explanation E .

5. Experiment

Dataset. We evaluate our method on the VQA-X dataset [5], an established benchmark that extends the Visual Question Answering (VQA) dataset with human-written explanations. VQA-X contains 33k question-answer pairs over 28k images sourced from the MSCOCO dataset [26]. Each question averages 7.5 words, and each explanation around 11 words, with a vocabulary size of approximately 10k. The data is split into training (29k), validation (1.4k), and test (1.9k) sets. Each question may have multiple valid answers. The scale and diversity of VQA-X make it well suited for assessing both answer accuracy and the quality of generated explanations.

Baselines. We compare our NMN-BART model against representative baselines, categorised by their use of answer conditioning (more discussion see Section 3). *Not answer-conditioned*: NLX-GPT [6] generates explanations without relying on the predicted answer. Other methods are *answer-conditioned*, where RVT [3] and our model NMN-BART are *with a given answer* to guide explanation generation, and the other methods apply a *filtered setting*: filtering those explanations where answers are correctly predicted by the model, assuming that explanations supporting incorrect answers are invalid and should be excluded from evaluation [6]. These methods include PJ-X [5], FME [4], e-UG [9], and S³C [8], where most methods condition the explanation directly on answers, and in S³C the explanation is rewarded by the correct answer using reinforcement learning.

Automatic evaluation metrics. We use two types of automatic metrics for explanation evaluation:

N-gram-based metrics: These metrics assess lexical-level similarity, measuring word and phrase overlap without considering deeper meaning. We use BLEU-1 [27] to measure the unigram precision, capturing word overlap between the generated and reference texts, with a brevity penalty for short outputs. ROUGE-L [28] uses the Longest Common Subsequence to assess sentence-level structural similarity. Both are primarily sensitive to n-gram overlap.

Semantic-based metrics: These metrics evaluate a deeper semantic alignment between the generated and reference text. METEOR [13] goes beyond n-gram matching by incorporating stemming, synonym matching, and paraphrase tables. SPICE [14] converts both explanations into scene graphs and evaluates the alignment of objects, attributes, and relationships.

Implementation. We use the pretrained BART-base from Facebook [29] as our language backbone. The encoder consists of 6 layers, with 3 of them configured as fusion layers corresponding to the three NMN modules. The model is trained for 15 epochs (17,730 steps) using a batch size of 24 and a learning rate of $2e-5$, on a single Quadro GV100 GPU for approximately 4.8 hours. For cross-modal fusion, we apply a one-hidden-layer MLP that projects the concatenation of token nodes from LM and NMN outputs to a 288-dimensional space, with a dropout rate of 0.2.

5.1. Results and Discussion

Results. The results of all methods on the VQA-X dataset are summarised in Table 1. The methods are categorised first by whether the explanation generation is *Not answer-conditioned* or *Answer-conditioned*. Among the latter, the methods are further categorised by applying the *filtered setting*: filtering explanations only when the answers are correct [8] or *with given answer*: a given answer is an input for generating the explanation. It can be observed that recent methods such as NLX-GPT and S³C, achieve high scores on BLEU-1 and ROUGE-L, indicating strong lexical-level similarity with the reference explanations. Our approach obtains substantially higher scores on semantic-based metrics, achieving a METEOR of 38.5 and a SPICE of 25.9, with respective improvement of 61% and 13% compared

Table 1

NMN-BART shows substantially stronger performance on semantic-based metrics, despite lower scores on n-gram-based metrics. These results suggest that our method better captures the underlying semantic of the explanations, even when the generated text differs in surface from the reference. This is supported by our case studies and human evaluations discussed in Section 5.2. **Bold:** best; underline: second-best.

