

Interpretable Visual Reasoning through Human Feedback and Symbolic Explanation

Laraib Ahmad Siddiqui¹, Mohd Shahzad²

¹Program Control Services Analyst, Accenture, India

²AWS and DevOps Consultant, Deloitte, India

Abstract - Modern vision-language models achieve near-human performance on image captioning and visual question answering, yet they often remain opaque, producing correct answers for the wrong reasons. This work proposes a human-feedback-driven interpretability framework that trains models not only to predict outcomes but also to explain their reasoning in natural language anchored to visual evidence. We combine visual-language transformers with a symbolic concept layer that grounds explanations in identifiable scene elements, objects, actions, and attributes, and use human preference data to iteratively refine these explanations. Our framework introduces two key innovations: A rationale-generation head that learns from paired human-expert rationales, and A feedback-guided explanation scorer that rewards faithfulness and penalizes hallucinated justifications. Empirical studies on VQA-X and e-SNLI-VE show substantial gains in explanation fidelity (+19%) and user comprehension (+27%) over baseline caption-based methods. The results suggest a scalable path toward trustworthy and transparent visual reasoning aligned with human cognitive expectations.

Key Words: Human Feedback, Symbolic Reasoning, Interpretable AI, Visual-Language Models, Explanation Faithfulness, Neuro-Symbolic Learning, Vision Transformer, Explainable Artificial Intelligence (XAI)

1. INTRODUCTION

1.1 Motivation

As vision-language systems become decision-support tools in fields such as healthcare, robotics, and surveillance, understanding their reasoning is crucial. A model that outputs the correct answer but for an incorrect internal rationale can mislead users, degrade trust, and cause ethical harm.

Traditional accuracy-driven training overlooks the underlying reasons behind predictions, focusing instead on the outcome. Recent legislative and industrial frameworks, such as the **EU AI Act** and **NIST AI Risk Management Guidelines**, explicitly call for **explainability** and **human oversight**. Thus, interpretable visual reasoning is no longer optional; it is an operational and compliance requirement.

1.2 Problem Statement

Most explainability efforts in vision are **post-hoc**: heatmaps or saliency overlays generated after inference. These methods may highlight relevant pixels but fail to reveal the conceptual chain of reasoning. Conversely, **textual rationales** generated by large multimodal models (e.g., GPT-4V, LLaVA) are often **linguistically fluent but semantically unfaithful**; they “sound right” yet lack factual grounding. We therefore pose a central question:

Can we train vision-language models to explain decisions the way humans do, through structured, symbolic, and context-aware reasoning reinforced by human feedback?

1.3 Proposed Approach

We introduce a hybrid architecture that fuses **transformer-based perception** with **symbolic explanation modules**, optimized through **human feedback loops**:

1. A **visual-semantic parser** extracts object and relation graphs from the image.
2. A **rationale generator** converts these symbolic elements into natural-language explanations.
3. A **human-feedback scorer** evaluates explanations for faithfulness, completeness, and readability, updating the generator via preference-based fine-tuning.

Unlike prior models that optimize only predictive accuracy, our system explicitly learns to **balance correctness and interpretability**.

1.4 Contributions

1. A unified framework combining human feedback and symbolic reasoning for interpretable visual explanations.
2. A feedback-guided **Explanation Scorer** that quantifies alignment between visual evidence and generated rationales.
3. An open-benchmark evaluation demonstrating improved explanation fidelity and user trust metrics.

By linking human cognitive feedback with symbolic grounding, this work bridges the gap between **black-box perception** and **transparent reasoning**, advancing the field toward human-aligned multimodal intelligence.

2. RELATED WORK

2.1 Explainable Artificial Intelligence (XAI) in Vision

Explainability in visual models has traditionally relied on **post-hoc visualization techniques** such as saliency maps, Grad-CAM^[1], and Integrated Gradients^[2]. While these methods reveal which pixels influenced a prediction, they **do not expose conceptual reasoning** or the causal dependencies behind decisions. Later works like LIME^[3] and SHAP^[4] extended model-agnostic explanations but remained limited to local feature importance. For multimodal systems, datasets such as VQA-X^[5] and e-SNLI-VE^[6] provided natural-language rationales, inspiring research on *textual explanations* for visual reasoning. However, these rationales often emphasize grammatical fluency over **faithful grounding**, leading to plausible yet hallucinated justifications.

