

DRAWINGVQA: A Real-World Benchmark for Multi-Depth Visual–Textual Reasoning on Construction Drawings

Supplementary Material

Contents

A Full Main Results	2
B Author contributions	2
C Limitations and future work	3
D Benchmark details	3
D.1. Accessing DRAWINGVQA	3
D.2. General Guidelines	3
D.3. Question curation guidelines	4
D.4. Dataset VQA Structure	4
D.5. Training contamination mitigation	5
D.6. Question Format: Multiple-Choice vs. Open-Ended	5
D.7. Generalization to Unseen Drawing Styles	5
D.8. Ethical Consideration	5
E Dual Categorization Taxonomy	5
F. Experiments details	6
F.1. Evaluation Prompts and Parsing	6
F.2. Model details	7
F.3. Human baseline	7
F.4. Dense and MoE models	8
F.5. No-Image Ablations	8
F.6. Breakdown results on Dual Category Mapping	9
F.7. More experiments on Image resolution	10
F.8. More experiments on on Option Permutation and Prefix Sensitivity	10
F.9. Ablation Study on Input Modality: The PDF Hybrid Setting	11
G Error Analysis	13
G.1. Visual Perception errors	13
G.2. Visual and Text Perception errors	14
G.3. Text Perception (OCR) errors	15
G.4. Knowledge errors	16
G.5. Reasoning errors	17
G.6. Instruction Adherence and Output Formatting	19
G.7. Parametric Bias vs. Visual Grounding	19
G.8. Visual Attention	21

A. Full Main Results

Table 1 presents the comprehensive evaluation of all models across three reasoning depths (R1–R3), four MLLM dimensions, and seven construction-specific domains.

- **Gaps Between Theory and Practice:** Previous studies have demonstrated that models like GPT-4o pass construction certification exams with nearly 90% accuracy [11]. However, our results show a significant performance drop when these models are tasked with deciphering engineering drawings. For instance, GPT-4o achieves 48.9% on DRAWINGVQA. This highlights that while current MLLMs possess strong domain knowledge retrieval capabilities, they fundamentally lack the spatial and semantic visual reasoning required to execute real-world engineering workflows. This discrepancy aligns common notion that MLLMs excel at *Knowledge* (their strongest dimension), rather than the visual reasoning required to execute real-world engineering workflows.
- **Multimodal Architecture:** Among the evaluated architectures, Gemini and Qwen family achieve state-of-the-art performance. This suggests that the native multimodal architecture of the family may have better visual grounding and model architecture for dense technical documents compared to others. Especially, Gemini-3-pro-preview¹ shows improved knowledge and OCR capability in their modality, leading to the best result (77.2%), which now moves much closer to the graduate and young professionals’ benchmark score.
- **Benchmarking Against Human Expertise:** Most MLLMs significantly lag behind Graduate & Young professionals (78.2%) and the Professional Expert baseline (94.9%). This gap between the best AI and industry experts indicates that while MLLMs are approaching the competency of an entry-level engineer, they are not yet reliable enough for autonomous professional practice.

Detailed model size impacts and qualitative error analysis are provided in the subsequent sections.

Table 1. Full Main Results. **MLLM Dimensions** (VP: Visual Perception, K: Knowledge, R: Reasoning, OCR: Optical Character Recognition); **Construction Eng. Domain Aspects** (Adm.: Administration, EI: Element Identification, LU: Domain Language Understanding, DR: Dimensional Recognition, SU: Design Semantic Understanding, QTO: Quantity Take Off, Comp.: Compliance). Underscored values indicate the highest score for each model within the MLLM Dimensions and CE Domain Aspects categories.

