Table 1: Overview of existing generalized VQA surveys

Name	Year	Challenges & Open Problems	Contributions
Wu et al. [33]	2017	Question Constraints, Visual and Textual Understanding, External Knowledge, Preference for Computer Vision-based methods	Comprehensive, the most cited VQA survey Categorization and generalization of early datasets and methods Discusses emerging works like Structure Scene Text Annotation [80]
Kafle and Kanan [32]	2017	Method superiority, Dataset Bias, Attention in VQA, Open-Ended (OE), and Multiple Choice (MC) Evaluation	Emphasis on problem formulation related to Vision and Language tasks Qualitative comparison of methods and evaluation metrics Elaborative discussion on research challenges with recommendations
Gupta [81]	2017	Answer Type Prediction Models, Hybrid models, Answer Generation	Simple and straightforward survey providing an introductory view Focuses on a few important datasets and methods
Teney et al. [82]	2017	Dataset Bias, Zero-Shot VQA, External Knowledge, Modular Architectures	Reviews fundamental techniques with phase-by-phase generalization Emphasizes attention-based and memory-augmented architectures Highlights advanced methodologies and domain trends
Hassantabar [83]	2018	Complex reasoning, short-term memory and counting-based questions	Introductory survey similar to [81] focusing on few datasets and models
Manmadhan and Kovoor [84]	2020	R-CNN [85] based Image Featurization, Out of Vocabulary words, Transformer-based Architectures, Sentence-based Embeddings	Guide for newcomers expositing fundamental concepts Highlights computer vision subtasks to solve VQA Phase-wise comparison of methods in different VQA architectures
Sharma and Jalal [86]	2021	OE and MC Evaluation, Dataset Bias, real VQA Image Featurization, Conversational, and Scene Text Questions	Compares traditional VQA models to scene text VQA models Detailed result analysis on 13 prominent VQA datasets Introduces a few open challenges in the domain
Srivastava et al. [87]	2021	Incorporating CV and NLP strategies, Real-life Datasets	Highlights major breakthroughs in VQA Discussions and analysis based on architectural paradigms

Table 4: Contributions and limitations of traditional VQA datasets

Name	Contributions	Limitations
DAQUAR [93]	First VQA benchmark to attempt Visual Turing Test [98] Various question categories with extensible templates	Insufficient data to train large models Limited to indoor scenes with unfavorable lighting conditions Questions are restricted to templates and answers are limited to classes Complicated model evaluation due to multiple metrics
COCO-QA [24]	Larger dataset with standardized image source [67] QA algorithm is extensible to other image captioning datasets [118, 119] Easier evaluation due to formulation as classification problem	Unnatural and grammatically inaccurate questions Limited question diversity Answers are limited to a single word only
Visual Madlibs [104]	Proposes the novel task of fill-in-the-blanks (FITBS) with multiple choices Diversified questions prompts	Insufficient answers for open-ended evaluation FITBs based on declarative sentences are easily answerable
FM-IQA [105]	Multilingual Dataset (English and Chinese) Free-form questions with diversified answer choices Rigorous quality assurance for Chinese QA pairs	Visual turing test-based manual evaluation is unscalable English QA pairs may not be accurate due to automated translation
VQA [1]	Benchmark for free-form VQA used for evaluating many models Diversified dataset with realistic and synthetic images High answer-to-question ratio with automatic evaluation	Unbalanced dataset resulting in questions answerable without images Lack of reasoning-based and complex questions Subjective questions without a single correct answer
Visual7W [101]	Introduces the task of visual grounding QA diversity corresponding to multiple standard vision tasks	Lacks binary (yes/no) questions Wide performance gap between humans and AI
Visual Genome [80]	Largest free-form dataset based on QA pairs Diversified QA pairs on multiple image regions Attribute and Relationship-based QA pairs using Scene Graphs	Difficult to evaluate long answers Inherently too large resulting in preference for its subset [101]
VQA v2 [10]	Introduces counter-examples to create a balanced dataset Reduction of Language bias as seen in VQA [1] Counter-examples can be used as a modality for model explainability	Lacks questions on general knowledge Insufficient reasoning-based questions especially on synthetic data Question category biases[1]might result in poor real-world performance [109]
TDIUC [109]	Wide category of questions with a balanced distribution Absurd/meaningless questions for image-based reasoning Evaluation strategy counters question-category biases	Similar questions belonging to a certain category, especially colors The majority (around 40%) are binary questions on object presence Manual annotations come from a small sample space

[1]An abundance of a certain category of questions like "Is/Are" questions will result in models being trained better in that particular question category.