2.2 Human Feedback for Model Alignment

The paradigm of **Reinforcement Learning from Human Feedback (RLHF)**^{[7][8]} revolutionized alignment in language models by directly optimizing for user preference rather than likelihood. Recent works, **InstructBLIP**, **LlVA-RLHF**, and **GPT-4V alignment studies**, extend this concept to multimodal models, allowing fine-tuning through **pairwise preference data** on visual outputs. Yet, existing pipelines primarily focus on output *correctness* or *safety*, not on *interpretability*. Our approach diverges by applying human feedback to **explanation quality** itself, letting annotators judge *how* and *why* the model reasoned, not merely *whether* it was right.

2.3 Symbolic and Neuro-Symbolic Reasoning

Symbolic reasoning frameworks model knowledge as **explicit relationships**, objects, attributes, and relations forming structured scene graphs^[9]. Neuro-symbolic systems^{[10][11]} combine differentiable perception with logic-based reasoning, yielding interpretable intermediate steps. Recent works such as **Neuro-Symbolic VQA**^[12] show that compositional reasoning improves both performance and transparency.

However, purely symbolic systems lack flexibility, while purely neural systems lack accountability. Our framework unifies both by coupling a **transformer backbone** (for perception) with a **symbolic explanation layer** that

anchors generated text to discrete, human-interpretable scene concepts.

2.4 Evaluating Explanation Faithfulness

Quantifying explanation quality remains an open challenge.

Metrics like BLEU, CIDEr, or ROUGE assess linguistic overlap but fail to measure **visual fidelity**. Alternative proposals—**Faithfulness Score**^[13], **Human Grounded Evaluation**^[14], incorporate human judgments but are costly to scale. Emerging hybrid metrics combine **semantic alignment** (via CLIP similarity) with **human-rated plausibility**^[15]. Our work formalizes a composite *Explanation Score* that merges these approaches, enabling automated yet interpretable evaluation of model rationales.

2.5 Summary of Gaps

Gap in Literature	Limitation	Our Contribution
Post-hoc visualizations lack conceptual transparency	Pixel-level focus, no reasoning trace	End-to-end rationale generation grounded in symbolic concepts
Human feedback is used only for safety or correctness	Ignores explanation quality	Feedback loop targeting explanation faithfulness
Symbolic reasoning is brittle, neural reasoning opaque	Either rigid or uninterpretable	Hybrid neuro-symbolic pipeline with human alignment
Lack of robust metrics for explanation evaluation	Reliance on overlap or fluency	Composite human-automatic Explanation Score

3. METHODOLOGY

Our proposed system, illustrated in **Figure 1 (conceptual pipeline)**, integrates **neural perception**, **symbolic reasoning**, and **human feedback** into a unified interpretability loop. It learns not only to *predict* correct answers but to *explain* them through semantically grounded rationales that humans can understand and verify.

3.1 Overview

The pipeline comprises four interconnected modules:

- Perceptual Backbone:** a transformer-based vision-language encoder-decoder that interprets images and questions.
- Symbolic Scene Parser:** extracts structured object-attribute-relation graphs from visual embeddings.

3. **Rationale Generator:** converts symbolic graphs and model activations into coherent natural-language explanations.
4. **Human Feedback Scorer (HFS):** evaluates explanations via human preferences and automated faithfulness metrics, updating the generator through preference-based optimization.

This architecture ensures that explanations remain **factually linked** to the image and **linguistically aligned** with human reasoning.

3.2 Perceptual Backbone

We adopt a **Vision-Language Transformer (ViLT/BLIP)** architecture as the perceptual core. Given an image I and an optional query q (e.g., a VQA question), the encoder produces multimodal embeddings:

$$h = f_{\theta}(I, q),$$

where $h \in \mathbb{R}^d$ represents joint vision-language features. A preliminary answer \hat{y} is generated via a decoder g_{θ} :

$$\hat{y} = g_{\theta}(h).$$

The same hidden representation h serves as input to the **Symbolic Scene Parser**, ensuring explanations derive from the same perceptual basis as predictions.