Model	Overall	Reasoning Depth			MLLM Dimensions				Construction Eng. Domain Aspects						
		R1	R2	R3	VP	K	R	OCR	Adm.	EI	LU	DR	SU	QTO	Comp
GPT-4o	48.9	61.3	44.1	40.7	47.1	59.4	44.3	48.7	56.3	32.4	70.8	41.9	50.0	25.0	42.9
o3	58.7	64.5	61.8	48.2	58.8	<u>65.6</u>	54.1	59.2	62.5	56.8	54.2	48.4	61.5	25.0	42.9
Gemini-2.5-pro	71.7	80.6	76.5	55.6	67.6	<u>78.1</u>	67.2	72.4	<u>87.5</u>	56.8	79.2	67.7	75.0	41.7	57.1
Gemini-2.5-flash	66.3	77.4	73.5	44.4	64.7	<u>81.2</u>	57.4	63.2	<u>87.5</u>	51.4	79.2	51.6	67.3	33.3	71.4
Claude-4.5-Sonnet	57.6	64.5	64.7	40.7	58.8	<u>59.4</u>	57.9	54.1	<u>68.8</u>	48.7	66.7	41.9	57.7	41.7	57.1
Claude-4.5-Haiku	54.3	51.6	73.5	33.3	52.9	<u>65.6</u>	52.5	54.0	<u>62.5</u>	43.2	50.0	48.4	51.9	33.3	42.9
LLaVA-OneVision-1.5-8B-Instruct [1]	40.2	61.3	41.2	14.8	38.2	<u>46.9</u>	31.1	39.5	43.8	24.3	<u>62.5</u>	25.8	42.3	0.0	0.0
LLaVA-v1.6-Mistral-7B [5]	28.3	25.8	29.4	29.6	29.4	<u>37.5</u>	27.9	23.7	31.2	24.3	16.7	<u>32.3</u>	30.8	25.0	14.3
LLaVA-v1.6-34B [5]	42.4	38.7	44.1	44.4	50.0	<u>53.1</u>	42.6	40.8	<u>43.8</u>	37.8	41.7	35.5	40.4	33.3	42.9
Llama-3.2-11B-Vision-Instruct [6]	39.1	45.2	47.1	22.2	<u>41.2</u>	40.6	34.4	38.2	<u>50.0</u>	29.7	45.8	32.3	42.3	8.3	28.6
Llama-4-Scout-17B-16E-Instruct [7]	53.3	58.1	58.8	40.7	52.9	<u>62.5</u>	52.5	55.3	<u>62.5</u>	37.8	58.3	41.9	48.1	50.0	42.9
Qwen3-VL-8B-Instruct [9]	53.3	51.6	61.8	44.4	50.0	53.1	45.9	<u>55.3</u>	<u>81.2</u>	40.5	66.7	45.2	46.2	33.3	42.9
Qwen3-VL-32B-Instruct [9]	58.7	48.4	79.4	44.4	55.9	56.3	55.7	<u>63.2</u>	<u>75.0</u>	48.7	66.7	67.7	57.7	33.3	42.9
Qwen3-VL-30B-A3B-Instruct [9]	47.8	51.6	58.8	29.6	38.2	<u>53.1</u>	41.0	48.7	<u>62.5</u>	29.7	<u>62.5</u>	45.2	48.1	16.7	28.6
InternVL3.5-8B[10]	50.0	58.1	58.8	29.6	44.1	<u>53.1</u>	49.2	50.0	56.3	37.8	<u>58.3</u>	41.9	38.5	<u>58.3</u>	28.6
InternVL3.5-30B-A3B[10]	41.3	51.6	35.3	37.0	26.5	<u>43.8</u>	42.6	42.1	50.0	40.5	<u>45.8</u>	45.1	46.1	8.3	14.3
InternVL3.5-38B[10]	50.0	48.4	64.7	33.3	41.2	<u>59.4</u>	47.5	54.0	<u>62.5</u>	35.1	<u>62.5</u>	54.8	44.2	16.7	14.3
Phi-4-multimodal-instruct [8]	40.2	41.9	47.1	29.6	29.4	40.6	<u>42.6</u>	40.8	37.5	40.5	<u>45.8</u>	35.5	40.4	25.0	28.6
Random	27.2	25.8	17.7	40.7	23.5	<u>37.5</u>	27.9	29.0	37.5	18.9	29.2	29.0	23.0	16.6	<u>42.8</u>
Human	68.4	75.6	72.6	59.9	75.8	81.4	<u>84.9</u>	75.0	46.8	71.9	82.7	82.1	<u>99.2</u>	83.3	66.2
* Undergraduate	62.8	71.8	70.2	53.8	64.3	66.7	<u>74.9</u>	53.0	43.6	49.7	70.7	75.4	<u>97.4</u>	69.4	46.2
* Graduate & Young Professionals	78.2	85.4	76.6	68.8	73.9	81.9	<u>83.8</u>	76.0	50.0	70.0	81.9	75.0	<u>100.0</u>	83.7	62.5
* Professionals	94.9	90.0	85.0	93.3	89.1	95.6	<u>96.0</u>	<u>96.0</u>	46.7	96.0	95.6	96.0	<u>100.0</u>	96.6	90.0
Gemini-3-pro-preview	77.2	80.6	79.4	70.4	76.5	<u>81.3</u>	72.1	80.3	<u>93.8</u>	75.7	83.3	77.4	78.8	33.3	85.7

B. Author contributions

- **Project Conception:** Y.J., J.F., M.G.

