Accessible VQA Evaluation Using Reading-Level Controlled GQA

Michael Dong[1]

Received March 27, 2026
Accepted July 25, 2026
Electronic access September 15, 2026

This paper introduces an Accessibility Focused GQA Variant dataset, AFGV, with the aim of assessing the linguistic accessibility of Visual Question Answering systems for VQA in educational, real-world, and assistive contexts. AFGV performs rewrites of the original prompt that can help assess the models' sensitivity to lexical variation while holding the image and ground truth constant. Prompt variants are created using a Word2Vec- based semantic substitution pipeline. A meaningful word is swapped with a semantically equivalent word selected based on Word2Vec similarity and readability-based rankings for each question. The original and re-written prompts are scored based on Flesch-Kincaid reading level, and those scores are aggregated as features in splits. The dataset comes with reproducible code to generate variants of a question and running the models, the images and questions as part of our evaluation pipeline, and finally the table of our human evaluation on 100 randomly picked samples. Applying AFGV to GQA, we investigate the impact of lexical and syllabic variation on the accuracy of VQA, and obtain hypotheses on the accessibility issues faced by AI tools targeting local communities. In both BLIP, InstructBLIP and LLaVA-1.5, reworded prompts resulted in a performance decrease compared to the performance obtained with the original prompts, which may indicate that existing VQA systems are to some extent sensitive to lexical variation, and thus can unveil prompts' related robustness problems due to accessibility. Effects are also broken down by question type to determine if shifts in performance vary by types of visual reasoning.

Keywords: Accessibility; Visual Question Answering (VQA); Lexical Complexity; Dataset Benchmarking; Language Bias Robustness, Answer Semantic Similarity

Introduction

Improvements on the standard VQA are unable to allow us to distinguish between models that are grounded in visual reasoning or linguistic shortcuts[1]. This is important for community-facing systems: a system that performs well when aggregated may perform unreliably when a user presents the same query in different terms[2]. Prior work demonstrates that models can predict answers from the question itself—by picking up on conditional biases, and correlations specific to the dataset—thus making an educated guess without referencing the image faithfully[3]. For community-facing applications—users with a wide range of reading levels and vocabulary—this shortcut introduces a second problem: linguistic accessibility. In this work, we study VQA sensitivity to lexical changes and variations in readability. The lens of accessibility motivates us: if models are sensitive to changes in phrasing, they may not behave reliably for users who phrase their queries differently from how the models are trained. We need to verify whether such sensitivity exists before making any accessibility-oriented claims. Recently, domain-generalization efforts such as VQA-GEN also emphasize evaluating VQA robustness to distribution shifts both in the visual as well as the textual domain[4].

The GQA dataset was recently introduced to promote real-world object recognition and compositional reasoning using scene graphs[5]. Questions are associated with a scene graph describing the objects, attributes, and relations in the image, encouraging compositional, multi-step reasoning over question template completion. Like CLEVR, GQA uses a semantic representation for question generation, and the answer distribution is audited to reduce class imbalance (e.g., an over-representation of yes/no answers). Conditional biases still creep in, though, and an issue largely neglected in VQA dataset design is linguistic accessibility. Like most VQA datasets, GQA poses a single canonical way to ask each question; actual users have diverse reading comprehension levels, vocabularies, language abilities, and needs[6,7]. To better serve students, English language learners, and users with language-related disabilities, VQA datasets should explicitly evaluate how well a model performs when questions are rewritten at easier and more challenging levels, and how well meaning is preserved in the rewriting. We contribute AFGV, an accessibility-focused variant of GQA, by generating simple rewrite pairs (simplified, baseline, complex) through our Word2Vec-based semantic substitution pipeline and grade-level readability formula (Fig. 1).

[1] *Archbishop Mitty High School, USA*

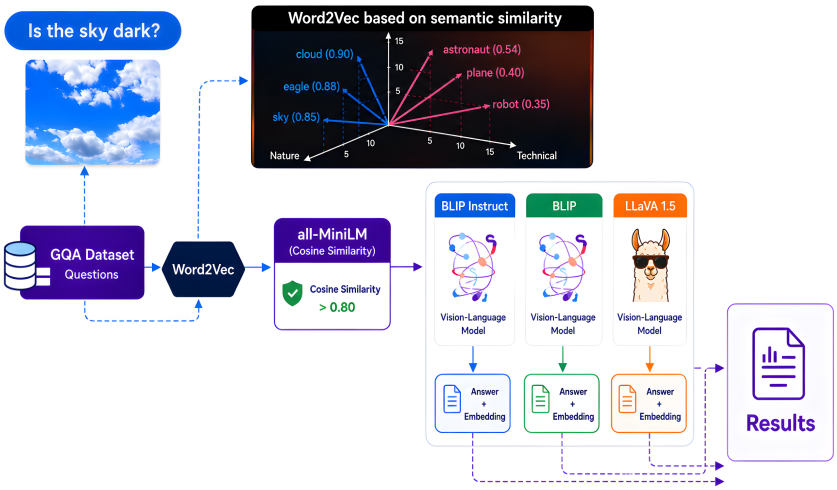

Figure. 1 Overview of the AFGV dataset creation

This pipeline takes questions from the GQA dataset and uses Word2Vec to estimate linguistic complexity. Selected questions are then rewritten in two directions into a simplified variant and a more complex variant while the original questions are retained as a reference. The original and rewritten questions are compared using all-MiniLM sentence embeddings, and cosine similarity ($\geq$0.80) is used to estimate intent preservation. All three versions (original, simplified, and complex) are evaluated with three models: BLIP, InstructBLIP, and LLaVA-1.5 on the same GQA images to measure how question formulation influences reasoning behavior and accuracy.

