Investigating Compositional Reasoning and Systematic Generalization in Visual Question Answering: A Multimodal Transformer Approach

Eve Riskin

***Abstract*—Visual Question Answering (VQA) has achieved impressive benchmark performance with transformer-based multimodal architectures [1], [2]; however, these models exhibit critical limitations in compositional reasoning and systematic generalization [3]. This research investigates whether contemporary VQA systems truly understand visual scenes or merely exploit superficial statistical cues, particularly when processing novel combinations of visual concepts and linguistic structures requiring multi-step inference. We hypothesize that state-of-the-art models will demonstrate substantial performance degradation ($\geq$15% accuracy drop) on compositional tasks and exhibit poor systematic generalization ($\leq$40% of in-distribution performance) under rigorous out-of-distribution evaluation. Employing a mixed-methods experimental design combining diagnostic benchmarking, controlled ablation studies, and interpretability analysis, we will evaluate LXMERT, ViLBERT, and CLIP-based models on CLEVR, GQA challenge splits, and newly constructed synthetic challenge sets that systematically separate visual primitives from linguistic operators. Our methodology integrates neural module networks with transformer backbones, structured scene graph representations, and interpretability analysis via attention visualization and feature attribution. Expected contributions include: (1) a systematic empirical analysis of reasoning failure modes with human performance baselines; (2) a novel multimodal architecture demonstrating $\geq$20% improvement in systematic generalization while enhancing interpretability through attention alignment with human reasoning chains; and (3) an open-source evaluation toolkit with standardized metrics and baseline implementations for community adoption.**

I. Introduction

Visual Question Answering (VQA) represents a critical frontier in artificial intelligence, requiring unified capabilities in visual perception, linguistic understanding, and cross-modal reasoning [1], [4], [5]. While contemporary multimodal transformer architectures achieve compelling benchmark performance [6], these successes belie a fundamental limitation: models predominantly exploit superficial statistical regularities rather than acquiring genuinely composable representations that support systematic reasoning [2]. This research addresses a crucial gap in understanding whether state-of-the-art VQA systems can perform human-like compositional reasoning—the ability to systematically combine known visual concepts and linguistic operators to answer novel, out-of-distribution queries.

The significance of this investigation extends beyond benchmark metrics. As multimodal AI systems increasingly deploy in high-stakes domains from medical imaging to autonomous navigation, their capacity to generalize to unfamiliar scenarios becomes paramount. Current VQA models demonstrate proficiency in object recognition and attribute classification, yet exhibit pronounced deficits on questions requiring relational inference, counting, or logical deduction—capabilities humans accomplish through systematic composition of simpler primitives. This performance asymmetry suggests that high benchmark scores may reflect dataset-specific biases rather than robust reasoning abilities, potentially precipitating failures in real-world deployment where distributional shifts are inevitable.

Systematic generalization, defined as the capacity to understand and produce novel combinations of familiar components [7], [8], constitutes a core hallmark of human cognition that remains elusive for artificial systems. Recent evidence demonstrates that even neural networks with strong in-distribution performance exhibit dramatic generalization failures when evaluated on systematically novel combinations [3], a phenomenon directly relevant to VQA where models encounter question-answer pairs involving visual concepts and linguistic structures not co-occurring during training. Such failures reveal whether architectures learn truly compositional representations or merely memorize spurious correlations.

This proposal investigates three central questions: (1) How do state-of-the-art VQA models perform on compositional questions requiring multi-step visual reasoning compared to simple descriptive queries? (2) To what extent do current architectures exhibit systematic generalization when evaluated on out-of-distribution question-answer pairs involving novel combinations of visual concepts and linguistic structures? (3) Can incorporating explicit reasoning modules and structured knowledge representations enhance both performance on complex compositional tasks and model interpretability?

We hypothesize that state-of-the-art models will demonstrate significant performance degradation ($\geq$15% accuracy drop) on compositional reasoning tasks, exhibit poor systematic generalization ($\leq$40% of in-distribution performance), and that architectural enhancements integrating explicit reasoning pathways with transformer backbones will improve systematic generalization by $\geq$20% while increasing interpretability. Through controlled experiments on diagnostic datasets and novel challenge splits, we will evaluate transformer-based architectures, integrate neural module networks with structured scene graph representations, and develop an open-source evaluation toolkit for systematic generalization assessment. The subsequent sections detail our methodology, expected contributions, and broader impacts.