¹This was released on 11/19/2025 at the time of writing supplementary materials.

- **Task Definition & Benchmark Design:** Y.J., J.F., M.G.
- **Data Curation & Management:** Y.J., J.F., M.G.
- **Benchmark Question Generation:** Y.J., J.F., M.G.
- **Model Evaluation:** Y.J., J.F.
- **Qualitative & Quantitative Analysis:** Y.J., J.F.
- **Writing & Visualization:** Y.J., J.F.
- **Supervision:** Y.J., J.F., M.G.

C. Limitations and future work

- **Dataset Scale, Density, and the Expert Bottleneck** Our final dataset consists of 92 high-quality, professionally annotated samples. We acknowledge that this scale is smaller than general-domain VQA benchmarks or DesignQA [3]. However, as highlighted in comparisons (in the main paper), there is a distinct trade-off between dataset size and reasoning depth. High-level cognitive tasks especially toward real-world engineering workflows require annotations from licensed experts, creating an “expert bottleneck” that limits rapid scaling. Crucially, despite the modest total count, DRAWINGVQA is larger and more diverse than the construction-engineering subsets of existing generalist benchmarks. As shown in Figure 4 in the main paper, the relevant subsets of MM-Vet (4 samples), ScienceQA (24 samples), and RBench-M (54 samples) are significantly smaller. Furthermore, where other datasets typically focus on a single question type (e.g., from textbook, exam), DRAWINGVQA spans 7 distinct construction-engineering workflow aspects. While smaller in volume, our dataset offers a higher “information density” and a more comprehensive evaluation of professional versatility than existing alternatives.
- **Closed vs. Open-Ended** To ensure robust and reproducible comparisons, the majority of our quantitative evaluation relies on Multiple-Choice Questions. We recognize, however, that real-world engineering workflows operate in an open-ended setting—practitioners do not select from a list of options but must generate solutions derived from compliance checks to spatial and quantity calculations. While we included an open-ended subset in our benchmark, it remains a challenge. Future iterations of DRAWINGVQA aim to establish a rigorous question set for open-ended engineering VQA.
- **Scope of Engineering Disciplines** DRAWINGVQA currently emphasizes structural engineering and construction management, utilizing real-world IFC-level structural drawings. While structural discipline is the backbone of any construction, the domain is multidisciplinary. Our current scope excludes other disciplines such as Architectural, Civil, and other critical system disciplines such as Mechanical, Electrical, and Plumbing (MEP). While the way to read IFC drawings and the main rationale are similar across disciplines, visual perception and the reasoning process differ. MEP drawings rely heavily on schematic symbols, deciphering complex relations and connections, rather than the physical scale and dimensioning critical to structural drawings. Future work must expand the dataset to cover these modalities. This would allow for a more holistic assessment of an MLLM’s ability to understand construction engineering drawings in a meaningful way.

D. Benchmark details

D.1. Accessing DRAWINGVQA

DRAWINGVQA is an expertly curated benchmark datasets that utilizes Issued for Construction (IFC) level drawings. In this iteration of DRAWINGVQA, the focus was made on structural discipline drawings, as structural drawings include the backbone of any building structures that become the basis for coordination with other disciplines.

The QA sets within the drawings focus on practical questions that would be asked by construction industry professionals to better understand design intent and design requirements. Contrary to many dataset that may require ‘intrinsic’ or ‘latent’ knowledge, DRAWINGVQA focuses on reasoning within the given context window of a drawing or a portion of the drawing, to truly test MLLMs capability in understanding local contexts without hallucinating.

The DRAWINGVQA is shared on <https://huggingface.co/datasets/S2-MIND/DrawingVQA>. It is distributed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) license. Please note that most of the drawings do contain private information such as the architects, engineers, and general contractors. Any information that indicates private information were redacted.

D.2. General Guidelines

There are two types of questions asked within this QA set: Multiple-Choice Questions, and Open-Ended Questions. The multiple choice questions included True/False, and choices of more than 4 options with one correct option. The closed ended questions encompassed short descriptive responses, however, the questions were asked in a way to ensure that long answers

were avoided. Each questions had at least one input of a construction drawing image, where it was either a full drawing image, or a proportion of the full drawing image.