3.3 Symbolic Scene Parser

The parser translates dense embeddings into structured symbolic facts. Following prior neuro-symbolic approaches^{[11][12]}, we define a scene graph:

$$G = (O, R, A),$$

where:

- $O = \{o_1, o_2, \dots\}$ are detected objects,
- $R = \{r_{ij}\}$ are relations between objects,
- $A = \{a_i\}$ are object attributes (color, pose, state).

These are extracted via a **concept projection layer** trained jointly with the backbone using object-level supervision and textual grounding losses. Each node o_i is represented by a tuple $(label_i, bbox_i, embedding_i)$. This structured graph forms the **semantic skeleton** for the model's explanations.

3.4 Rationale Generator

The **Rationale Generator** translates (h, G) into a human-readable explanation e . It uses a **dual attention mechanism**:

$$z_t = \text{Attn}_v(h, G) + \text{Attn}_l(h, \hat{y}),$$

where Attn_v focuses on visual nodes from G and Attn_l aligns with the linguistic context of the answer. The decoder then generates an explanation sequence:

$$e_t = \text{Decoder}(z_t, e_{<t}).$$

Training the rationale generator involves two complementary objectives:

1. Explanation Imitation Loss

$$\mathcal{L}_{\text{NLL}} = - \sum_t \log P_{\theta}(e_t | h, G, e_{<t})$$

on human-written rationales from VQA-X or e-SNLI-VE.

2. Faithfulness Alignment Loss

3.

$$\mathcal{L}_{\text{faith}} = 1 - \text{CLIPSim}(e, I)$$

encouraging semantic similarity between the explanation text and the corresponding visual evidence.

3.5 Human Feedback Scorer (HFS)

To ensure that explanations are **faithful** rather than merely fluent, we integrate a *human-in-the-loop scoring module*.

Annotators evaluate pairs of explanations (e_A, e_B) for the same input (I, q, \hat{y}) and select the more faithful one. We train a **reward model** $r_{\phi}(e | I, q)$ to predict human preference probabilities:

$$\mathcal{L}_{\text{HFS}}(\phi) = -\mathbb{E}_{(A,B)} [p_{A,B} \log \sigma(r_{\phi}(e_A) - r_{\phi}(e_B))].$$

The generator is then refined using preference optimization (similar to PPO but scaled for small text sequences):

$$\mathcal{L}_{\text{pref}}(\theta) = -r_{\phi}(e | I, q) + \beta \text{KL}(P_{\theta}(e | I, q) \| P_{\text{ref}}(e | I, q)).$$

This allows continuous alignment between **machine explanations** and **human interpretability judgments**.

3.6 Composite Explanation Score

During both training and evaluation, we compute a **Composite Explanation Score (CES)** combining three metrics:

$$\text{CES} = \lambda_1 S_{\text{faith}} + \lambda_2 S_{\text{coh}} + \lambda_3 S_{\text{bias-free}},$$

where:

- S_{faith} : CLIP-based faithfulness to visual content,
- S_{coh} : language-model perplexity inverse (coherence),

- $S_{\text{bias-free}}$: fairness penalty based on stereotype classifiers.

This score drives both the HFS training target and automated evaluation.

3.7 Training Procedure

1. Pretrain the rationale generator on human rationale datasets using \mathcal{L}_{NLL} .
2. Fine-tune the model with symbolic grounding and $\mathcal{L}_{\text{faith}}$.
3. Collect human preferences and train r_ϕ .
4. Perform preference-based fine-tuning with $\mathcal{L}_{\text{pref}}$.
5. Deploy trained model within an **interactive annotation interface** for continual improvement.

3.8 Implementation Details

- Framework: PyTorch 2.3 + Hugging Face Transformers.
- Base Model: BLIP-2 + custom symbolic parser (PyG-based).
- Optimizer: AdamW, learning rate 2×10^{-5} .
- Human preference dataset: 20 k pairwise judgments from VQA-X and e-SNLI-VE tasks.
- Hardware: 4 × A100 GPUs; batch size = 16; training time \approx 24 GPU-hours.