II. Background

Visual Question Answering (VQA) represents a foundational challenge at the intersection of computer vision and

Manuscript received April 19, 2021; revised August 16, 2021.

natural language processing, requiring systems to provide accurate textual answers to questions about visual content [1]. While initial VQA approaches achieved promising results through sophisticated attention mechanisms and bottom-up/top-down architectures [2] [9], contemporary transformer-based models demonstrate remarkable performance on descriptive queries yet exhibit significant degradation when confronted with compositional questions requiring multi-step inference and relational reasoning. This discrepancy exposes a critical gap between statistical pattern recognition and genuine systematic understanding, suggesting that standard benchmarks may inadvertently reward superficial heuristics rather than robust reasoning capabilities.

The theoretical concept of systematic generalization—the ability to productively combine known primitives into novel structures—provides a rigorous framework for evaluating VQA systems beyond conventional accuracy metrics [7] [3]. Human cognition exhibits strong systematicity, enabling interpretation of entirely novel sentences from familiar linguistic components, yet state-of-the-art VQA models trained on conventional datasets show poor generalization when evaluated on out-of-distribution question-answer pairs involving novel combinations of visual concepts and linguistic operators. This deficiency indicates that current architectures rely heavily on statistical shortcuts and co-occurrence patterns rather than learning truly composable representations that capture underlying visual-linguistic structure, raising concerns about their robustness and reliability.

Recent architectural interventions have attempted to address these limitations through explicit reasoning pathways. Neural module networks decompose questions into sequences of specialized neural programs, while structured scene graph representations provide explicit symbolic grounding of visual relationships. Methods incorporating chain-of-thought reasoning demonstrate that intermediate reasoning steps can enhance performance on complex scientific question answering tasks [10], suggesting that explicit reasoning traces may benefit compositional VQA. However, these approaches remain underexplored in the VQA domain, particularly when integrated with modern transformer backbones. Moreover, the trade-offs between model performance and interpretability have not been systematically quantified, leaving open questions about whether explicit reasoning modules fundamentally alter information processing or merely provide superficial performance gains.

The emergence of large-scale vision-language pretraining and prompting techniques has further complicated the evaluation landscape [11] [12]. While these methods show improved few-shot learning capabilities, their systematic generalization remains poorly characterized, and the absence of standardized evaluation protocols specifically designed to test compositional reasoning and systematicity has hindered direct comparison across architectural paradigms. Existing compositional reasoning datasets such as CLEVR and GQA challenge splits often lack the granularity needed to diagnose specific failure modes in transformer-based systems, limiting progress toward more robust architectures.

This research directly addresses these gaps through three synergistic contributions that map directly to our research questions: (1) a comprehensive empirical analysis quantifying specific reasoning failure modes and systematic generalization deficits in contemporary multimodal transformers, providing diagnostic benchmarks against human performance to answer how current models perform on compositional versus simple queries; (2) novel architectures integrating explicit reasoning modules with structured knowledge representations, predicting $\geq$20% improvement in systematic generalization while enhancing interpretability to address whether architectural enhancements can bridge these gaps; and (3) an open-source evaluation toolkit featuring standardized metrics and challenge datasets for assessing compositional reasoning, enabling community adoption and reproducible research on systematic generalization in VQA systems.

III. Method

Methodology

This research employs a mixed-methods experimental design combining controlled benchmarking, architectural ablation studies, and human-centered interpretability analysis to systematically investigate compositional reasoning failures in Visual Question Answering systems.

3.1 Experimental Design and Data Curation

We adopt a phased experimental protocol to isolate specific reasoning capabilities. Phase 1 involves diagnostic evaluation on established benchmarks: CLEVR for synthetic compositional reasoning, GQA challenge splits for real-world scene understanding, and VQA v2.0 [1] as a baseline performance reference. To rigorously assess systematic generalization, we construct novel train-test splits that enforce separation between visual primitives and linguistic operators [7]. Specifically, we partition data such that test sets contain novel combinations of object-attribute pairs (e.g., ”spotted cylinder”) and relational operators (e.g., ”behind the”) absent from training distributions, preventing models from exploiting statistical shortcuts. Additionally, we generate a synthetic challenge set of 50,000 question-answer pairs using a grammar-based procedural engine that systematically varies compositional depth (1-5 reasoning steps) and vocabulary overlap ratios (0-100%), enabling fine-grained analysis of failure modes. All datasets undergo expert validation by three independent annotators to ensure question-answer pair correctness and reasoning chain annotations, with inter-annotator agreement (Cohen's > 0.8) required for inclusion.

