# BridgeEQA: Virtual Embodied Agents for Real Bridge Inspections

## Supplementary Material

## 1. Evaluating Image Citation Relevance for Human Alignment

To validate that our Image Citation Relevance metric aligns with human judgment, we conducted a manual annotation study on a randomly sampled set of 100 question-answer pairs from BridgeEQA. For each sample, we randomly perturbed the reference images by introducing varying numbers of random images from the original PDF document set and then randomly removing a varying number of images. This perturbation process generated image sets spanning the full relevance spectrum—from completely irrelevant to fully relevant to the question.

Three annotators independently labeled each sample on a 5-point scale, which we normalized to the 0.0-1.0 range to match the Image Citation Relevance output range. We then computed Image Citation Relevance scores for the same dataset using Gemini-2.5-flash as the evaluator model.

The Spearman correlation between the averaged human annotations and Image Citation Relevance scores was 0.817, demonstrating strong alignment between our automated metric and human judgment of image relevance.

## 2. Dataset Example

We provide an example from BridgeEQA in Figure 1. The reference images are images parsed from the source report and the condition rating is extracted from the answer. This particular scene graph for structure 0010 in Chelsea has a total of 53 nodes.

## 3. Dataset Creation Details

### 3.1. Example Source Inspection Reports

We provide sample pages from Vermont bridge inspection report in Figure 2 as an example source report for BridgeEQA.

### 3.2. Data Collection

We collected bridge inspection reports from the Vermont Agency of Transportation (VTrans) public database, which contains unstructured PDF inspection reports covering bridges across Vermont. Each report documents the condition of a single unique bridge and includes inspector observations, condition ratings, and photographic documentation.

### 3.3. Stage 1: Preprocess and Filter

The preprocessing stage applies quality control filters to ensure that selected reports contain sufficient visual documentation for a meaningful infrastructure assessment.

**Report-Level Filtering.** Initial qualitative evaluation of the inspection reports revealed significant variability in visual documentation quality and comprehensiveness. To ensure sufficient visual coverage for meaningful condition assessment, we applied a minimum threshold of 20 images per report. Reports failing to meet this criterion were excluded based on several quality indicators:

- **Incomplete visual coverage:** Reports with fewer images often documented only limited perspectives of the bridge, missing critical structural components necessary for comprehensive assessment.
- **Low image quality:** Sparse image sets frequently exhibited poor resolution, unfavorable lighting conditions, or obstructed views that would hinder reliable condition evaluation.
- **Non-standard outlier conditions:** Some reports documented bridges that were demolished, under major reconstruction, or otherwise not representative of typical operational infrastructure.

**Page Filtering.** Page filtering removes the first two pages of each report, which typically contain administrative cover pages, title pages, and summary information without detailed inspection content or photographic documentation.

**Image-Level Filtering.** Within the quality-controlled reports, individual images underwent additional filtering. Images smaller than 200x200 pixels were systematically removed from the dataset. This threshold was established through empirical observation that sub-threshold images predominantly contained:

- Organizational logos and branding elements
- Document headers and administrative markings
- Thumbnails and preview images lacking structural detail

Such images provide minimal information for infrastructure condition assessment and could introduce noise into model training or evaluation.

**Random Sampling.** From the filtered pool of quality-controlled reports, we employed a random sampling strategy to select 200 reports for the final dataset. This sampling approach ensures representative coverage of Vermont's bridge inventory while maintaining computational tractability for annotation and evaluation.

### 3.4. Stage 2: Extract

The extraction stage processes filtered PDFs to obtain textual and visual content. Text extraction parses inspector notes, observations, and structured fields. Image extraction retrieves photographs meeting quality criteria, preserv-

BridgeEQA Example: Structure 0010 in Chelsea

**Question:** Given the transverse cracking in the asphalt wearing surface where no joints are present, what are the resulting deterioration patterns—such as cracking and abrasion—observed on the abutments, and what is the overall condition rating for the substructure?