The prompts within each tier exemplify lexical rewrites while keeping the image and the original answer constant, allowing for a more focused analysis on how VQA models behave with rewritten questions. Our manual analysis shows that some rewrites do not perfectly preserve the intent, and we quantify this shortcoming.[8] A careful analysis over three vision language models, BLIP, InstructBLIP, and LLaVA-1.5-7B, reveals significantly deteriorated accuracy with rewritten prompts, indicating a lack of robustness with respect to lexical variations. Due to imperfect intent preservation, the results should be read with some nuance as a sensitivity to lexical rewrites rather than a pure reading-level effect.[9] In addition to the questions asked above, AFGV brings forth questions often left unanswered in VQA practice: *Are VQA models subject to the same accessibility-driven robustness issues as humans when the questions are simplified? Does complicating the language provide artificial cues that exacerbate shortcuts? Are the popular VQA benchmarks inaccessible for the very people who can provide the most useful insights and the ones that can benefit the most from AI-based VQAs?*[10]

Methods

Pipeline

Figure 1 provides an overview of the AFGV construction pipeline which proceeds in four stages: First, questions are sampled from the GQA dataset and classified as the baseline tier. Second, Word2Vec semantic similarity is used to rewrite each question into two directions: complex and simplified. Third, all three versions are passed through GQA images across three different vision-language models: BLIP, InstructBLIP, LLaVA-1.5, producing answers for each tier. Finally, prompt intent similarity is estimated using all-MiniLM cosine similarity (threshold > 0.80)[11], and a manual review sample is used to assess how well this automated filter captures intent preservation.

GQA Dataset Overview

The GQA dataset, illustrated in Figure 2, provides a visual reasoning benchmark in the form of real images, scene graphs of the images, and a set of question-answers about those images grounded in the scene-graphs. GQA questions are generated using an automatic engine that composes a set of linguistic and structural templates operating on the scene-graph content, resulting in diverse questions that explore the objects, attributes, and relations present in each image[5,12]. Each GQA question is derived from the scene graph–nodes being objects, edges being their relation– such that it challenges the model to provide a multi-step reasoning path. This reasoning path provides diagnostics on failure modes and where the model's reasoning path diverges from the actual path. Thus, GQA provides a benchmark for compositional and relational reasoning that

Figure. 2 Examples of AFGV. Readability scores are reported in parentheses.

Original 1: What animal is eating from the field? (3.997)
Simplified 1: What animal *eats* from the field? (2.311)
Complex 1: What animal is eating from the *forest*? (5.628)
Original 2: What animal is in front of the grass? (2.280)
Simplified 2: What animal is in front of the *hay*? (2.280)
Complex 2: What animal is *reflected* in the grass? (5.230)

curbs the model from taking language shortcuts such as the exploitation of answer priors (e.g., when asked about the color of the sky, the model answers "blue"). For our AFGV evaluation pipeline, we use the subset of GQA's balanced training split comprising 943,000 questions. For our benchmark, we sample $N = 20{,}000$ unique image question pairs. We select this split because the balanced split was explicitly constructed by the GQA authors to reduce conditional biases and overrepresented answer distributions (e.g., skewed set of yes/no answers). This split alleviates the risk of the reported results being driven solely by imbalanced classes in the dataset, though not all dataset (or training data) biases are eliminated[13].

Metrics

GQA argues that accuracy obscures the genuine evaluation of VQA and that metrics such as answer plausibility, consistency, and image grounding are better suited for this task[5]. In accordance with their motivation, this paper reports (1) exact-match accuracy of VQA answers; (2) lexical complexity of the rewritten prompts measured by FK and ASW; (3) Prompt Intent Similarity, the cosine similarity between the original and the rewritten question embeddings; and (4) Answer Semantic Similarity, the cosine similarity between the model predictions and ground-truth answer embeddings. Accuracy evaluates the final performance of the model on the task, but does not assess linguistic perturbations. A decline in accuracy does not distinguish between a model that fails because it is faced with a prompt of a higher reading level and a model that fails because the original prompt is altered in the word swap process. To address these variables, we design two similarity measurements. Prompt Intent Similarity calculates the cosine similarity between the baseline and rewritten prompts' embeddings, reflecting the sensitivity of the prompt intents under lexical changes. Answer Semantic Similarity calculates the cosine similarity between the predicted answer and the ground-truth answer, reflecting the semantically correctness of the model output even if the exact match score is zero. We report Flesch–Kincaid Grade Level[10] as a measure of lexical complexity, not reading level. Given the brevity of the questions in the GQA dataset and that the pipeline changes only one word per question, FK is primarily affected by the change in the number of syllables of the replaced word. Thus, we consider ASW the main readability statistic and report FK for the sake of comparison. Average Syllables per Word (ASW) is computed as follows:

$$ASW = \frac{\text{Total Syllables}}{\text{Total Words}}$$

On the other hand, within the complex tier, Word2Vec substitutions commonly replace single syllable, high-frequency words with multi-syllabic, abstract words (e.g., "field" substituted by "environment").

Implementation Details

The baseline GQA training questions are run through a Word2Vec model on Google Colab. Although Word2Vec cosine similarity may provide a local approximation for semantic similarity, there is no guarantee that it preserves question intent. Since candidate substitutions are generated based on proximity in vector space independently of the question context, this can lead to semantic drift that may change the visual reasoning task. Readability is calculated using **textstat**, a Python package used to compute the readability; lower float values correspond to easier readability and the inverse as well[10]. Each candidate has its readability computed using textstat. We pick among candidates with the following ranking rules: (1) the candidate giving the lowest Flesch–Kincaid Grade Level is preferred; (2) if more than one candidate results in the same grade level, the shortest word is preferred (as a proxy for a fewer number of syllables and simpler vocabulary); and (3) if there is still a tie, the word with higher frequency is chosen. Word frequencies are computed over the GQA training dataset, the least frequent word in each question is the substitution target. Replacements are generated using semantically similar equivalents using Word2Vec and their readability scores are calculated. The word length is recorded as a proxy for lexical complexity.

The choice of grade level first, word length second, and

word frequency third, reflects the stated goal of the paper to manipulate the Flesch-Kincaid scores for the different levels of word complexity. The Flesch-Kincaid Grade Level formula relies on the structural features of total words over total sentences and the surface lexical features of total syllables over total words. Word length is used second in the event of a tie; when two semantic replacements yield the same structural readability score, the preference for the longer word (or shorter) uses word length as a surface-level proxy for lexical complexity. Finally, word frequency represents familiarity more than it does structural complexity. Less frequently used words result in longer fixation durations because they are less predictable. But, because infrequently used words do not necessarily translate to greater structural reading difficulty (i.e. short archaic words), it is relegated to the role of final tie-breaker.

Results

Baselines

We use a standard GQA benchmark for the main evaluation of compositional visual reasoning. Models are trained using a sigmoid-based classifier and the Adam optimizer to train the models for 15 epochs. For the Memory, Attention, and Composition (MAC) model, we use an open-source implementation. Before employing it for the AFGV readability tiers, we first verify that the implementation is correctly configured by using it on the standard VQA benchmark. This implementation reports **67%**, which is comparable to the performance reported in the original MAC paper[5]. Blind models that only observe the question text are evaluated to quantify how much performance can be obtained from linguistic cues alone[3] The large gap between blind and vision-based models suggests that our evaluation uses visual information beyond the language cues in the question alone, though it is not immune to language bias[9].To generate variations in prompts, we use the question generation framework of GQA, sampling across compositional question types and grammar templates to generate structurally similar prompts[5]. Simplified prompts are curated with the help of Word2Vec, a technique used to efficiently compute dense vector representations for words in a space where semantically similar words appear in close proximity[14]. To reduce the reading level of a prompt, we substitute uncommon words with more frequent words, a proxy for reduced vocabulary. Given vector representations of the original word A and replacement word B we calculate the cosine similarity:

$$\textit{cosine similarity}(A,B) = \frac{A \cdot B}{\|A\|\|B\|} \tag{1}$$

The ultimate size of the AFGV dataset consists of 20k distinct image-question pairs. Considering there are 943k questions in the train split of GQA balanced, our sample of 20k questions from GQA constitutes 2.12% of the split. Since our sample size is exactly $N = 20{,}000$, for each reported accuracy, we can compute SE and CI for the standard Wald formula for binomial proportions as follows:

$$SE = \sqrt{\frac{p(1-p)}{N}} \tag{2}$$

$$95\%\ CI = p \pm 1.96 \times SE \tag{3}$$

For each accuracy value, we report a 95% confidence interval using the Wald approximation for binomial proportions. These intervals provide an estimate of sampling uncertainty for each tier[15]. This paper applies classical statistical methods based on likelihood ratios to construct 95% confidence intervals, demonstrating how modern AI systems still rely on classical methods to ensure reliable predictions.

We then use Z, a standardized metric that indicates how many standard deviations an observed data point differs from the mean of a standard normal distribution. In the context of this experiment, in order for the results to be statistically significant a Z-score of 1.96 or higher is conventionally treated as evidence of a statistically detectable difference.

