Urdu-CLEVR: A Novel Benchmark for Visual Reasoning in an Under-Resourced Linguistic Context

Sohail Ashraf, Adeel Zafar, Slawomir Nowaczyk, Ahthasham Sajid
Riphah International University, Islamabad, Pakistan
CAISR, Halmstad University, Halmstad, Sweden
Fazaia Bilquis College of Education for Women, Islamabad, Pakistan
cheemasohail286@gmail.com, adeel.zafar@hh.se,
slawomir.nowaczyk@hh.se, ahthashamsajid@gmail.com

Abstract

Visual Question Answering (VQA) bridges the gap between computer vision and natural language processing, yet progress remains largely confined to high-resource languages. For low-resource languages such as Urdu, research is severely hindered by the absence of large-scale reasoning-based datasets. To address this critical gap, we introduce the first synthetic Urdu VQA dataset, modeled on the CLEVR framework and specifically designed to evaluate complex, multi-step visual reasoning. We conduct a rigorous comparative analysis using both transformer-based architectures (VisualBERT, LXMERT, ViLT) and neuro-symbolic models. The evaluation of these architectures validate that Urdu-CLEVR provides a meaningful and non-trivial benchmark for compositional reasoning. This work establishes a primary benchmark for Urdu VQA, demonstrating that hybrid reasoning architectures provide a robust and scalable solution for advancing multimodal AI in under-resourced linguistic contexts.
The dataset is publicly available at Urdu-CLEVRso how the

Keywords: Low Resource Language, Urdu VQA, Synthetic Dataset, Multimodal AI

1. Introduction

The field of VQA presents a unique intersection of linguistic complexity and computer vision, requiring models to navigate semantic, syntactic, and pragmatic layers simultaneously (Antol et al., 2015). For the Urdu language, this task is particularly formidable due to its rich morphological structure and intricate syntactic dependencies. Despite its status as a widely spoken language, the integration of Urdu linguistic elements with visual context remains a significant bottleneck in multimodal research. This disparity not only hinders the development of robust natural language understanding (NLU) models but also limits the deployment of AI in critical sectors such as accessibility, information retrieval, and digital education.

Furthermore, Urdu VQA serves as a vital tool for cultural preservation and historiographic study, enabling the archival and comprehension of visual content deeply rooted in linguistic heritage (Maryam et al., 2024). Advancing VQA technologies for Urdu is therefore not merely a technical challenge but a necessity for technological inclusivity. Improvements in this domain lead to more sophisticated AI systems capable of handling the complex configurations and unconventional spatial orientations inherent in South Asian scripts and visual contexts.

However, the advancement of such systems is currently stalled by the total absence of datasets designed to test complex visual reasoning in Urdu. Most existing resources for low-resource languages are limited to simple object identification or automated translations that lack structural rigor. To address this gap, we introduce **Urdu-CLEVR**, the first reasoning-aware benchmark for the Urdu language.

This work makes two primary contributions to the field of multimodal AI. First, we introduce Urdu-CLEVR, the first human-verified, complexity-controlled VQA benchmark for the Urdu language. Unlike existing ad-hoc datasets, Urdu-CLEVR utilizes a structured template-based generation process that allows for the isolation and systematic study of multi-step logical reasoning, effectively bridging a critical resource gap for underrepresented languages. Second, we establish the first relational VQA benchmarks for Urdu, demonstrating that neuro-symbolic architectures (85.3%) significantly outperform transformer-based models by up to 7.1%. These results provide a robust methodological foundation for developing equitable vision-language technologies in low-resource linguistic contexts.

2. Related Work

2.1. Development and Importance of VQA Datasets

VQA datasets serve as the primary vehicle for evaluating multimodal integration, reasoning depth, and model robustness. As the field has evolved, these benchmarks have addressed critical challenges such as language bias, cultural specificity,

Proceedings of SIGUL 2026 Joint Workshop with ELE, EURALI, and DCLRL @ LREC 2026, pages 74–82
11–12 May 2026.