3.2 Model Architectures and Interventions

Three transformer-based baselines are evaluated: LXMERT, ViLBERT, and CLIP-ViT fine-tuned models, selected for their demonstrated strong performance on standard VQA benchmarks [12]. To test our third hypothesis, we develop two enhanced architectures: (1) **Transformer-Module Networks**, which integrate neural module networks with pre-trained transformer backbones, dynamically assembling reasoning subnetworks based on question parse trees through a layout predictor; and (2) **Graph-Reasoning Transformers**, which incorporate structured scene graph encoders alongside visual transformers, enabling explicit relational reasoning through

graph attention mechanisms [2]. Both interventions introduce explicit reasoning pathways while maintaining end-to-end differentiability. Training employs standard cross-entropy loss with curriculum learning, progressively increasing compositional complexity across epochs. Models are implemented in PyTorch with HuggingFace Transformers, trained on 4×A100 GPUs for two weeks per configuration.

3.3 Evaluation and Analysis Framework

Performance is quantified using top-k accuracy, balanced accuracy (to mitigate class imbalance), and consistency metrics across question paraphrases. For systematic generalization assessment, we report **compound divergence scores** measuring performance degradation relative to in-distribution baselines as a function of linguistic and visual novelty: $CDS = (1 - (OOD_{acc}/ID_{acc})) \times 100\%$, where OOD_{acc} represents out-of-distribution accuracy and ID_{acc} represents in-distribution accuracy on matched question templates. Statistical significance is determined via bootstrap resampling with 95% confidence intervals.

Interpretability analysis employs a multi-level approach: (1) attention visualization to examine whether enhanced models attend to semantically relevant image regions, with alignment scores computed against human-annotated reasoning chains; (2) Integrated Gradients to attribute model predictions to specific visual concepts and linguistic tokens; and (3) human evaluation of generated reasoning chains using a 5-point Likert scale assessing logical validity and plausibility, following established protocols for chain-of-thought evaluation [10]. We recruit 20 domain experts to evaluate 500 randomly sampled reasoning chains, measuring inter-rater reliability (Cohen's) to ensure robust assessment. All human subject protocols will follow institutional IRB guidelines for minimal risk research.

3.4 Limitations and Alternative Approaches

Potential limitations include computational resource constraints for large-scale synthetic data generation, the challenge of obtaining high-quality human reasoning annotations, and potential biases in source datasets that may affect OOD generalization. Alternative approaches include leveraging prompting strategies [11] to elicit compositional reasoning from frozen large multimodal models, which we will explore if architectural interventions fail to achieve target performance improvements.

IV. Experimental Setup

Experimental Design

This research employs a three-phase experimental pipeline designed to systematically evaluate compositional reasoning capabilities in VQA models. All experiments will utilize a unified evaluation framework implemented in PyTorch with HuggingFace Transformers, ensuring reproducibility through fixed random seeds (42, 123, 456, 789, 101112), version-controlled datasets, comprehensive logging via Weights & Biases, and Docker containerization.

Phase 1: Diagnostic Benchmarking Analysis Addressing Research Question 1, we conduct controlled experiments comparing model performance across question complexity levels. Using the CLEVR dataset [1] for synthetic compositional reasoning and GQA for real-world complexity, we categorize questions into four tiers: (1) object recognition, (2) attribute identification, (3) relational reasoning, and (4) multi-step inference. Three pretrained multimodal architectures—following bottom-up attention mechanisms [2]—will be evaluated on identical test splits. Primary metrics include overall accuracy, per-tier performance delta (= tier - tier), and **consistency scores** measuring answer stability under 10 visual perturbations (Gaussian noise, contrast shifts). Statistical significance will be assessed via paired t-tests with Bonferroni correction (α=0.01) and effect size calculations (Cohen's d).