Method			N-gram-based		Semantic-based	
			BLEU-1	ROUGE-L	METEOR	SPICE
Not answer-conditioned		NLX-GPT [6]	<u>64.2</u>	<u>51.5</u>	23.1	22.1
Answer-conditioned	Filtered setting	FME [4]	59.1	47.1	20.4	18.4
		PJ-X [5]	57.4	46.0	19.7	17.1
		S ³ C [8]	64.7	52.1	<u>23.9</u>	<u>23.0</u>
		e-UG [9]	57.3	45.7	22.1	20.1
	With given answer	RVT [3]	51.9	42.1	19.2	15.8
		NMN-BART (Ours)	36.8	31.3	38.5	25.9

to the second best baseline S³C. On the other hand, NMN-BART has BLEU-1 (36.8) and ROUGE-L (31.3) scores that are notably lower than many baselines.

Discussion. By comparing the two types of metrics across all methods, we can conclude that semantic-based metrics are generally more challenging than n-gram-based metrics. On average, most methods score high on n-gram-based metrics but relatively low on semantic-based ones (100 means perfect.).

The n-gram-based metrics primarily focus on word or phrase overlap. BLEU-1 captures unigram overlap, while ROUGE-L focusses on the longest subsequence. From the results, we can infer that NLX-GPT and S³C generate explanations that closely overlap with the reference, while NMN-BART performs less well on lexical overlap. However, this does not imply worse performance for NMN-BART; rather, it reflects that NMN-BART generates more different lexical content than the reference.

METEOR extends n-gram overlap by incorporating semantic alignment. It also considers recall and penalises brevity, meaning that explanations lacking key information or being overly short will score lower. METEOR thus balances lexical similarity and semantic alignment, capturing subtleties that n-gram metrics might miss. Observing that methods such as NLX-GPT and S³C score higher on n-gram-based metrics but lower on METEOR, we can postulate it is because their explanations tend to be shorter or omit important details. While NMN-BART generates explanations that are relatively longer and cover more important information overlapping with the reference.

SPICE, designed for evaluating image captions, constructs scene graphs from both generated and reference texts, comparing entities, relationships, and attributes. It considers both precision and recall, and is robust to lexical variation. As the most challenging metric, SPICE scores tend to be low in all methods. Methods with high n-gram-based scores also tend to perform poorly on SPICE, probably because they generate shorter explanations with less coverage of key information. Despite scoring relatively high on METEOR, NMN-BART also exhibits a relatively low SPICE score. It can be that the specific way SPICE constructs scene graphs affects the evaluation, or it may indicate that the explanations of NMN-BART contain more information than the reference, affecting the alignment.

In conclusion, NMN-BART demonstrates superior performance in generating semantically aligned explanations compared to the baselines. To verify the underlying reasons we postulated, we conduct case studies and human evaluations (Section 5.2).

5.2. Case Study with Human Evaluation

Case study. To further understand the automatic evaluation results and verify our interpretation, we analyse several representative cases (Figure 3). These cases show that NMN-BART indeed generates explanations with good semantic alignment with references, despite mismatches at the lexical level. In some cases, the generated explanations even contain more relevant information than the reference.

Case 1: Skateboarding. For the question “What is the man doing?” with the answer “skateboarding,” the reference explanations describe actions such as riding, balancing, or performing tricks on a skateboard. Our model-generated explanations, while differing in wording, capture key information by describing trick performance and movement on a ramp. The generated explanations also provide additional