Category	Dataset	Lang.	Synth.	Programs	Bias-Ctrl.	Knowledge	Scale
Diagnostic Reasoning	CLEVR	Eng	✓	✓	✓	✗	100K images
	FigureQA	Eng	✓	✓	✓	✗	100K+ figs
	DVQA	Eng	✓	✓	Partial	✗	3M QA pairs
Real-World / Bias-Ctrl.	VQA 2.0	Eng	✗	✗	Partial	✗	1.1M QA pairs
	GQA	Eng	✗	Partial	Partial	✗	22M Qs
	VQA-CP	Eng	✗	✗	✓	✗	438K Qs
Knowledge Based	FVQA	Eng	✗	✗	✗	✓	5.8K QA pairs
	OK-VQA	Eng	✗	✗	✗	✓	14K Qs
	KVQA	Eng	✗	✗	✗	✓	183K QA pairs
Urdu Multimodal	Urdu Scene-Text	Urdu	✗	✗	✗	✗	∼1K images
	Urdu-CLEVR (Ours)	Urdu	✓	✓	✓	✗	**85K Img, 850K QA**

Table 1: Comparison of key VQA benchmarks. Urdu-CLEVR is the first diagnostic, reasoning-oriented benchmark designed specifically for the Urdu language.

and the integration of external knowledge. As summarized in Table 1, the landscape of VQA research can be categorized into three distinct paradigms: synthetic diagnostic reasoning, large-scale real-world perception, and knowledge-intensive inference.

To isolate the challenges of compositional reasoning from the complexities of natural image noise, synthetic benchmarks provide a controlled environment for systematic evaluation. The CLEVR dataset (Johnson et al., 2017) pioneered this approach by utilizing automatically generated scenes and functional program annotations to track multi-step logical operations. Subsequent works like FigureQA (Kahou et al., 2017) and DVQA (Kafle et al., 2018) extended this logic to relational comparisons and structured data visualizations. While these datasets excel at bias control and providing transparency through symbolic grounding, they are traditionally limited to high-resource languages, leaving a significant gap in diagnostic tools for under-represented languages.

Conversely, real-world datasets focus on the richness of natural imagery and linguistic plurality. VQA 2.0 (Goyal et al., 2017) introduced balanced answer distributions to mitigate the influence of language priors, while GQA (Hudson and Manning, 2019) utilized scene graph annotations to bridge the gap between real images and compositional logic. VQA-CP (Ramakrishnan et al., 2018) further refined this by re-distributing answers to test model resistance to dataset biases. Despite their scale, these datasets often lack the explicit reasoning programs required to evaluate the intermediate logical steps of a model's decision-making process.

The third category transcends visual perception by requiring external world knowledge. FVQA (Wang et al., 2017) integrates structured knowledge bases to answer fact-oriented queries, while OK-VQA (Marino et al., 2019) and KVQA (Shah et al., 2019) demand commonsense and named-entity reasoning, respectively. While these benchmarks push models toward cognitive-level understanding, they rarely offer the structured reasoning annotations or bias-controlled environments necessary for evaluating compositional generalization in low-resource settings.

2.2. Neural and Symbolic Approaches for VQA

Early VQA frameworks relied on foundational feature extractors, employing Convolutional Neural Networks (CNNs) for visual encoding and Recurrent Neural Networks (RNNs) for linguistic processing. While these models established the basis for multimodal integration, they were inherently limited in capturing the intricate, high-order interactions between modalities (Mao, 2024). A significant architectural shift occurred with the introduction of attention mechanisms, which enabled models to dynamically weight relevant spatial regions and textual tokens. This selectively-weighted approach significantly improved model interpretability and performance (Zhang et al., 2021), a capability further refined by co-attention mechanisms that jointly align features across both domains (Yu et al., 2018).

Recent advancements have transitioned toward Transformer-based architectures. By leveraging self-attention, these models extract fine-grained relational features, facilitating a more holistic understanding of scene context (Li et al., 2021). Modern trends continue to refine these networks to model spatial-sequential dependencies, effectively integrating local visual features with global semantic dependencies (Liu et al., 2021).

However, despite the success of purely neural transformers, they often lack explicit logical grounding. Neuro-symbolic computing addresses this by fusing the robust perceptual capabilities of neural networks with the formal reasoning and

transparency of symbolic AI (Wang et al., 2024). This hybrid paradigm is particularly relevant for diagnostic datasets like CLEVR, where multi-step logical inference is required. By decoupling perception from reasoning, neuro-symbolic models provide a pathway for interpretable and scalable VQA, particularly in linguistic contexts where massive end-to-end training data may be scarce.