**Answer:** Asphalt over the structure has transverse cracking forming in various areas scattered throughout both sides over the abutments. The substructure below shows related deterioration. Reinforced concrete abutment #1 has multiple settlement cracks present, and several areas along the lower portions exhibit minor abrasion. Reinforced concrete abutment #2 has a full-height vertical crack, and its lower portions have heavy abrasion/scaling that penetrates as deep as approximately 15". Because all primary structural elements are sound but have minor section loss and cracking, the substructure is in **Fair Condition** with a rating of **5**.

**Reference Images:**

- 636045817baf4010d986b01a55baca1d.png
- 1de1058cd997112823da497df50cc66b.png
- 11a2bf060c346794067d43bc7718caad.png
- 3c435b2e6814c354430fee1480befc8d.png

**Extracted Condition Rating:** 5

**Scene Graph:**


```
{
  "nodes": [
    {
      "image_name": "13b13f0ccab54674d56471f83082c47c.png",
      "central_focus": "Upstream view of Jail Brook from under the bridge.",
      "image_description": "View from under the bridge looking upstream...",
      "edges": [
        {
          "connected_to": "95b6875b4d3e1bcc4d524926a13b0157.png",
          "description_of_connection": "adjacent_to"
        },
        {
          "connected_to": "be36bcfe248d002bf1b6d2a62e6b20ae.png",
          "description_of_connection": "provides_context_for"
        }
      ]
    },
    ...
  ]
}
```


Figure 1. We provide an example from BridgeEQA on structure 0010 in Chelsea. This particular example has a scene graph with 53 nodes.

ing metadata about location and context. This stage yielded 9,586 images across 200 reports, averaging 47.93 per report.

### 3.5. Stage 3: Transform

The transformation stage structures the extracted content into standardized formats with ground truth annotations. We employ several vision-language models as zero-shot parsing tools to extract structured information from the inspection reports. Gemini 2.5 Flash and Gemini 2.5 Pro [3] serves as the primary extraction model for its efficiency and quality, with no fine-tuning or training on the bridge inspection data. The models function purely as information extraction tools, parsing existing content rather than learning dataset-specific patterns. We found Gemini 2.5 Flash to frequently have parsing errors or hallucinations as context size's grew in this stage, as such we fall back to Gemini 2.5 pro to reprocess when these errors occur and drop reports if errors persist.