All three models suffer significant accuracy declines for both the rewritten simple and complex tiers, with no overlap in confidence intervals, indicating these results are not due to chance. BLIP has the largest declines (Δ vs Baseline = -0.110 for simplified prompts, -0.105 for complex prompts; $Z > 21$), followed by InstructBLIP (Δ vs Baseline = -0.069; $Z > 13$), and then LLaVA-1.5 (Δ vs Baseline = -0.040; $Z > 4$) where the lower Z-scores are due to the smaller number of samples rather than the smaller effect size. Curiously, the simplified prompts, which use a more common vocabulary, lead to larger accuracy declines than the complex prompts for BLIP. At present, we find it difficult to make a strong case for any one reason why this asymmetry occurs, and it may be due to necessary grammatical scaffolding being removed. The Minimum CI separation, equivalent to the minimum possible difference in performance accounting for uncertainty in the confidence intervals, is not applicable in some cases (N/A) because no comparison is made for the baseline tier as it is the reference.

Readability Analysis

To comprehend how the Word2Vec-based rewriting process affects prompt complexity, we conduct a comparative analysis of the original, simplified, and complex prompts of GQA questions. The simplified prompts replace the infrequent

Table 1 Exact match accuracy, 95% confidence intervals, and comparison of baseline, simplified and complex prompt tiers across BLIP, InstructBLIP, and LLaVA-1.5 are shown.

Model	Tier	Accuracy	95% CI	Δ vs Baseline	Z-score	Min. CI separation
BLIP	Baseline	0.533	0.526–0.540	N/A	N/A	N/A
BLIP	Simplified	0.423	0.416–0.430	−0.110	22.15	9.60
BLIP	Complex	0.428	0.421–0.435	−0.105	21.13	9.10
InstructBLIP	Baseline	0.503	0.496–0.510	N/A	N/A	N/A
InstructBLIP	Simplified	0.435	0.428–0.442	−0.068	13.72	5.45
InstructBLIP	Complex	0.434	0.427–0.440	−0.070	14.04	5.61
LLaVA-1.5	Baseline	0.283	0.271–0.296	N/A	N/A	N/A
LLaVA-1.5	Simplified	0.246	0.234–0.258	−0.038	4.29	1.34
LLaVA-1.5	Complex	0.241	0.229–0.253	−0.042	4.83	1.81

words in the original prompts with more frequent, meaning-wise carefully chosen words, resulting in a reduced Flesch–Kincaid grade level[14].

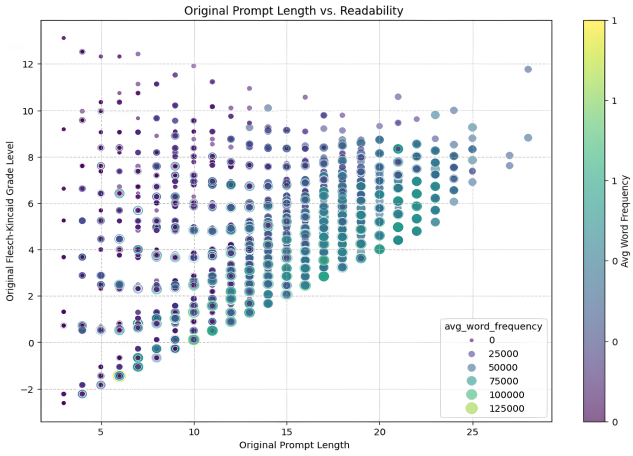

Figure. 3 Prompt length vs readability for original questions. Marker and color size indicate average word frequency.

The complex prompts, as shown in Figure 3, replace the meaning-wise selected common words in the original prompts with less frequent or more abstract words, resulting in an increased Flesch–Kincaid grade level. The simplified prompts exhibit consistently lower grade levels, verifying that the rewriting procedure decreases linguistic complexity. In contrast, the complex prompts increase the grade level.

For all three tiers of questions, we compute Flesch-Kincaid (FK) grade level, as well as raw counts of syllables, for 20,000 sampled questions in GQA. Given that 99.1% of the questions in all tiers have fewer than 20 words, FK is known to no longer function as a reading level beyond that point. We report FK scores for the sake of completeness, but use the syllable counts as our primary indicator of readability.

In Table 2, original questions have 10.53 syllables (FK = 2.10) , simplified questions have 9.94 (FK= 1.2), and the complex questions have 11.70 (FK = 3.94). The difference between the simplified and complex tiers–1.76 syllables– highlights that the two types of rewrites differ in their linguistic complexity. These results are very similar to those produced by the initial $N = 1{,}000$ run, indicating that the readability signal is fairly stable across sample sizes.

Table 2 Average syllable count by question tier.

Tier	Mean Syllables	Standard Deviation
Original	10.53	±4.29
Simplified	9.94	±4.18
Complex	11.70	±4.34

VQA Models on AFGV

Figure 3 shows that prompt readability is dominated by the length of the sentence. For long sentences, prompt readability is entirely driven by sentence length, as naturally complex, low-frequency words are present. For short sentences, prompt readability varies considerably with word choice and phrasing, as sentence length plays a comparably small role. Prompts of varying readability were tested on BLIP, InstructBLIP[9,16], and LLaVA-1.5[17].