Construction drawings utilized in the industry are typically saved as PDFs for portability. For the purpose of benchmarking, these drawings were converted into PNG images. As part of a process to export from PDF to images, there are various resolutions that can be selected in the unit of pixels per inch (ppi). For the DRAWINGVQA, range of resolution from 25 to 125 were extracted for testing. Practically, the resolutions can defer based on the usage, where for lightweight purpose, the ppi can be as low as 75 ppi, but in other cases such as for printing applications, this can be up until 300 ppi.

The table below shows the average width and height of the images used, as it contains full drawing images to portion of the drawing image, to a very small patch for symbol detection.

D.3. Question curation guidelines

To curate the dataset, 33 Issued for Construction level structural discipline drawings were collected. These drawings became the basis for brainstorming potential questions that can be asked, specifically in the sense that if this was a real project, what sort of questions would professionals ask first when seeing the drawings. These questions and answer pairs were recorded first to ensure that the complexity of the questions were realistic, and matched what was needed for a meaningful benchmark.

Upon completion for the first round of Question-Answer creation, a more scrutinized evaluation was made to come up with plausible options that may arise as a result of misreading or misunderstanding the drawings. These options were added as part of the Question-Answer pair. Finally, these questions were evaluated and designated for its dual categorization to possibly create meaningful insights upon completion of the benchmark testing.

Table 2. Drawing image size variations in pixels

	Minimum	Median	Maximum
Width	54	4064	4800
Height	82	3000	3600

D.4. Dataset VQA Structure

The code block below indicates the dataset schema and structure used to conduct the test in various models this work has benchmarked.

```
{
  "id": Integer,
  "image_name": String,
  "image2_name": String,
  "question": String,
  "options": {
    "A": String,
    "B": String,
    "C": String,
    "D": String
  },
  "answer": String,
  "explanation": String,
  "cv_field": List[String]
  "cv_subfield": List[String]
  "ce_field": List[String]
  "ce_subfield": List[String],
  "topic_difficulty": String,
  "question_type": String
},
```

D.5. Training contamination mitigation

A critical challenge in benchmarking MLLMs is the risk of data contamination, where test samples inadvertently appear in the model’s pre-training corpus. This leakage allows models to solve problems through memorization rather than reasoning.

To guarantee the integrity of our evaluation, DRAWINGVQA is designed to be strictly contamination-free through two key mechanisms:

- **Proprietary and Offline Image Sources:** Unlike benchmarks derived from public internet crawls, the IFC construction engineering drawings in our dataset were sourced from private construction projects and have not been visible to public search engines.
- **Novel Annotation and Temporal Separation:** The Question-Answer (QA) pairs were generated *de novo* by our research team in late 2025, following industry practices. This creation date post-dates the knowledge cutoff of all models tested.

Consequently, we can confirm that DRAWINGVQA represents a true zero-shot evaluation environment. The performance metrics reported in this study reflect the models’ genuine capability to contextualize novel visual engineering context window of a drawing, rather than their ability to recall training data.

D.6. Question Format: Multiple-Choice vs. Open-Ended

To ensure rigorous and reproducible evaluation, DRAWINGVQA employs a Multiple-Choice Question (MCQ) format. The distractors within these MCQs are meticulously expert-designed to reflect realistic cognitive and procedural errors commonly made by practitioners when interpreting IFC (Issued for Construction) drawings. This design ensures the benchmark serves as a reliable proxy for real-world engineering challenges.

To verify that the closed-ended format does not artificially alter the inherent difficulty of the tasks, we conducted an ablation study by converting a subset of 20 MCQs into open-ended questions. An evaluation of MLLM performance across these two formats yielded no statistically significant difference in accuracy (paired t-test, $p = 0.09$). This confirms that the MCQ format in DRAWINGVQA effectively evaluates reasoning capabilities without introducing a format-based advantage or bias.

D.7. Generalization to Unseen Drawing Styles

We prioritize reasoning density over repetitive volume. To further validate the generalization capabilities of our dataset, we conducted an additional evaluation using unseen data. Specifically, we compiled a supplementary set of 49 new VQA pairs derived from 7 newly introduced drawings, distributing 7 QA sets evenly across 7 distinct Construction Engineering (CE) dimensions.

Comparing model performance on this unseen set against the original benchmark, a paired t-test revealed no statistically significant difference in overall MLLM accuracy ($p = 0.07$). This confirms that the performance benchmarked by our dataset remains consistent and generalizes effectively when models are confronted with different drawing styles, due to our drawing QA is careful engineering reasoning required not just extracting information.

D.8. Ethical Consideration

- **Copyright and Licensing:** We maintain strict adherence to all applicable copyright and licensing regulations.
- **Data Privacy and Anonymity:** All project-specific identifiers (e.g., client names, specific site addresses) were redacted prior to inclusion.