**Image Reference Mapping.** This component links pho-

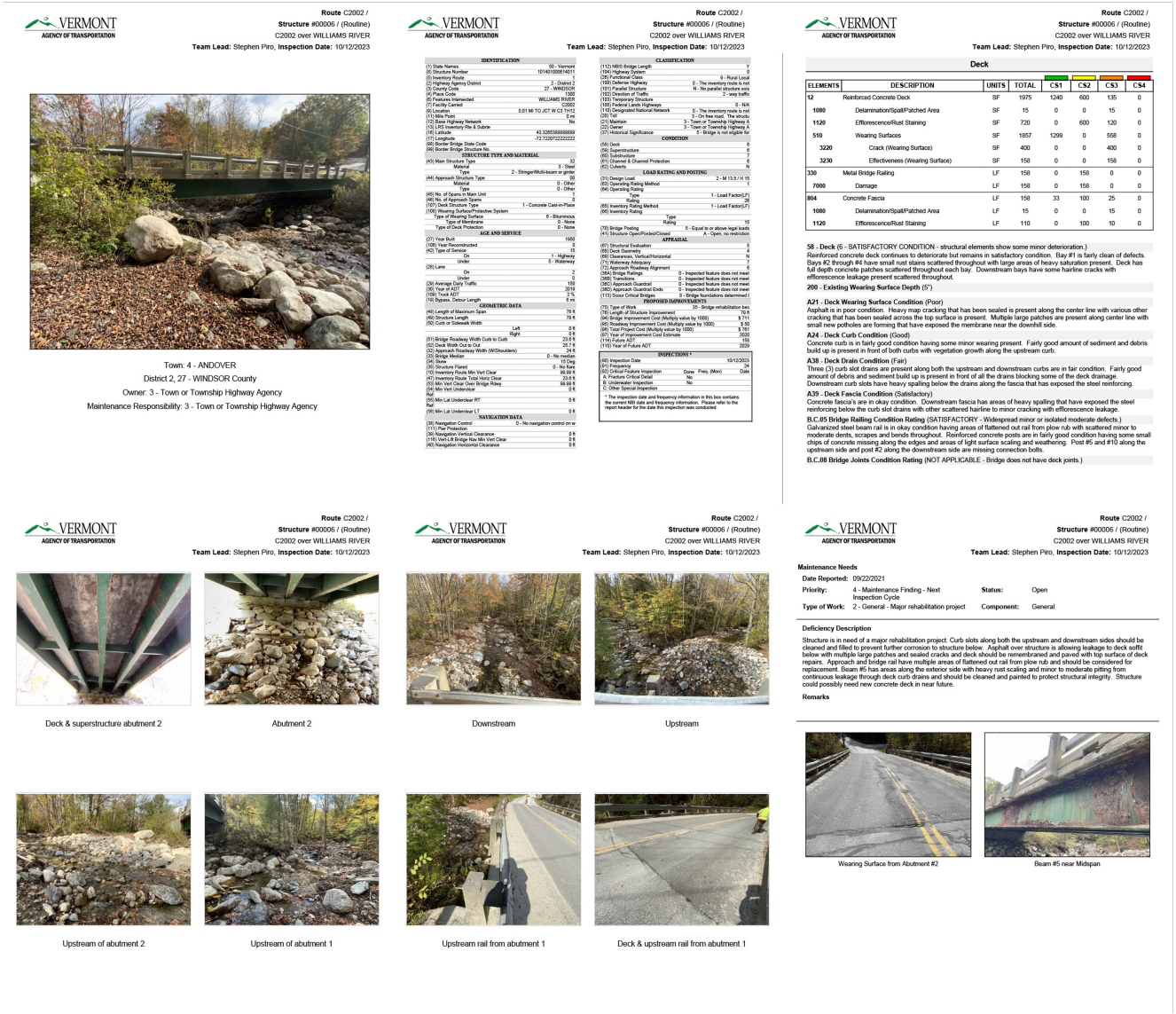

Figure 2. Sample pages of BridgeEQA source report originating from the Vermont Agency of Transportation's (VTrans) inspection report for Structure 00006, located in Andover.

tographs to corresponding textual descriptions in inspector notes, supporting scene formation where multiple images document the same infrastructure component. This step is required to allow grounded questions that use real references for component names, such as Abutment 1.

**Condition Rating Extraction.** This component parses component-level NBI ratings from inspector assessments, providing ground truth labels on the standardized 0-9 scale [5].

**Inspector Note Preservation.** Inspector notes are preserved to maintain the original context and rationale for condition assessments, ensuring that ground truth annotations remain grounded in the source documentation. We leverage these notes to ensure QA generation is grounded to real statements in the report.

### 3.6. Stage 4: Validate

Human quality control checks verify data integrity before QA generation. Additionally we test for any false or hallucinated image references. Parsing error detection identifies reports with corrupted text extraction, malformed condition ratings, or broken image-text mappings. When parsing errors or missing image references are detected, the report is automatically reprocessed using Gemini 2.5 Pro [3] as a fallback model for more robust extraction. Both Flash and Pro are used solely as zero-shot parsing tools without any

training or fine-tuning, ensuring that evaluation results reflect genuine visual reasoning capabilities rather than memorization. Reports that fail validation after reprocessing were removed from the dataset.

### 3.7. Stage 5: Generate QA

The final stage generates structured question-answer pairs for evaluation. Using a Gemini 2.5 Flash and Pro, we create questions grounded in the inspection report content, spanning condition assessment, component identification, and defect description tasks. Each answer includes the ground truth response sourced from inspector notes, references to supporting images, and the associated NBI condition rating when applicable. Quality checks verify that all referenced images exist and that answers are properly grounded in the available evidence.

### 3.8. Data Generation Validation

To ensure QA quality, we employ several evaluation metrics. We use the RAGAs [4] metrics: Faithfulness, which measures how well answers are grounded in the provided context, and Answer Relevancy, which assesses how effectively answers address the posed questions. We also incorporate the Answerability metric from RAGalyst [6] to determine whether questions can be adequately answered given the available context. To assess domain specificity, we employ LLM-as-a-Judge to determine Inspector Relevancy (0.0-1.0). This score measures the direct applicability of the question and its associated answer for bridge inspectors.