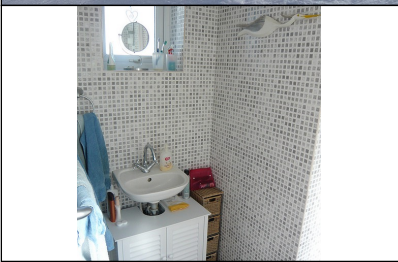
	<p>Question: "What is the man doing?" Answer: "skateboarding"</p> <p>Reference Explanations: "he is riding his skateboard on a skate rail, he is doing a trick on his board, he is balancing on a board on top of a rail"</p> <p>Predicted Explanations: "he is doing a trick on a skateboard. he is jumping a trick off the edge of a ramp on top of a board that has wheels attached to his skateboard. he is at the edge and he is on the edge of a ramp"</p>
	<p>Question: "What are the people doing?" Answer: "snowboarding"</p> <p>Reference Explanations: "it is their favorite sport, they all are riding on a snowboard, they are on a snow-filled mountain with boards under their feet"</p> <p>Predicted Explanations: "they are on a snowboard going down a snowy hill. they have snowboards strapped to their feet and are on snowboards. they're attached to boards and are touching the snow"</p>
	<p>Question: "Is the sink clean?" Answer: "clean"</p> <p>Reference Explanations: "ther is nothing in the sink, it is white and does not have any residue inside it, it is sparkling white with no dirty spots"</p> <p>Predicted Explanations: "there is no dirt on the sink and the sink is covered with no dirt or grime on it. there is not dirt or dirt or stains in the sink. there are no dirt of the sink in the bowl of it, next to the toilet."</p>

Figure 3: Case study: image, question, answer, reference explanations, and the predicted explanations by MNM-BART. a. Case 1: Skateboarding; b. Case 2: Snowboarding; c. Case 3: Sink Cleanliness.

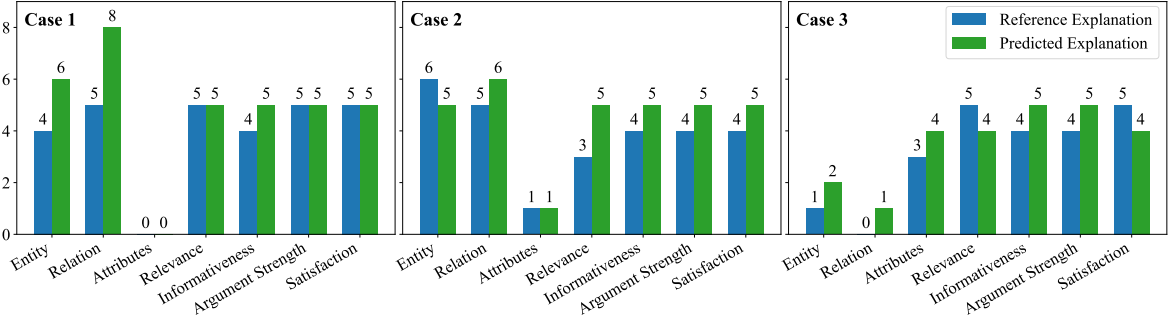


Figure 4: Human evaluation scores comparing reference and predicted explanations across three cases using seven metrics (metrics definition see Section 5.2). Higher scores reflect better performance.

information, such as details describing the trick: jumping, edge, ramp, wheels. In particular, the predicted explanation contains more objects not in the reference such as *ramp*, *wheels*, *edge*, and additional relations, e.g., in *(he, is_jumping, trick)* and *(trick, off, edge)*.

Case 2: Snowboarding. When asked “What are the people doing?” with the answer “snowboarding,” the reference explanations focus on the general activity of riding snowboards on a snowy mountain. Our model similarly identifies the key elements of the scene, describing that the individuals are moving downhill with snowboards strapped to their feet. Notably, the reference explanation “it is their favorite sport” cannot be directly seen from the image. It is a rather subjective interpretation. The predicted explanations has more details, such as *(they, going_down hill)* and *(they, are_touching, snow)*.

Case 3: Sink Cleanliness. For the question “Is the sink clean?” with the answer “clean,” the reference explanations describe the absence of dirt or residue. A key challenge here is handling negation. If we allow negation in the predicate, such as *(sink, covered_with_no, dirt)*, this would lead to an infinite number of possible tail entities. To address this, we treat *covered_with_no_dirt* as a relevant attribute in this context, which captures the intended meaning of cleanliness, rather than treating dirt or grime as entities. With this design, the generated explanation is semantically correct and aligned with the reference in detecting relevant attributes, even though it provides extra contextual details, such as mentioning a toilet that is not directly visible in the image.