2.3. Developments of Urdu VQA datasets

One of the earlier academic attempts at Urdu VQA was the dataset presented by Maryam et al. (2024), which covers Urdu natural scene text detection, recognition, and VQA. It is the first benchmark specifically designed for Urdu scene-text-based VQA, as it comprises 1,000 or more real-world images, annotated text instances, and related Urdu question-answer pairs. It helps to combine multimodal reasoning to complex real-world and text rich environment.

Other Urdu VQAs have been created outside the academic standards as a result of community and industrial efforts. FutureBee AI provides a larger proprietary dataset that contains around 5,000 images and 35,000 pairs of Urdu QA, but it is not peer-reviewed. A few text-only QA datasets are available, which is useful as a base of multilingual reasoning transfer, despite the data not being images. There are very few image captioning datasets that have been recently created, such as the COCO-Urdu dataset developed by Hassan (2025), which contains 59000 images and 319000 Urdu captions. Another image captioning dataset developed by Muzaffar et al. (2025), which contains 159,816 captions and is inspired by the Flickr30k Plummer et al. (2015) dataset. Nevertheless, the current Urdu VQA resources are small in size and scope, and none of them offers synthetic bias-controlled settings or organized reasoning supervision by annotating the programs. This poses a tremendous loophole in an assessment of compositional and interpretable VQA models in Urdu.

2.4. Challenges in Urdu Language Visual Question Answering

Urdu script presents significant challenges due to its cursive nature and context-dependent character shapes Anjum and Khan (2023). Urdu's rich morphology and orthography add layers of complexity to language processing tasks such as tokenization, part-of-speech tagging, and parsing Fatima et al. (2007). This limitation affects the growth of reliable natural language processing (NLP) systems Ullah et al. (2024). However, the overall lack of resources remains a significant barrier to advancing Urdu language processing Basit et al. (2024). The existing literature primarily discusses VQA in a broader sense, emphasizing the challenges of integrating visual and linguistic data to generate meaningful responses to questions posed in natural language Barra et al. (2021).

Despite the integration of Urdu NLP into Visual Question Answering (VQA) and related tasks such as image captioning, the incorporation of Urdu NLP remains challenging primarily due to the richness of the Urdu language Both of these linguistic and resource-level issues also lead to the desire to have controlled, synthetic, and reasoning-conscious Urdu VQA benchmarks capable of measuring compositional generalization in a systematic way.

3. Materials and Methods

Our methodology comprises the systematic synthesis of the Urdu-CLEVR dataset and a comparative evaluation of transformer-based and neuro-symbolic architectures. While these models excel on English benchmarks, we investigate their capacity to navigate Urdu's morphological nuances and resolve complex visual reasoning tasks in a low-resource context. The pipeline is divided into three primary stages: (1) **Dataset Synthesis**, utilizing structured templates to control linguistic and reasoning variance; (2) **Cross-modal Alignment**, involving specialized preprocessing to map Urdu text to visual features; and (3) **Evaluation and Analysis**, where models are assessed on accuracy, interpretability, and cross-linguistic generalization between the Urdu and original English CLEVR environments.

3.1. Development of Urdu-CLEVR Dataset

The Urdu-CLEVR dataset is modeled after the Compositional Language and Elementary Visual Reasoning (CLEVR) framework (Johnson et al., 2017). The original benchmark utilizes synthetically generated images featuring objects of three distinct shapes (cube, sphere, and cylinder), two sizes (small and large), and two material types (metal and rubber) in eight colors. These objects are positioned within a three-dimensional space to establish four primary spatial relationships: left, right, behind, and in front.

As illustrated in Figure 1, our development process began by defining the reasoning scope and linguistic parameters required for the Urdu language. To maintain visual consistency with existing benchmarks, we utilized the original CLEVR image set while focusing our efforts on gener-