After evaluating all QA's with Gemini-2.5-flash, we reach a Faithfulness of 0.997, an Answer Relevancy of 0.997, an Answerability of 0.996, and an Inspector Relevancy of 0.980. These high scores across all metrics demonstrate the overall high quality of the dataset.

### 3.9. Human Validation

To validate the automated filtering and processing pipeline, human evaluation was conducted on a random subsets of the processed reports. This manual inspection verified that the quality-controlled reports met the following standards:

- Sufficient visual coverage of critical bridge components
- Adequate image quality for condition assessment
- Consistency with typical operational bridge inspection documentation
- Accurate representation of the condition rating labels

## 4. Effects of Scene Graph Connectivity on Condition Rating

We provide the accuracy heatmap of each method and VLM by the number of nodes in the scene graph in Figure 3 and by the number of edges in the scene graph in Figure 4.

Table 1. Condition rating exact match accuracy (%) on open-source VLMs evaluated on BridgeEQA instances with fewer than 30 images.

| Method | Qwen3-VL 8B-Thinking[1]† | Qwen3-VL 30B-A3B[1] | Nemotron-3 Nano-30B-A3B[2] |
|---|---|---|---|
| Multi-Frame VLM | 9.1 | 8.5 | 11.0 |
| Socratic LLM w/ SG | 36.4 | 6.1 | 11.0 |
| EMVR VLM w/ SG Only | 40.9 | **29.3** | **23.2** |

†Due to its limited context window, the Qwen3-VL 8B model was evaluated only on scenes with fewer than 30 images, a small fraction of the full dataset. These results are not directly comparable to other models reported in this paper.

Table 2. Condition rating within $\pm 1$ accuracy (%) on open-source VLMs evaluated on BridgeEQA instances with fewer than 30 images.

| Method | Qwen3-VL 8B-Thinking[1]† | Qwen3-VL 30B-A3B[1] | Nemotron-3 Nano-30B-A3B[2] |
|---|---|---|---|
| Multi-Frame VLM | 81.8 | 23.2 | 58.5 |
| Socratic LLM w/ SG | 72.7 | 30.5 | 48.8 |
| EMVR VLM w/ SG Only | 81.8 | **76.8** | **70.7** |

†Due to its limited context window, the Qwen3-VL 8B model was evaluated only on scenes with fewer than 30 images, a small fraction of the full dataset. These results are not directly comparable to other models reported in this paper.

Generally, the performance across methods decreases as the number of edges and nodes increase. This is due to the increased context sizes which is known to reduce VLM performance. However, EMVR performance degrades less at higher node and edge counts since EMVR mitigates the "lost in the middle" problem.

## 5. Open-Source Model Results

We extend our evaluation to open-source VLMs (Vision-Language Models) to assess generalizability beyond proprietary models. Given the large context windows required by our dataset, we omit Multi-Frame VLM w/ SG and EMVR VLM w/ Images + SG, as both require encoding images alongside the scene graph, which exceeds the context window of these models. Additionally, these models exhibited high failure rates in structured output generation, hallucinated function calls, and repeated the same actions in loops during agent execution. Due to these limitations, only a fraction of BridgeEQA could be tested. Results in 1 and 2 should therefore not be compared against the main paper results.