Accuracy and Answer Semantic Similarity Analysis

While accuracy drops, semantic similarity remains high, indicating a dissonance in the evaluation method: BLIP predictions are often semantically equivalent or close in meaning (car vs. automobile) but are considered incorrect under a strict string-matching metric[18,19]. Recent work on more flexible generative VQA evaluation argues that exact-match scoring is overly restrictive to open-ended VQA, and semantic evaluators are needed to more fully assess model output[20]. Answer Semantic Similarity remains high across all prompt

variants, indicating that BLIP is consistently able to provide similar answers even when the exact match differs. Answer Semantic Similarity pertains to the cosine-similarity between predicted and ground truth embeddings. The relatively minor difference in Figure 3 indicates that the rewritten prompts are more similar to the original prompts than unrelated prompts, but this does not guarantee the preservation of the intent of the prompt. Word2Vec cosine similarity does not guarantee the preservation of the intent of the question. Visual inspection of prompts in Figure 2 shows that a few substitutions lead to a change in the meaning of the question (e.g. field vs. forest, in front of vs. reflected in), potentially changing the correct answer to the question. So, the drop in accuracy observed in the complex tier cannot be solely attributed to linguistic fragility of the model but also reflects structural difficulty and a change in the task. Future work must be mindful to either enforce a close match in contexts where the original sentence is altered or change the ground truth alignments to fully isolate linguistic fragility.

We show some examples of the answers produced by BLIP for original tier questions. Question: "*Which clothing item is gray?*", BLIP answer: "*shirt*", Ground truth answer: "*t-shirt*". Exact match evaluation returns False for this answer even though the answer is equivalent to the ground truth. Question: "What is the vehicle to the left of the fence?", BLIP answer: "*suv*", Ground truth answer: "*car*". 'False' is returned for an exact match here even though both answers refer to the same object in the scene. The above examples show that Answer Semantic Similarity captures the equivalence that exact match fails to capture and the difference in accuracy between tiers is partially due to the brittleness of the evaluation metric.

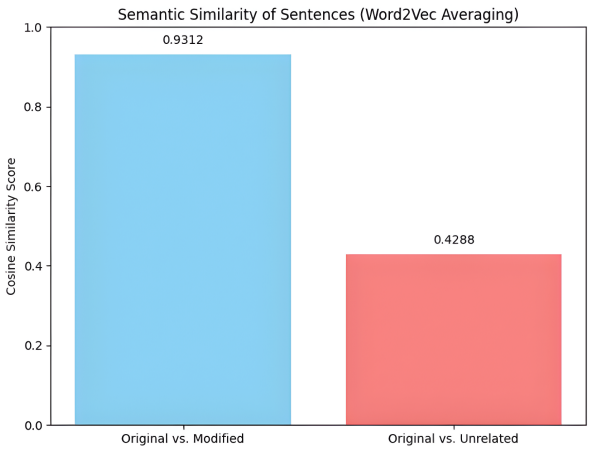

Figure. 4 Prompt Intent Similarity between Original and Modified Prompts.

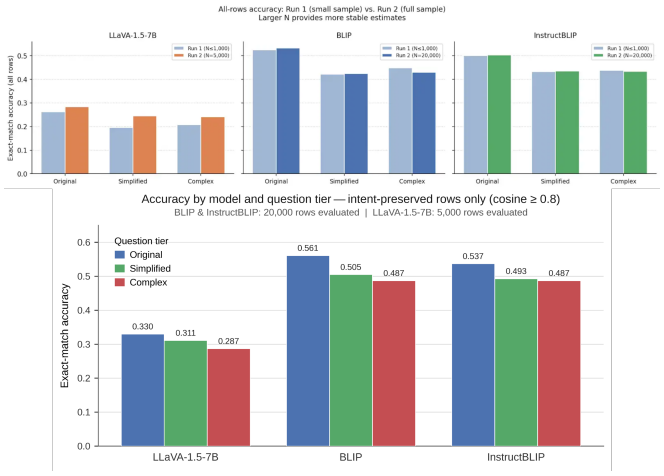

Figure. 5 Exact-match accuracy for original, simplified, and complex prompts across BLIP, InstructBLIP, and LLaVA-1.5.