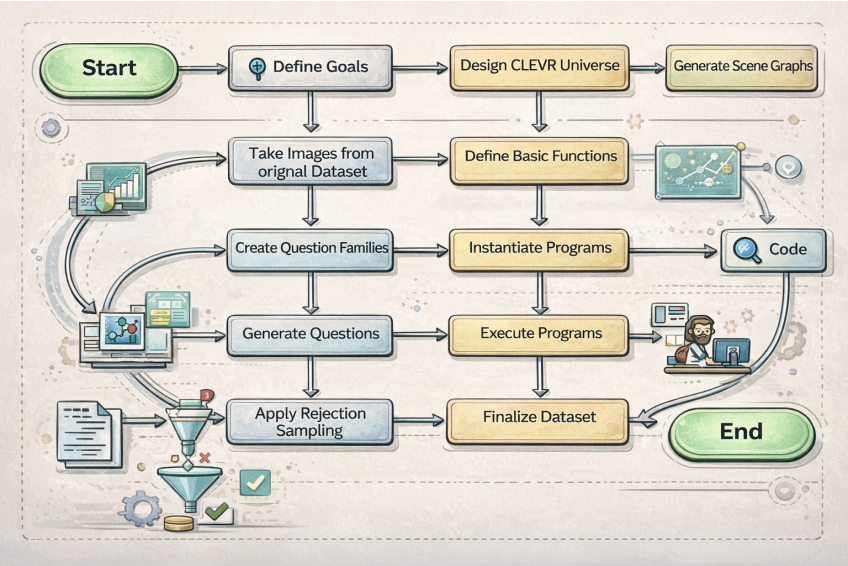

Figure 1: Pipeline of end-to-end dataset generation of the Urdu CLEVR benchmark

ating a novel linguistic corpus. This involved defining a set of functional modules for compositional reasoning—such as filtering, relating, and counting—and designing question families that reflect these logic templates.

To ensure semantic validity, we instantiated these templates into runnable programs executed against the scene graphs of the original images. These generated questions were rigorously filtered to eliminate ill-posed or degenerate instances. Furthermore, we employed rejection sampling to achieve a balanced answer distribution and mitigate inherent dataset biases. The final Urdu-CLEVR dataset comprises these validated, human-verified question-answer pairs, providing a robust tool for evaluating multi-step visual reasoning in Urdu.

3.1.1. Image generation

The original CLEVR visual assets are language-agnostic and rendered using Blender, we repurposed the existing image set to maintain cross-lingual consistency. Each scene contains a randomized configuration of three to ten objects varying in color, material, shape, size, and spatial coordinates. To ensure a robust evaluation, we utilized the standard CLEVR split: a training set of 70,000 images and a testing set of 15,000 images. By maintaining the original dataset's visual scale and parameters, we ensure that our performance comparisons specifically isolate the challenges of Urdu linguistic reasoning rather than visual variance.

3.1.2. Scene Graphs Representations

Each image is accompanied by a comprehensive scene graph that provides a structured representation of the visual environment. These metadata files encode the precise spatial coordinates and physical attributes—such as shape, color, and material—of every object within the scene. Figure 2 illustrates a comparative sample of the ground-truth scene graphs, showcasing the alignment between the original English labels and our corresponding Urdu translations. By leveraging these structured annotations, we ensure that the logical programs used for question generation are grounded in accurate, human-verified Urdu semantics.

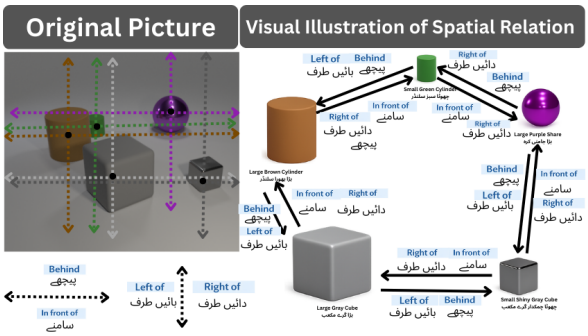

Figure 2: Sample image from the training dataset

3.1.3. Question-answer pairs generation

To ensure logical consistency and systematic coverage of reasoning tasks, we utilized a template-based generation framework. We designed nine distinct question families, each containing a variety of question types that target specific cognitive abilities, such as attribute identification, spatial relationship reasoning, and integer comparison. These templates are instantiated using functional

programs that traverse the image scene graphs to produce precise question-answer pairs. A detailed breakdown of the question families and their corresponding reasoning tasks is provided in Table 2.

There are 90 question families with 426 different questions, and to further diversify the dataset of questions, we created a set of synonyms for each shape, size, position, material, and color that helped us to generate more and more unique questions.

3.1.4. Statistics of the Dataset

The corpus is partitioned into training and testing subsets to ensure robust evaluation. As shown in Table 3, the distribution of question-answer pairs reflects significant linguistic variety and scale, providing the structural integrity required for diagnostic reasoning. To maximize generalization, the dataset features a high volume of unique, synthetically composed questions with minimal redundancy, facilitating the training of deep multimodal architectures on complex, multi-step Urdu scene descriptions.