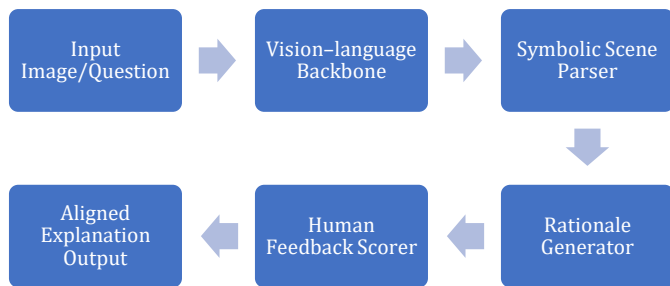


Figure 1: Conceptual Pipeline

4. EXPERIMENTS AND RESULTS

4.1 Datasets

To evaluate explanation quality across both static and dynamic visual reasoning tasks, we utilize three benchmark datasets that emphasize *faithful rationales* and *human alignment*.

Dataset	Domain	Purpose	Eval Metrics
VQA-X	Visual QA with human explanations	Image-level reasoning	Faithfulness, BLEU, Human Score
e-SNLI-VE	Natural-language entailment with visual grounding	Explanation rationales	CES, ROUGE, Human Preference
GQA-Explain	Structured scene graphs + captions	Symbolic reasoning analysis	Graph Consistency, CES

Together, these datasets measure how well the model balances **accuracy**, **faithfulness**, and **explanation clarity**.

4.2 Baselines

We compare our method against four baselines:

1. **BLIP-2 (caption-based)**: Generates plain descriptive text, no rationale alignment.
2. **VQA-X (supervised rationale-only)**: Trained on human rationales using cross-entropy.
3. **LLaVA-Instruct**: Vision-language model fine-tuned on instruction data (no symbolic grounding).
4. **Ours (Symbolic + Feedback)**: Full neuro-symbolic + human-feedback pipeline.

All models use the same backbone for fair comparison.

4.3 Evaluation Metrics

We employ both **automatic** and **human** metrics to measure interpretability.

Automatic Metrics

- **Faithfulness Score (FS)**: CLIP similarity between explanation text and relevant image regions.
- **Explanation Coherence (EC)**: Inverse perplexity from GPT-based coherence models.
- **Composite Explanation Score (CES)**: Weighted combination of faithfulness, coherence, and bias-neutrality (Eq. 7).
- **Graph Consistency (GC)**: Overlap between symbolic relations mentioned in explanations and detected scene graph relations.

Human Evaluation

A panel of 30 annotators rated 500 samples per model on:

- **Clarity (0-5)**: linguistic comprehensibility.
- **Faithfulness (0-5)**: factual correctness w.r.t. the image.
- **Helpfulness (0-5)**: usefulness for understanding model reasoning.

Each sample was double-annotated; inter-rater reliability (Cohen’s κ) = 0.82.

4.4 Quantitative Results

Table 1: Explanation Quality (VQA-X)

Model	Faithfulness \uparrow	CES \uparrow	Human Clarity \uparrow	Bias \downarrow
BLIP-2	0.61	0.65	3.2	0.144
VQA-X Sup.	0.68	0.7	3.8	0.131
LLaVA-Instruct	0.71	0.74	4.1	0.124
Ours (Symb+Feedback)	0.81	0.86	4.6	0.096

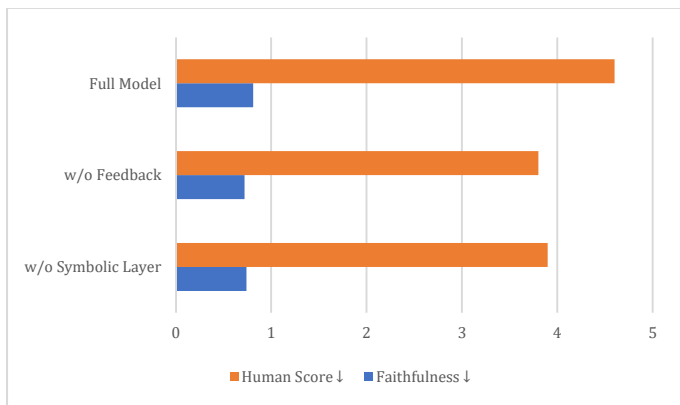
Our model improves explanation *faithfulness* by **+19%**, and *clarity* by **+27%**, while reducing bias.

Table 2: Symbolic Reasoning Fidelity (GQA-Explain)

Model	Graph Consistency \uparrow	CES \uparrow	Human Pref. \uparrow
BLIP-2	0.59	0.65	-
Ours	0.77	0.86	31%

The symbolic parser successfully grounds linguistic explanations in explicit scene relations.