Discussion

Early work on VQA primarily combined CNNs for images with RNNs for text. Teney, Liu, and den Hengel argue that a CNN+LSTM model does not capture multiple objects and are unable to perform multi-step reasoning. Their approach explicitly makes use of representations in the form of graphs[21]. In order to better push for the desired reasoning capabilities on datasets that are not sufficiently explored by existing datasets, new datasets are proposed. Among the CLEVRdatasets[22,23], CLEVR3D extends CLEVR to 3D focusing on transforming real scenes to feature spatial relationships. It is paired with a transformer-based model (TransVQA3D) to better address relational questions[24]. CRIC explores compositional reasoning across visual recognition to common-sense attributes. Automated generation of questions and construction of scene-graphs are leveraged to scale the construction of the dataset[25]. These datasets explore reasoning capabilities not adequately addressed by previous datasets. While prior work focuses on the structure of reasoning in VQA tasks, we focus on how lexical rewriting affects VQA performance. We on top of that evaluate the success of rewrites in preserving the original poster's intent. The gap between blind and vision models is large, indicating that language alone cannot capture the visual grounding that is relevant to human-driven visual reasoning[3].

For instance, an elementary prompt like "Bus color?" is less grammatically formed, while a more intricate prompt like, "What shade is the bus," substitutes less common synonyms and is typically longer[14]. Though both prompts convey the same meaning, BLIP fails on the rewritten prompts. Such differences suggest a bias towards familiar language structure and phrases over semantically strong concepts being grounded. The mistakes made with rephrasings that preserve

meaning indicate a disconnect that still needs to be closed between the performance of familiar phrases and visual comprehension[9]. This suggests that BLIP is sensitive to syntactical rephrasings and deviations in language structure[9].Near-uniform counts across the detailed, semantic,and structural categories.

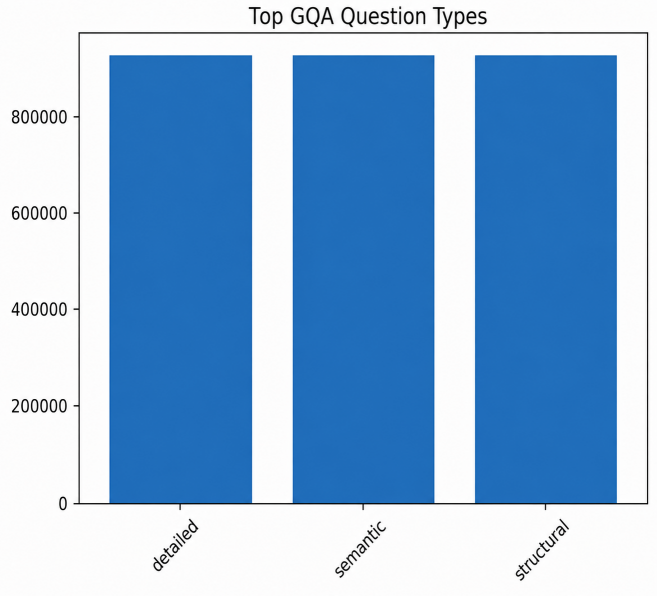

Figure. 6 Distribution of GQA question types.

Finally, we consider a second step to confirm that the intent is preserved to avoid intent drift using sentence level embeddings (all MiniLM-L6-v2, cosine 0.80)[11]. Across the full set of 20,000 questions, this filter removes 32.0% of complex rewrites that appear to have shifted the intent. Snippets like ”field” ”forest” and ”in front of” ”reflected in” are representative of ones removed by this filter. Importantly, the errors on intent shifted complex rewrites are 2.8x larger than the errors on intent preserved complex rewrites for BLIP ($\Delta = +0.183$ vs. $+0.066$), 2.5x larger for InstructBLIP ($\Delta = +0.183$ vs. $+0.066$), 2.3x larger for LLaVA-1.5-7B ($\Delta = +0.070$ vs. $+0.030$) showing that much of the original signal was confused by the difference in question semantics.

In Figure 7's red bars BLIP observes a 18.4% accuracy degradation attributed to the model answering a question that is not semantically equivalent. Still after removing the contaminated rows, there persists an accuracy drop for all three models (+0.066, +0.048, +0.030). Experiments on the filtered dataset show a smaller accuracy drop that still remains after variation in lexical construction, but the automatic filter is unable to fully be semantically equivalent to human evaluation, so these results must be taken cautiously . Running this test on a classification-head model (BLIP), an encoder-decoder (InstructBLIP), and a generative decoder (LLaVA), all ranging in size from 400M to 7B parameters, strengthens the argument that this is a generalized feature of the evaluation rather than a feature of one model. From a random sample of 100 questions (29% simplified, 50% complex), manual inspection found that the original intents could not be preserved for a significant fraction of questions. In 66% of simplified questions and 61% of complex questions, the cosine 0.80 sentence-embedding filters reject the questions, indicating that automatic filtering is not nearly as accurate a proxy for preserving the original question intent. Hence, the subsequent differences in accuracy across the tiers should be taken as a measure of sensitivity to lexical differences in general and not to linguistic complexity alone.