3.1.5. Question Type Distribution

To ensure a comprehensive evaluation of cognitive performance, we analyzed the distribution of reasoning types within the corpus. These categories target distinct logical operations, ranging from basic attribute retrieval to complex arithmetic inference. Figure 3 illustrates the proportion of each question type across the entire dataset.

The analysis of the distribution reveals a balanced emphasis on relational and compositional reasoning. Counting-based queries represent the largest segment at 25%, requiring models to quantify objects under specific conditional constraints. Existence-based questions (14%) assess the model's ability to verify the presence of objects within a scene. A significant portion of the dataset is dedicated to comparison tasks—spanning shape, material, color, and size—which underscores the benchmark's focus on relational logic. Attribute queries regarding specific object features each comprise between 4% and 10% of the data, while integer comparison questions (e.g., "greater than" or "equal to") account for 3% of the set, specifically testing logical and arithmetic inference. This diversified distribution ensures that the Urdu-CLEVR dataset provides a robust environment for training and analyzing models across a wide spectrum of cognitive tasks.

3.2. Models Evaluated

We evaluate three representative transformer-based architectures and one neuro-symbolic

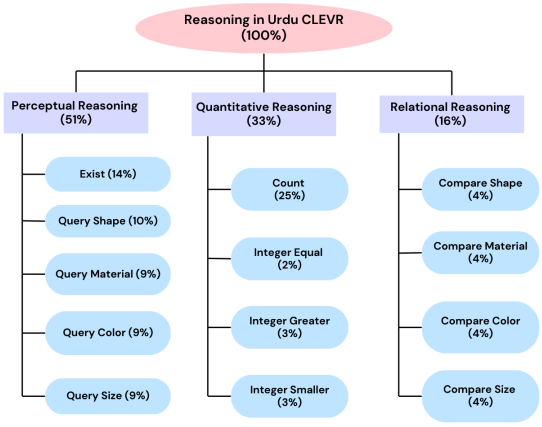

Figure 3: Distribution of Question Types in Urdu CLEVR Dataset

framework to assess their efficacy on Urdu-CLEVR. Among the transformers, **VisualBERT** (Li et al., 2019) serves as a single-stream baseline for holistic scene interpretation; **LXMERT** (Tan and Bansal, 2019) utilizes a dual-stream encoder for fine-grained cross-modal alignment; and **ViLT** (Kim et al., 2021) provides a computationally efficient, end-to-end alternative. In contrast, we implement a **Neuro-Symbolic** framework (Yi et al., 2018) to decouple perception from reasoning. This hybrid approach uses neural networks for attribute extraction and a symbolic engine to execute parsed Urdu questions against a structured scene graph, offering higher transparency and explicit logical reasoning compared to purely attention-based models.

4. Experimental Evaluation

4.1. Preprocessing and Data Alignment

Preprocessing was critical to bridge the gap between the English-centric CLEVR framework and the structural requirements of the Urdu language. While the original language-agnostic visual assets were retained to provide controlled object layouts, the textual and logical components underwent significant modification.

4.1.1. Visual Input Preparation

Visual inputs were decomposed into structured scene graphs encoding object properties (color, shape, size) and spatial coordinates. For the neuro-symbolic model, these graphs served as the ground-truth symbolic representation, while for transformer models, they provided the necessary metadata for regional feature alignment.