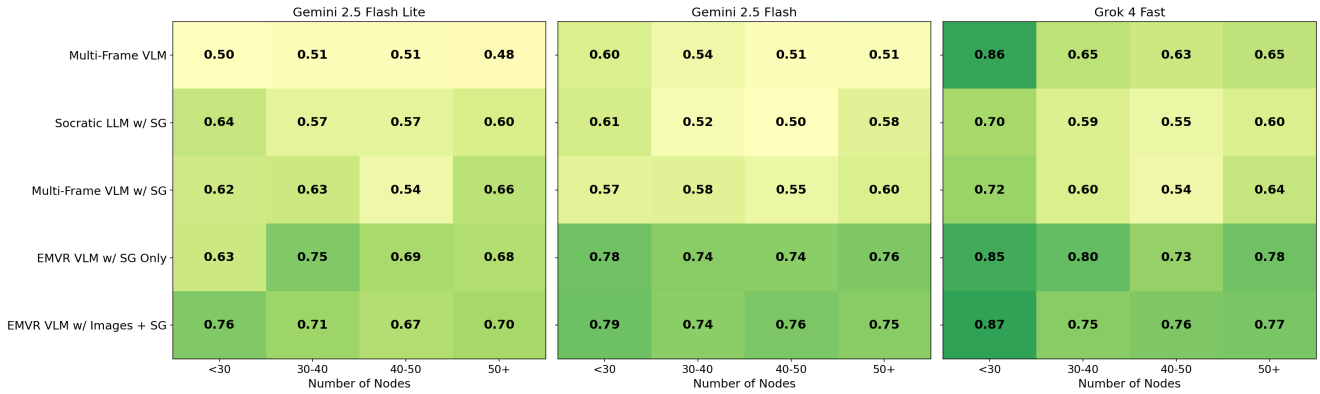


Figure 3. Condition rating within $\pm$ 1 accuracy heat map across method, VLM, and number of nodes.

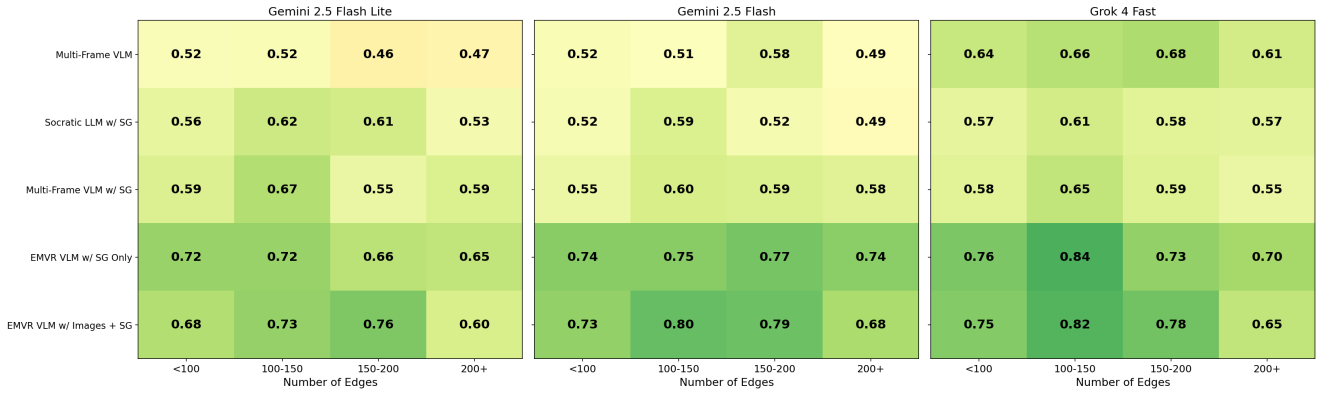


Figure 4. Condition rating within $\pm$ 1 accuracy heat map across method, VLM, and number of edges.

## References


[1] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025. 4

[2] Aaron Blakeman, Aaron Grattafiori, Aarti Basant, Abhibha Gupta, Abhinav Khattar, Adi Renduchintala, Aditya Vavre, Akanksha Shukla, Akhiad Bercovich, Aleksander Ficek, et al. Nvidia nemotron 3: Efficient and open intelligence. *arXiv preprint arXiv:2512.20856*, 2025. 4

[3] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 2, 3

[4] Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. RAGAs: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, St. Julians, Malta, 2024. Association for Computational Linguistics. 4

[5] Federal Highway Administration. Recording and coding guide for the structure inventory and appraisal of the nation's bridges. Technical Report FHWA-PD-96-001, U.S. Department of Transportation, Federal Highway Administration, 1995. 3

[6] Joshua Gao, Quoc Huy Pham, Subin Varghese, Silwal Saurav, and Vedhus Hoskere. Ragalyst: Automated human-aligned agentic evaluation for domain-specific rag. *arXiv preprint arXiv:2511.04502*, 2025. 4