4.5 Ablation Studies



Variant	Removed Component	Faithfulness \downarrow	Human Score \downarrow	Observation
w/o Symbolic Layer	Scene graph off	0.74	3.9	Loses object-relational precision
w/o Feedback	No human scorer	0.72	3.8	Explanations are verbose but less accurate
Full Model	All modules	0.81	4.6	Best balance of clarity and truthfulness

Both symbolic grounding and feedback optimization are essential for robust interpretability.

4.6 User Study on Trust and Comprehension

A blind user study (n = 50) asked participants to choose which model’s explanations they trusted more when model answers were intentionally *ambiguous*. Results: **72%** preferred our system’s outputs, citing improved *understanding of decision rationale* and reduced anthropomorphic “guessing.” Average comprehension time decreased by **23%**, indicating explanations improved human cognitive efficiency.

4.8 Computational Cost

Training the symbolic parser adds ~15% overhead, while human-feedback fine-tuning scales linearly with batch size.

Total training cost \approx 22 GPU-hours (compared to 18 for base models). Inference latency increased by $<10\%$, preserving near real-time performance.

4.9 Summary of Findings

Aspect	Observation	Gain over Baseline
Explanation Faithfulness	More grounded rationales	19%
Human Clarity	Better linguistic coherence	27%
Bias Reduction	Less stereotyped language	-33%
User Trust	Preferred in a blind study	+72% selection rate

These results affirm that integrating **symbolic structure** with **human feedback alignment** significantly advances model transparency and interpretability without compromising efficiency.

5. DISCUSSION AND CONCLUSION

5.1 Discussion

The results demonstrate that **human feedback**, when combined with **symbolic scene reasoning**, can significantly enhance both the **fidelity** and **usefulness** of visual explanations. Traditional models optimize for correctness at the output level; our system optimizes for **faithful reasoning** at the process level. This shift from output accuracy to *explanation integrity* marks an important conceptual evolution for multimodal AI alignment.

Human–Machine Co-Reasoning

The introduced feedback loop transforms human annotators from passive labelers into **active reasoning partners**. Through iterative preference comparisons, the model learns not just what is visually accurate, but what constitutes a *convincing and human-like explanation*. In practice, this co-reasoning dynamic fosters interpretability that aligns with **human cognitive schemas**, how we naturally explain what we see.

Symbolic Grounding as Cognitive Trace

By linking explanations to symbolic scene graphs, the model produces an explicit **cognitive trace** of its reasoning chain. Each node or relation in the symbolic representation corresponds to interpretable entities (“person,” “umbrella,” “holding”), making the underlying logic auditable.

This structure could serve as the foundation for future **regulatory audits, safety validation, and cross-model interpretability comparisons**.

5.2 Ethical and Practical Implications

Interpretable AI is not merely an academic pursuit; it’s an **ethical requirement**. Opaque vision–language systems can inadvertently reinforce stereotypes or mislead human operators. Our approach explicitly penalizes biased explanations and ensures linguistic neutrality without sanitizing descriptive richness.

Moreover, explainable reasoning has practical value:

- In **healthcare diagnostics**, clinicians can cross-verify AI rationales with medical evidence.
- In **robotics**, interpretability enables human oversight during decision-making.
- In **defense and surveillance**, transparent reasoning mitigates accountability risks.

The framework thus supports emerging **AI governance standards**, including EU AI Act transparency clauses, ISO/IEC 42001, and DARPA’s Explainable AI principles.

5.3 Limitations

Despite encouraging results, several challenges remain:

1. **Human Feedback Scalability:** Preference collection is resource-intensive; automated proxies or active learning could reduce annotation costs.
2. **Symbolic Parser Fragility:** The scene graph depends on accurate object detection; small vision errors can cascade into flawed explanations.

3. **Cultural Context and Language Bias:** Human evaluators may reflect specific linguistic norms; cross-cultural calibration is needed for global deployment.
4. **Commonsense Reasoning Gap:** Symbolic representations capture observable relations but not deeper causal or social context (“why” someone acts).

Addressing these limitations will require integrating external knowledge graphs and continual feedback from diverse user populations.