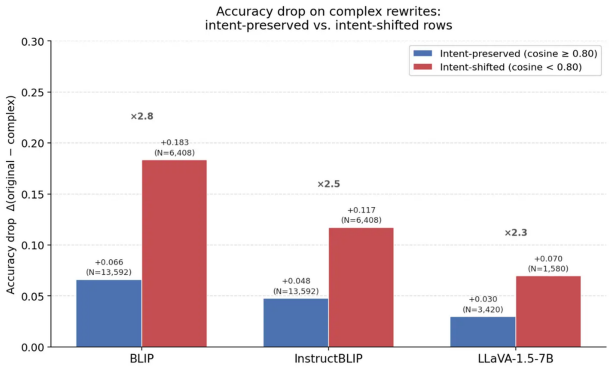

Figure. 7 Accuracy drop Δ (original-complex) across three models: BLIP, InstructBLIP, and LLaVA-1.5-7b. Red bars represent the accuracy reduction in complex rewrites ability to change the question's meaning. Blue bars represent the accuracy reduction after removing those contaminated rows.

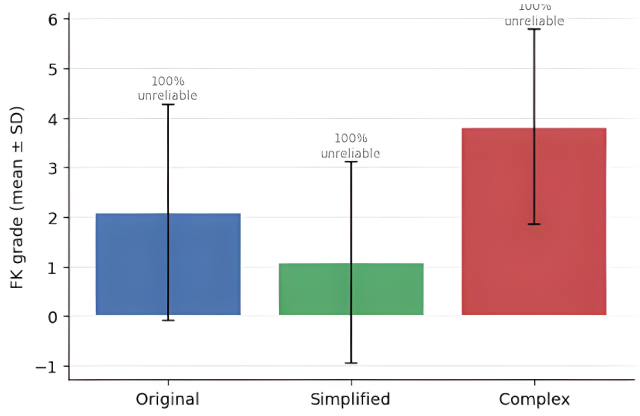

Figure. 8 Flesch–Kincaid readability scores for original, simplified, and complex prompts.(Average grade levels: 1, 0.5, and 2 respectively).

Table 3 Distribution of GQA question types. Accuracies across question types are near-uniform.

type	correct_base	correct_simp	correct_comp	delta_simp	delta_comp	N
structural	0.5607	0.5046	0.4867	0.0561	0.0740	7480
semantic	0.5607	0.5046	0.4868	0.0561	0.0739	7480
detailed	0.607	0.5045	0.4868	0.0562	0.0739	7480

While we report Flesch-Kincaid scores here on all three tiers, showing notable separation (mean grade levels of 0.5, 1.0, and 2.0 for simplified, original, and complex, respectively), we know that FK is not a reliable measure when applied to a single sentence. (Fig. 8) The average GQA question length is between 7 and 9 words, and the rewriting pipeline replaces one word from each question, resulting in the sentence length being approximately constant, and the resulting FK score correlating more so with the syllable count of the replaced word than with its readability. So, 100% of questions from all three tiers are below the sentence length where FK has been validated.

Conclusion

We introduce AFGV, a benchmark evaluating if varying linguistic difficulties through swapping words could expose robustness issues with VQA while maintaining the visual scene and ground truth constant. Our rewriting process aims for single word substitutions, but upon manual inspection the questions' meanings only approximately remain. Thus, differences in tiers can be attributed to sensitivity to linguistic changes as opposed to strictly isolated structural complexity. We test the accuracy of BLIP, InstructBLIP, LLaVA-1.5-7B and find that accuracy drops consistently for our intent-preserving rewrites Considering how the soft similarity is consistent across tiers, this indicates that the models struggle both due to question intent not being preserved and the model being prone to vocabulary shifts.

Acknowledgements

The author would like to thank Inspirit AI for its support and my mentor, Kimmy C. De Alba, for her guidance throughout this project.

References

1 H. Sterz, J. Pfeiffer, I. Vulic. DARE: Diverse Visual Question Answering with Robustness Evaluation. Transactions of the Association for Computational Linguistics. Vol. 13, 2025, https://aclanthology.org/2025.tacl-1.52/.

2 S. Whitehead, H. Wu, Y. R. Fung, H. Ji, R. Feris, K. Saenko. Learning from lexical perturbations for consistent visual question answering. arXiv preprint arXiv:2011.13406, 2020, https://arxiv.org/abs/2011.13406.

3 R. Cadene, C. Dancette, H. Ben-Younes, M. Cord, D. Parikh. RUBi: Reducing Unimodal Biases for Visual Question Answering. Advances in Neural Information Processing Systems. Vol. 32, 2019, http://arxiv.org/abs/1906.10169.

4 S. J. Unni, R. Moraffah, H. Liu. VQA-GEN: A Visual Question Answering Benchmark for Domain Generalization arXiv:2311.00807, 2023, *https://arxiv.org/abs/2311.00807*.

5 D. A. Hudson, C. D. Manning. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Pg. 6700-6709, 2019, http://arxiv.org/abs/1902.09506.

6 Y. Zhao, Y. Zhang, R. Xiang, J. Li, H. Li. VIALM: A Survey and Benchmark of Visually Impaired Assistance With Large Models. arXiv preprint arXiv:2402.01735, 2024, https://arxiv.org/abs/2402.01735.