#	Template Name	Urdu Description (Template)	English Translation	Classes	Types
1	Zero_hop	<Z> <C> <M> <S> کی تعداد کیا ہے؟	What is the number of ‹Z› ‹C› ‹M› ‹S›?	6	23
2	One_hop	<Z> جو <Z2> <C2> <M2> <S2> کے <R> ہے، اس کا سائز کیا ہے؟	What is the size of the ‹Z2› ‹C2› ‹M2› ‹S2› that is ‹R› of the ‹Z› ‹C› ‹M› ‹S›?	6	28
3	Two_hop	<Z> جو <Z2> <C2> <M2> <S2> کے <R2> ہے جو <Z3> <C3> <M3> <S3> کے <R> ہے، اس کا سائز کیا ہے؟	What is the size of the ‹Z3› ‹C3› ‹M3› ‹S3› that is ‹R2› of the ‹Z2› ‹C2› ‹M2› ‹S2› that is ‹R› of the ‹Z› ‹C› ‹M› ‹S›?	6	30
4	Three_hop	کیا کوئی <Z4> <C4> <M4> <S4> ہے جو <Z3> <C3> <M3> <S3> کے <R3> ہے جو <Z2> <C2> <M2> <S2> کے <R2> ہے جو <Z> <C> <M> <S> کے <R> ہے؟	Is there a ‹Z4› ‹C4› ‹M4› ‹S4› that is ‹R3› of the ‹Z3›... that is ‹R› of the ‹Z› ‹C› ‹M› ‹S›?	6	30
5	Single_or	<Z> <C> اشیاء کی تعداد کیا ہے جو یا تو <Z2> <C2> <M2> <S2> ہیں یا <Z> <C> <M> <S>؟	What is the number of objects that are either ‹Z› ‹C› ‹M› ‹S› or ‹Z2› ‹C2› ‹M2› ‹S2›?	8	30
6	Single_and	<Z3> <C3> <M3> <S3> کی تعداد کیا ہے جو <Z2> <C2> <M2> <S2> کے بھی <R2> ہیں اور <Z> <C> <M> <S> کے بھی <R> ہیں؟	What is the count of ‹Z3›... that are both ‹R2› of ‹Z2›... and ‹R› of ‹Z›...?	5	25
7	Comp_Int	کیا <Z> <C> <M> <S> اور <Z2> <C2> <M2> <S2> کی تعداد برابر ہے؟	Is the number of ‹Z› ‹C› ‹M› ‹S› and ‹Z2› ‹C2› ‹M2› ‹S2› equal?	9	21
8	Comparison	کیا <Z> <C> <M> <S> اور <Z2> <C2> <M2> <S2> کا سائز ایک جیسا ہے؟	Do ‹Z› ‹C› ‹M› ‹S› and ‹Z2› ‹C2› ‹M2› ‹S2› have the same size?	16	88
9	Same_Relate	کیا کوئی اور چیز ہے جس کا رنگ <Z> <C> <M> <S> جیسا ہو؟	Is there anything else that has the same color as the ‹Z› ‹C› ‹M› ‹S›?	28	151

Table 2: Question templates with Urdu descriptions and English translations.

Characteristic	Training Set	Testing Set
Total Questions	699,989	149,991
Unique Questions	600,269	138,812
Unique Answers	35	35
Vocabulary Size	664,916	163,880

Table 3: Statistical summary of the Urdu-CLEVR question and answer pairs.

4.1.2. Linguistic Localization

Linguistic transformation involved more than direct translation; it required a structural mapping of English logic templates into Urdu. We developed a collection of Urdu templates verified by native speakers to ensure syntactic accuracy and morphological consistency. To enhance model generalizability, we implemented a synonym mapping system for key attributes (e.g., *red* → *surkh/laal*), capturing the variety of natural discourse.

The textual pipeline utilized the NLTK library for Urdu-specific tokenization and normalization. All inputs were standardized to Unicode to ensure consistency across different Urdu input methods and diacritic usage, preventing potential character-level mismatches during model training.

4.2. Training and Evaluation

4.2.1. Hyperparameter Settings

All models were fine-tuned on the Urdu-CLEVR dataset using optimized hyperparameters derived from established literature (Li et al., 2019; Yi et al., 2018) and empirical validation. To ensure stability and convergence, we maintained uniform batch sizes and learning rate schedules across the transformer-based architectures, while the neuro-symbolic model was trained using a modular curriculum approach.

4.2.2. Performance Metrics

Model performance was evaluated based on **Accuracy**, defined as the percentage of correct predictions over the total test set, and **Computa-**

Models	Accuracy (%)	Processing Time (sec/img)
VisualBERT	72.5	0.75
LXMERT	78.2	1.10
ViLT	74.6	0.40
Neuro-Symbolic Model	**85.3**	**0.95**

Table 4: Comparison of model performance on the Urdu-CLEVR dataset.

tional Efficiency, measured as average processing time per image (s/img). These metrics assess both the reasoning capabilities of the architectures and their feasibility for deployment in resource-constrained environments.