Human evaluation metrics for case study. We adapt human evaluation metrics from [30] combining content-based and subjective metrics, as outlined below:

- *Entity*: the number of relevant distinct entities in the explanations.
- *Relation*: the number of relevant distinct relations in the explanations.
- *Attributes*: the number of relevant distinct attributes in the explanations.
- *Relevance*: the explanation only contains relevant information about the question and answer, a subjective score ranging 0-5 with lower number penalising superfluous information.
- *Informativeness*: the explanation adds additional relevant information beyond the Q&A, a subjective score ranging 0-5 with higher number indicating more relevant information.
- *Arguments strength*: the degree to which the explanation supports the answer. It reflects the number or strength of the correct arguments in the explanation with the values, a subjective score ranging 0-5 with higher number indicating stronger arguments.
- *Satisfaction*: subjective satisfaction about the explanation, a subjective score ranging 0-5 with higher number indicating higher satisfaction.

The human evaluation results comparing the predicted and reference explanations across three cases using both content-based and subjective metrics are presented in Figure 4. For content-based coverage (Entity, Relation, Attributes), the predicted explanations capture more or comparable relevant information compared to the references, particularly excelling in Entity and Relation. Subjective evaluations of explanation quality, including Relevance, Informativeness, Argument Strength, and Satisfaction, show that predicted explanations achieve competitive or superior performance in most cases, with multiple perfect scores (5) in these metrics. Based on these results, we can conclude that our approach produces explanations that are semantically rich and have high human satisfaction (4 to 5), even when the generated text deviates from reference explanations at the lexical level.

6. Conclusion and Outlook

In this paper, we introduce our ongoing work on NMN-BART, a novel architecture that combines Neural Module Network with BART to generate natural language explanations for visual question answering. Our model leverages compositional reasoning over scene graphs to capture deeper semantic relationships between the image, question, and answer, producing explanations that are semantically rich, persuasive, and with high human satisfaction. Experiments on the VQA-X dataset demonstrate that our method significantly outperforms baselines in capturing semantic content, despite lower lexical alignment with the references. Future work will focus on testing NMN-BART on additional datasets, performing larger-scale human evaluations, developing automatic metrics that better align with human rationales, ultimately contributing to more transparent and interpretable AI.

Declaration on Generative AI. The authors have used ChatGPT to assist with the polishing of human-authored text. The authors take full responsibility for the publication’s content.

References

- [1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, et al., VQA: Visual Question Answering, in: ICCV, IEEE, 2015, pp. 2425–2433.
- [2] R. Dua, S. S. Kancheti, V. N. Balasubramanian, Beyond VQA: Generating Multi-word Answers and Rationales to Visual Questions, in: CVPRW, IEEE, 2021, pp. 1623–1632.
- [3] A. Marasović, C. Bhagavatula, J. S. Park, R. L. Bras, et al., Natural Language Rationales with Full-Stack Visual Reasoning: From Pixels to Semantic Frames to Commonsense Graphs, in: Findings EMNLP, ACL, 2020, pp. 2810–2829.
- [4] J. Wu, R. J. Mooney, Faithful Multimodal Explanation for Visual Question Answering, in: ACL Workshop, ACL, 2019, pp. 103–112.
- [5] D. H. Park, L. A. Hendricks, Z. Akata, A. Rohrbach, et al., Multimodal Explanations: Justifying Decisions and Pointing to the Evidence, in: CVPR, IEEE, 2018, pp. 8779–8788.