7 A. Karamolegkou, M. Nikandrou, G. Pantazopoulos, D. Sanchez Villegas, P. Rust, R. Dhar, D. Hershcovich, A. Søgaard. Evaluating Multimodal Language Models as Visual Assistants for Visually Impaired Users. Association for Computational Linguistics. Vol. 1, pg. 25949-25982, 2025, *https://aclanthology.org/2025.acl-long.1260.pdf*.

8 Z. Lu, Q. Zeng, M. Lu, G. Chen, Y. Xia. Bridging the Semantic Gap in Medical Visual Question Answering with Prompt Learning. IEEE Transactions on Medical Imaging. Vol. 44, pg. 4605-4616, 2025, *https://pubmed.ncbi.nlm.nih.gov/40526558/*.

9 J. Li, D. Li, C. Xiong, S. C. Hoi. BLIP: Bootstrapping Language-Image Pre-Training for Unified Vision-Language Understanding and Generation. Proceedings of Machine Learning Research. Vol. 162, pg. 12888-12900, 2022..

10 L. Si, J. Callan. A Statistical Model for Scientific Readability. Proceedings of the Tenth International Conference on Information and Knowledge Management (CIKM). Pg. 574-576, 2001 *https://dl.acm.org/doi/10.1145/502585.502695*.

11 W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, M. Zhou. MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. Advances in Neural Information Processing Systems. Vol. 33, pg. 5776-5788, 2020, *https://arxiv.org/abs/2002.10957*.

12 S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, D. Parikh. VQA: Visual Question Answering. Proceedings of the IEEE International Conference on Computer Vision (ICCV). Pg. 2425-2433, 2015, *http://arxiv.org/abs/1505.00468*.

13 Z. Wang, L. Chen, H. You, K. Xu, Y. He, W. Li, N. Codella, K. Chang, S. Chang. Dataset Bias Mitigation in Multiple-Choice Visual Question Answering and Beyond. Association for Computational Linguistics: EMNLP. Pg. 8598-8617, 2023, *https://aclanthology.org/2023.findings-emnlp.576/*.

14 T. Mikolov, K. Chen, G. Corrado, J. Dean. Efficient Estimation of Word Representations in Vector Space. International Conference on Learning Representations (ICLR) Workshop, 2013, *https://arxiv.org/pdf/1301.3781*.

15 L. Sluijterman, E. Cator, T. Heskes. Likelihood-Ratio-Based Confidence

Intervals for Neural Networks. Machine Learning. Vol. 114, article 116, 2025, https://doi.org/10.1007/s10994-024-06639-3.

16 W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. N. Fung, S. C. Hoi. InstructBLIP: Towards General-Purpose Vision-Language Models with Instruction Tuning. Advances in Neural Information Processing Systems. Vol. 36, pg. 49250-49267, 2023, https://dl.acm.org/doi/10.5555/3666122.3668264.

17 H. Liu, C. Li, Y. Li, Y. J. Lee. Improved Baselines with Visual Instruction Tuning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Pg. 26296-26306, 2024, https://arxiv.org/abs/2310.03744.

18 J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, R. Girshick. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Pg.1988-1997, 2017, DOI: 10.1109/CVPR.2017.215.

19 S. Ging, M. A. Bravo, T. Brox. Open-ended VQA benchmarking of vision-language models by exploiting classification datasets and their semantic hierarchy. International Conference on Learning Representations, 2024, https://arxiv.org/abs/2402.07270.

20 H. Ji, Q. Si, Z. Lin, W. Wang. Towards Flexible Evaluation for Generative Visual Question Answering. arXiv preprint arXiv:2408.00300, 2024 https://arxiv.org/abs/2408.00300.

21 D. Teney, L. Liu, A. van den Hengel. Graph-Structured Representations for Visual Question Answering. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Pg. 1-9, 2017, http://arxiv.org/abs/1609.05600.

22 L. Salewski, A. Koepke, H. Lensch, Z. Akata. CLEVR-X: A Visual Reasoning Dataset for Natural Language Explanations. Springer International Publishing. Pg. 69-88, 2022, http://dx.doi.org/10.1007/978-3-031-04083-2_5.

23 A. Lindstrom, S. Abraham. CLEVR-Math: A Dataset for Compositional Language, Visual and Mathematical Reasoning. arXiv preprint arXiv:2208.05358, 2022, https://doi.org/10.48550/arXiv.2208.05358.

24 X. Yan, Z. Yuan, Y. Du, Y. Liao, Y. Guo, S. Cui, Z. Li. Comprehensive Visual Question Answering on Point Clouds Through Compositional Scene Manipulation. IEEE Transactions on Visualization and Computer Graphics. Vol. 30, pg. 7473-7485, 2024, https://dl.acm.org/doi/abs/10.1109/TVCG.2023.3340679.

25 D. Gao, R. Wang, S. Shan, X. Chen. CRIC: A VQA Dataset for Compositional Reasoning on Vision and Commonsense. IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol. 45, pg. 5561-5578, 2023, https://arxiv.org/abs/1908.02962.