5. Results

The performance of the evaluated models: VisualBERT, LXMERT, ViLT, and the neuro-symbolic architecture is summarized in Table 4 based on the identified key performance indicators. This comparative analysis elucidates the relative strengths of each model within the Urdu VQA domain, highlighting how different architectural paradigms navigate the challenges of linguistic complexity. By assessing both transformer-based and neuro-symbolic models across English and Urdu CLEVR benchmarks, this study provides critical insights into the scalability and cross-lingual potential of VQA systems.

The results of VisualBERT, LXMERT, and ViLT as depicted in Figure 4 indicate a definite performance difference between English and Urdu CLEVR. These models are below par on English CLEVR (e.g., LXMERT with 96.3% and ViLT with a mere 85%) but also average on Urdu CLEVR. The fact that this has decreased is indicative of several factors such as rich morphology, free word order, small pretraining corpora (since most transformers are trained on mostly English), and tokenization problems (subword segmentation might not be able to recognize semantic units in Urdu).

These results confirm that while transformer-based VQA systems are highly performant in high-resource settings, they lack zero-shot transferability to low-resource environments without extensive linguistic adaptation, domain-specific pre-training, or sophisticated cross-lingual alignment. In contrast, the neuro-symbolic model demonstrates superior robustness, achieving near-perfect accuracy on English CLEVR (99.8%) and a strong 85.3% on Urdu CLEVR as shown in Figure 4. While a performance degradation is observed when transitioning to Urdu, the neuro-symbolic architecture proves far more resilient than its transformer-based counterparts. Furthermore, the neuro-symbolic approach provides an explicit, interpretable line of reasoning, offering a level of transparency that remains elusive in the implicit attention mechanisms of transformer models.

The neuro-symbolic framework's performance highlights its potential as a backbone for language-agnostic VQA, especially in low-resource settings lacking massive pre-training corpora. Our analysis yields three critical insights: (1) **Benchmark Necessity**: the performance decay of English-optimized models on Urdu underscores the urgent need for language-specific benchmarks like Urdu-CLEVR; (2) **Symbolic Resilience**: structured logical operations prove more robust than dense embeddings in low-resource contexts; and (3) **Hybrid Potential**: combining transformer-based alignment with neuro-symbolic reasoning could enhance both accuracy and explainability. Ultimately, while transformers excel in data-rich environments, neuro-symbolic models offer the interpretability and adaptability required for the "long tail" of under-resourced languages, making Urdu-CLEVR a vital tool for inclusive multimodal research.

6. Limitations

While the Urdu-CLEVR dataset provides a controlled environment for evaluating visual reasoning, several limitations remain. First, the use of synthetic images and template-based question generation may not fully reflect the linguistic diversity and spontaneous nature of natural Urdu discourse. Despite incorporating synonyms and varied templates, the dataset's structural rigidity inherently limits its scope to formal reasoning tasks.

Second, the translation of spatial attributes and relational logic into Urdu carries the risk of semantic drift; English-centric spatial concepts may not always align perfectly with Urdu's cultural or linguistic framing of physical space. Consequently, while Urdu-CLEVR serves as a vital diagnostic benchmark, future work should focus on extending these reasoning tasks to more diverse, naturalistic visual and linguistic settings.