- [6] F. Sammani, T. Mukherjee, N. Deligiannis, NLX-GPT: A Model for Natural Language Explanations in Vision and Vision-Language Tasks, in: CVPR, IEEE, 2022, pp. 8312–8322.
- [7] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, et al., Learning Transferable Visual Models From Natural Language Supervision, in: ICML, PmLR, 2021, pp. 8748–8763.
- [8] W. Suo, M. Sun, W. Liu, Y. Gao, et al., S3C: Semi-Supervised VQA Natural Language Explanation via Self-Critical Learning, in: CVPR, IEEE, 2023, pp. 2646–2656.
- [9] M. Kayser, O.-M. Camburu, L. Salewski, C. Emde, et al., e-ViL: A Dataset and Benchmark for Natural Language Explanations in Vision-Language Tasks, in: ICCV, IEEE, 2021, pp. 1224–1234.
- [10] J. Shi, H. Zhang, J. Li, Explainable and Explicit Visual Reasoning over Scene Graphs, in: CVPR, IEEE, 2019.
- [11] J. Andreas, M. Rohrbach, T. Darrell, D. Klein, Neural Module Networks, in: CVPR, IEEE, 2017, pp. 39–48.
- [12] J. Johnson, B. Hariharan, L. Van Der Maaten, J. Hoffman, et al., Inferring and Executing Programs for Visual Reasoning, in: ICCV, IEEE, 2017, pp. 3008–3017.
- [13] S. Banerjee, A. Lavie, METEOR: An automatic metric for MT evaluation with improved correlation with human judgments, in: ACL Workshop, ACL, 2005, pp. 65–72.
- [14] P. Anderson, B. Fernando, M. Johnson, S. Gould, SPICE: Semantic propositional image caption evaluation, in: ECCV, IEEE, 2016, pp. 382–398.
- [15] J. Wei, X. Wang, D. Schuurmans, M. Bosma, et al., Chain-of-thought prompting elicits reasoning in large language models, in: NeurIPS, volume 35, 2022, pp. 24824–24837.
- [16] X. Wang, A. Amayuelas, K. Zhang, L. Pan, et al., Understanding reasoning ability of language models from the perspective of reasoning paths aggregation, in: ICML, PMLR, 2024, pp. 50026–50042.
- [17] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, et al., CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning, in: CVPR, IEEE, 2017, pp. 2901–2910.
- [18] R. Hu, J. Andreas, T. Darrell, K. Saenko, Explainable neural computation via stack neural module networks, in: ECCV, IEEE, 2018, pp. 53–69.
- [19] D. Suris, S. Menon, C. Vondrick, ViperGPT: Visual Inference via Python Execution for Reasoning, in: CVPR, IEEE, 2023, pp. 11888–11898.
- [20] T. Gupta, A. Kembhavi, Visual Programming: Compositional visual reasoning without training, in: CVPR, IEEE, 2023, pp. 14953–14962.
- [21] H. You, R. Sun, Z. Wang, L. Chen, et al., IdealGPT: Iteratively Decomposing Vision and Language Reasoning via Large Language Models, in: Findings EMNLP, ACL, 2023, pp. 11289–11303.
- [22] C. Mitra, B. Huang, T. Darrell, R. Herzig, Compositional Chain-of-Thought Prompting for Large Multimodal Models, in: CVPR, IEEE, 2024, pp. 14420–14431.
- [23] J. Shi, H. Zhang, J. Li, Explainable and explicit visual reasoning over scene graphs, in: CVPR, IEEE, 2019, pp. 8376–8384.
- [24] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, et al., BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, in: ACL, 2020, pp. 7871–7880.
- [25] P. Anderson, X. He, C. Buehler, D. Teney, et al., Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering, in: CVPR, IEEE, 2018, pp. 6077–6086.
- [26] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, et al., Microsoft COCO: Common objects in context, in: ECCV, IEEE, 2014, pp. 740–755.
- [27] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, BLEU: a method for automatic evaluation of machine translation, in: ACL, 2002, pp. 311–318.
- [28] C.-Y. Lin, Rouge: A package for automatic evaluation of summaries, in: Text summarization branches out, ACL, 2004, pp. 74–81.
- [29] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, et al., Huggingface bart-base code, 2025. URL: <https://huggingface.co/facebook/bart-base>, accessed 11 April 2025.
- [30] Y. Zhou, B. Zhou, H. Li, Q. Lyu, et al., Dataset for industrial question answering with explanation and scalable ensemble generation, in: WWW Companion, ACM, 2025, pp. 825–828.