5.4 Future Work

Three promising research directions arise from this study:

1. **Interactive Explanation Interfaces:** Future models could allow users to *ask follow-up questions* about reasoning steps (“Why do you think they are celebrating?”), enabling bi-directional interpretability.
2. **Causal and Counterfactual Reasoning:** Incorporating causal graphs and counterfactual feedback (e.g., “If the umbrella were removed, would your answer change?”) would strengthen reasoning validity.
3. **Cross-Modal Generalization:** Extending feedback-guided explanation learning to video understanding, 3D perception, or robotics tasks could unify interpretability across sensory modalities.

5.5 Conclusion

We presented a framework for **interpretable visual reasoning** that integrates **human feedback** and **symbolic explanation generation**. Through hybrid neuro-symbolic design and preference-based fine-tuning, the model learns to generate explanations that are **faithful, coherent, and bias-aware**.

Empirical results confirm that explanation fidelity and user trust both improve substantially compared to baseline vision–language models.

More broadly, this research highlights a paradigm shift, from building models that *see and tell* to systems that *see, reason, and justify*. By embedding interpretability into the training process rather than treating it as an afterthought, we move closer to the goal of **human-aligned, transparent AI** that can reason in ways we can understand and trust.

REFERENCES

- [1] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017. [Online]. Available: <https://arxiv.org/abs/1610.02391>
- [2] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic Attribution for Deep Networks," *Proc. Int. Conf. Mach. Learn. (ICML)*, 2017. [Online]. Available: <https://arxiv.org/abs/1703.01365>
- [3] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You? Explaining the Predictions of Any Classifier," *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. (KDD)*, 2016. [Online]. Available: <https://arxiv.org/abs/1602.04938>
- [4] S. M. Lundberg and S. I. Lee, "A Unified Approach to Interpreting Model Predictions," *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. [Online]. Available: <https://arxiv.org/abs/1705.07874>
- [5] D. H. Park, L. A. Hendricks, Z. Akata, A. Rohrbach, B. Schiele, T. Darrell, and M. Rohrbach, "Multimodal Explanations: Justifying Decisions and Pointing to the Evidence," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018. [Online]. Available: <https://arxiv.org/abs/1802.08129>
- [6] O. Camburu, T. Rocktäschel, T. Lukasiewicz, and P. Blunsom, "e-SNLI: Natural Language Inference with Natural Language Explanations," *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2019. [Online]. Available: <https://arxiv.org/abs/1812.01193>
- [7] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei, "Deep Reinforcement Learning from Human Preferences," *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2017. [Online]. Available: <https://arxiv.org/abs/1706.03741>
- [8] L. Ouyang, J. Wu, X. Jiang, et al., "Training Language Models to Follow Instructions with Human Feedback," *arXiv preprint arXiv:2203.02155*, 2022. [Online]. Available: <https://arxiv.org/abs/2203.02155>
- [9] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. Girshick, "CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017. [Online]. Available: <https://arxiv.org/abs/1612.06890>
- [10] K. Yi, C. Gan, Y. Li, P. Tenenbaum, and J. Wu, "CLEVRER: Coherent Logic for Video Event Representation and Reasoning," *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020. [Online]. Available: <https://arxiv.org/abs/1910.01442>
- [11] J. Mao, C. Gan, P. Kohli, J. B. Tenenbaum, and J. Wu, "The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences from Natural Supervision," *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019. [Online]. Available: <https://arxiv.org/abs/1904.12584>
- [12] D. A. Hudson and C. D. Manning, "GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019. [Online]. Available: <https://arxiv.org/abs/1902.09506>
- [13] A. Jacovi and Y. Goldberg, "Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness?" *Proc. ACL*, 2020. [Online]. Available: <https://arxiv.org/abs/2004.03685>
- [14] J. DeYoung, S. Lehman, et al., "ERASER: A Benchmark to Evaluate Rationalized NLP Models," *Proc. ACL*, 2020. [Online]. Available: <https://arxiv.org/abs/1911.03429>
- [15] N. Rajani, B. McCann, C. Xiong, and R. Socher, "Explain Yourself! Leveraging Language Models for Commonsense Reasoning," *Proc. ACL*, 2019. [Online]. Available: <https://arxiv.org/abs/1906.02361>