Phase 2: Systematic Generalization Assessment To evaluate true compositional understanding (Research Question 2), we design novel train-test splits that systematically separate visual primitives from linguistic operators. Following Lake & Baroni's meta-learning framework [3], we construct "systematicity splits" where test sets contain novel combinations of training concepts (e.g., new object-attribute pairs). We define a **novelty ratio** as the proportion of test concept combinations absent from training data (target: 30%, 50%, 70%). Each model will be trained on five random seeds with 5-fold cross-validation, measuring systematic generalization gap as performance difference between in-distribution and out-of-distribution (OOD) test sets. We will probe attention patterns using Integrated Gradients to identify statistical shortcut reliance, comparing attention distributions against human-annotated reasoning regions.

Phase 3: Architectural Intervention Study Addressing Research Question 3, we implement explicit reasoning modules integrated with the best-performing Phase 1 backbone: (1) Neural Module Networks with transformer-based module controllers, and (2) Structured Scene Graph Attention using bottom-up region features [2]. Both interventions incorporate learnable reasoning chains supervised by human-annotated explanations [10]. We evaluate using the same systematicity splits while measuring interpretability through attention alignment scores (IoU with ground-truth reasoning chains) and faithfulness metrics. Ablation studies will isolate contributions of symbolic reasoning pathways versus enhanced attention mechanisms, with component removal experiments.

Datasets and Resources Experiments require approximately 2,000 GPU hours on NVIDIA A100 clusters. We will release an open-source toolkit containing: (1) systematicity split generators for CLEVR and GQA, (2) pretrained model checkpoints, and (3) evaluation scripts with standardized metrics including **compositional accuracy** and **systematicity index** (SI = OOD/ID performance ratio). Validity threats are mitigated through: (a) internal validity via k-fold cross-validation and statistical power analysis (n≥1000 samples per condition), (b) construct validity through human baseline comparisons (n=50 participants per 1,000 test questions), and (c) external validity via cross-dataset evaluation on Science QA [10] and the newly constructed **VQA-Systematic** benchmark containing 15,000 novel concept combinations. Reproducibility will be ensured through full code release, model cards, and dataset documentation following ML reproducibility checklists.

V. Timeline

Timeline

This 24-month research project employs a mixed-methods experimental design across seven overlapping phases. **Phase 1 (Months 1-3)** establishes theoretical foundations through systematic literature review on compositional reasoning [3] and multimodal architectures [2], while configuring computational infrastructure. **Phase 2 (Months 4-6)** curates diagnostic datasets, constructing challenge splits for CLEVR and GQA that isolate visual primitives from linguistic operators, plus a novel synthetic dataset for systematic generalization assessment [1].

Phase 3 (Months 7-10) implements three transformer baselines (LXMERT, ViLBERT, CLIP) and conducts diagnostic benchmarking to test Hypothesis 1, quantifying performance degradation on compositional versus descriptive queries. **Phase 4 (Months 9-14)** develops architectural interventions integrating neural module networks with structured scene graphs [10], explicitly designed to validate Hypothesis 3 regarding improved systematic generalization and interpretability.

Phase 5 (Months 13-18) executes controlled ablation studies and systematic generalization experiments using novel visual-linguistic combinations to test Hypothesis 2, with human performance baselines. **Phase 6 (Months 16-21)** performs interpretability analysis using attention visualization, Integrated Gradients, and human evaluation of generated reasoning chains, while developing an open-source evaluation toolkit with standardized metrics. **Phase 7 (Months 19-24)** synthesizes empirical findings, prepares three journal/conference manuscripts, and disseminates challenge datasets, baselines, and the evaluation toolkit to foster community adoption. Each phase incorporates iterative experimentation cycles and risk mitigation for computational constraints.

VI. Expected Results

Anticipated Empirical Findings

H1: Performance Degradation on Compositional Tasks: We expect diagnostic benchmarking to reveal that state-of-the-art VQA transformers (LXMERT, ViLBERT, CLIP) suffer accuracy reductions of 15-25% on compositional reasoning tasks (CLEVR, GQA challenge splits) compared to standard VQA benchmarks. Error analysis will identify systematic failure modes: poor handling of relational chains (e.g., "left of the rightmost object"), counting inaccuracies beyond training distributions, and inability to resolve ambiguous attribute bindings. These results will confirm that high aggregate scores reflect dataset biases rather than robust visual understanding, with human performance expected to remain $>$90% on these same tasks [1] [2].