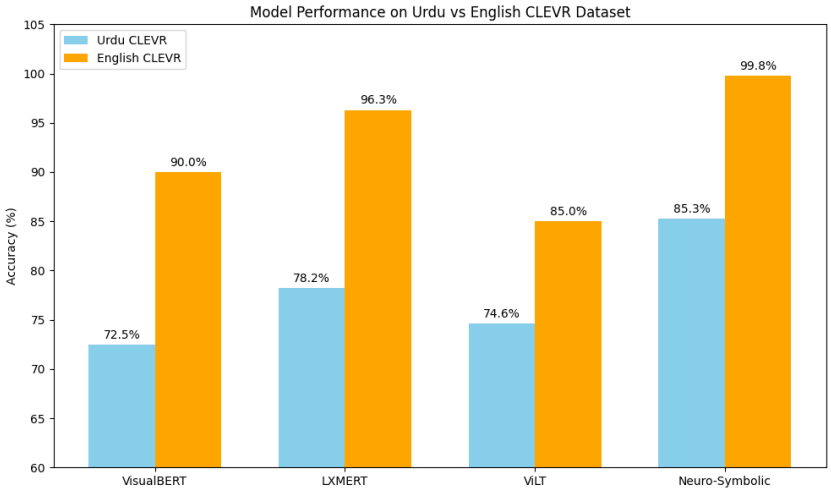

Figure 4: Performance of models on Urdu and English datasets

7. Conclusion

This paper introduces Urdu-CLEVR, a novel benchmark designed to evaluate multi-step visual reasoning in an under-resourced linguistic context. Through a comparative study of transformer-based and neuro-symbolic architectures, we demonstrate that while purely neural models excel in high-resource settings, they struggle to maintain accuracy and logical consistency when transitioned to Urdu. In contrast, our findings reveal that the neuro-symbolic approach achieves a superior accuracy of 85.3%, providing a more robust and interpretable framework for complex reasoning. The development of Urdu-CLEVR highlights critical linguistic gaps that remain obscured in English-centric benchmarks, underscoring the necessity of language-specific resources for equitable AI development.

8. References

Tayaba Anjum and Nazar Khan. 2023. Caltext: Contextual attention localization for offline handwritten text. *Neural Processing Letters*, 55(6):7227–7257.

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

Silvio Barra, Carmen Bisogni, Maria De Marsico, and Stefano Ricciardi. 2021. Visual question answering: Which investigated applications? *Pattern Recognition Letters*, 151:325–331.

Abdul Basit, Abdul Hameed Azeemi, and Agha Ali Raza. 2024. Challenges in urdu machine translation. In *Proceedings of the The Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*, pages 44–49.

Tayyaba Fatima, R Islam, and M Anwar. 2007. Morphological and orthographic challenges in urdu language processing: A review. In *Proc. 13th Workshop Asian Lang. Resour.*, 1, pages 44–51.

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

Umair Hassan. 2025. Coco-urdu: A large-scale urdu image-caption dataset with multimodal quality estimation. *arXiv preprint arXiv:2509.09014*.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910.

Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. 2018. Dvqa: Understand-

ing data visualizations via question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5648–5656.

Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. 2017. Figureqa: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300*.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Yue Li, Jin Liu, and Shengjie Shang. 2021. Wma: a multi-scale self-attention feature extraction network based on weight sharing for vqa. *Journal on Big Data*, 3(3):111.

Yun Liu, Xiaoming Zhang, Qianyun Zhang, Chaozhuo Li, Feiran Huang, Xianghong Tang, and Zhoujun Li. 2021. Dual self-attention with co-attention networks for visual question answering. *Pattern Recognition*, 117:107956.

Keyu Mao. 2024. Enhancing visual question answering through bi-modal feature fusion: Performance analysis. In *Proceedings of the 2024 6th International Conference on Image Processing and Machine Vision*, pages 115–122.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204.

Hiba Maryam, Ling Fu, Jiajun Song, Tajrian ABM Shafayet, Qidi Luo, Xiang Bai, and Yuliang Liu. 2024. Dataset and benchmark for urdu natural scenes text detection, recognition and visual question answering. In *International Conference on Document Analysis and Recognition*, pages 279–292. Springer.

Rimsha Muzaffar, Syed Yasser Arafat, Junaid Rashid, Jungeun Kim, and Usman Naseem. 2025. Uicd: A new dataset and approach for urdu image captioning. *PLoS One*, 20(6):e0320701.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. 2018. Overcoming language priors in visual question answering with adversarial regularization. *Advances in neural information processing systems*, 31.

Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. 2019. Kvqa: Knowledge-aware visual question answering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8876–8884.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

Fida Ullah, Alexander Gelbukh, Muhammad Tayyab Zamir, Edgardo Manuel Felipe Riverón, and Grigori Sidorov. 2024. Enhancement of named entity recognition in low-resource languages with data augmentation and bert models: a case study on urdu. *Computers*, 13(10):258.

Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. 2017. Fvqa: Fact-based visual question answering. *IEEE transactions on pattern analysis and machine intelligence*, 40(10):2413–2427.

Wenguan Wang, Yi Yang, and Fei Wu. 2024. Towards data-and knowledge-driven ai: a survey on neuro-symbolic computing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. 2018. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. *Advances in neural information processing systems*, 31.

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

Sheng Zhang, Min Chen, Jincai Chen, Fuhao Zou, Yuan-Fang Li, and Ping Lu. 2021. Multimodal feature-wise co-attention method for visual question answering. *Information Fusion*, 73:1–10.