H2: Systematic Generalization Deficits: Models will demonstrate $\leq$40% relative performance on out-of-distribution splits containing novel visual-linguistic compositions, confirming that statistical learning fails to capture systematicity [7] [3]. We anticipate particularly poor generalization to: (1) novel attribute combinations (e.g., "striped red sphere" when only "striped blue cube" appears in training), measured by novel combination accuracy $<$35%, (2) unseen question templates requiring identical reasoning (consistency scores $<$0.4), and (3) synthetic challenge sets with longer reasoning chains (performance degradation scaling linearly with chain length).

H3: Architectural Intervention Benefits: Neural module networks with structured scene graph conditioning will improve systematic generalization by $\geq$20% (absolute) while enhancing interpretability. Attention alignment with human-annotated reasoning chains will increase from baseline 0.35 to $\geq$0.65 (Spearman correlation), and feature attribution will reveal more focused reasoning pathways [10]. However, we anticipate computational overhead of 15-30% and potential overfitting to synthetic distributions, limiting real-world transfer without additional regularization.

Alternative Interpretations: Should baselines generalize better than expected, this may indicate emergent compositional capabilities in large-scale pretrained models, requiring re-evaluation of scaling versus architectural priors. Conversely, if interventions yield minimal gains, this could suggest that explicit reasoning modules require fundamentally different optimization objectives or that current evaluation benchmarks inadequately measure genuine compositionality. All findings will inform the development of our open-source evaluation toolkit and challenge dataset for community adoption.

VII. Conclusion

This research confronts a fundamental challenge in multimodal AI: despite high benchmark performance, contemporary VQA models lack genuine compositional reasoning capabilities and systematic generalization. We propose to systematically diagnose this gap by evaluating transformer-based architectures (LXMERT, ViLBERT, CLIP) on compositional reasoning tasks, testing three core hypotheses about performance degradation, systematic generalization deficits, and architectural enhancements [1], [3]. Our mixed-methods approach combines diagnostic benchmarking on CLEVR and GQA challenge splits, architectural interventions integrating neural module networks with structured scene graphs, and interpretability analysis through attention visualization and human evaluation. Expected contributions directly address our research questions: (1) comprehensive failure mode analysis revealing specific reasoning bottlenecks, (2) novel multimodal architectures with enhanced systematic generalization ($\geq$20% improvement) and interpretability, and (3) an open-source evaluation toolkit for community adoption. This work bridges connectionist and symbolic paradigms [2], [10], advancing toward AI systems that demonstrate human-like systematic generalization rather than statistical pattern memorization. The broader impact spans robotics, scientific discovery, and accessible technologies, fostering more trustworthy and interpretable multimodal intelligence.

References

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "Vqa: Visual question answering," *Unknown*, 2015.

[2] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. J. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," *ANU Open Research (Australian National University)*, 2018.

[3] B. M. Lake and M. Baroni, "Human-like systematic generalization through a meta-learning neural network," *Nature*, 2023.
[4] D. Xue, S. Qian, and C. Xu, "Variational causal inference network for explanatory visual question answering," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 2515–2525.
[5] ——, "Integrating neural-symbolic reasoning with variational causal inference network for explanatory visual question answering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 7893–7908, 2024.
[6] D. Xue, S. Qian, Q. Fang, and C. Xu, "Linin: Logic integrated neural inference network for explanatory visual question answering," *IEEE Transactions on Multimedia*, vol. 27, pp. 16–27, 2024.
[7] R. J. O'Hara, "Systematic generalization, historical fate, and the species problem," *Systematic Biology*, 1993.
[8] D. Xue, S. Qian, and C. Xu, "Few-shot multimodal explanation for visual question answering," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 1875–1884.
[9] H. Nam, J.-W. Ha, and J. Kim, "Dual attention networks for multimodal reasoning and matching," *Unknown*, 2017.
[10] P. Lu, S. Mishra, T. Xia, L. Qiu, K.-W. Chang, S. Zhu, O. Tafjord, P. Clark, and A. Kalyan, "Learn to explain: Multimodal reasoning via thought chains for science question answering," *arXiv (Cornell University)*, 2022.
[11] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Computing Surveys*, 2022.
[12] M. A. K. Raiaan, M. S. H. Mukta, K. Fatema, N. M. Fahad, S. Sakib, M. M. J. Mim, J. Ahmad, M. E. Ali, and S. Azam, "A review on large language models: Architectures, applications, taxonomies, open issues and challenges," *IEEE Access*, 2024.