E. Dual Categorization Taxonomy

To move beyond aggregate accuracy metrics and diagnose specific model bottlenecks, we developed a dual-category taxonomy. This framework cross-references the *Domain-Specific Dimension* (the engineering intent of the query) with the *Cognitive Capability* (the underlying mechanism required by the MLLM to solve it).

The taxonomy classifies queries into seven domain dimensions, ranging from *General Administrative* tasks to complex *Code and Specification Compliance verification*. As detailed in Table 3, each dimension is associated with the corresponding primary MLLM capabilities necessary for resolution:

- **OCR (Text Recognition/Understanding):** Essential for interpreting dense technical annotations, tables, and callouts (i.e., other cross-references).
- **Visual Perception:** Low-level identification of graphical entities (symbols, lines, shapes), delineation of major spaces within drawings such as sub-drawings or details, and object counting.
- **Reasoning:** Higher-order processing, including *Spatial* (2D-to-2/3D reasoning), *Alignment* (cross-referencing between visual and textual content), and *Visual* (compositional analysis) reasoning. An example can include understanding

spatial composition of elements within the model (e.g. this element A is on the left of element B, element C exists between grid line D and E).

- **Knowledge:** Retrieval of external domain facts or latent knowledge (e.g., standard acronyms, drawing standards, industry best practice, specifications or other engineering standards) not explicitly visible in the pixel data.

Table 3. The DRAWINGVQA Domain-Capability Map. We define seven domain dimensions and identify the primary MLLM cognitive capabilities typically required to solve tasks within each category.

Domain-Specific Dimension	Task Description	Primary Cognitive Dependencies
1. General Admin.	Identifying project participants, sheet metadata, dates, and drawing types.	OCR, Visual Perception
2. Domain Element Identification	Interpreting specific architectural, structural element symbols, callout symbols and other administrative components.	Visual Perception, Reasoning (Alignment, Visual), OCR
3. Discipline Language Understanding	Identifying discipline-specific acronyms, abbreviations, and technical nomenclature.	OCR, Knowledge, Reasoning (Visual, Alignment)
4. Dimensional Understanding	Mapping sections to plan views, locating intersections, and understanding spatial offsets.	Reasoning (Alignment, Spatial), OCR
5. Design Semantic Understanding	Inferring coordination needs, distinguishing design alternatives, and interpreting context.	Knowledge, OCR, Reasoning (Visual, Spatial)
6. Quantity Take-Off (QTO)	Counting elements, measuring linear/area quantities, and extracting categories from schedules.	Reasoning (Alignment), Visual Perception, OCR
7. Compliance & Specs	Matching components to specifications, identifying scope gaps, or verifying code compliance.	Knowledge, Reasoning (Alignment, Visual, Spatial), OCR

F. Experiments details

F.1. Evaluation Prompts and Parsing

To evaluate the model’s performance, we structured the input prompts to encourage Chain-of-Thought (CoT) reasoning and specified a strict output format. Depending on the question type, we appended the following instructions to the input.

Multiple Choice:

The following is a multiple choice question.
Think step by step and then output the answer in the format of “The answer is (X)” at the end.

{{[QUESTION]}}

Options:

{{[CHOICES]}}

Open-Ended:

The following is an open-ended question (with an explicit numeric answer).
Think step by step and then output the answer in the format of `\“The answer is (X)\”` at the end.

{{[QUESTION]}}

Despite these explicit formatting instructions, the models occasionally generated heterogeneous output formats. Common variations included:

- Standard sentences: “The answer is A.” or “The answer is (A)”
- LaTeX formatting: “The final answer is `\boxed{A}`.”
- Markdown emphasis: “The answer is **A**.”
- Minimalist output: “A” or “A: XX”

To robustly extract the predicted label across these variations, we implemented a regular expression pattern designed to capture the answer key while ignoring surrounding formatting or punctuation. Other variations, except the above list, are not handled in the parsing process.

F.2. Model details

Table 5 details the specific versions and API endpoints utilized in our experiments to ensure full reproducibility of our results.

Table 4. MLLM API endpoints and sources.

Model	API Endpoint	Source
GPT-4o	gpt-4o-2024-08-06	OpenAI API
o3	o3-2025-04-16	OpenAI API
Gemini-2.5-pro	gemini-2.5-flash	Google Gemini API
Gemini-2.5-flash	gemini-2.5-pro	Google Gemini API
Claude-4.5-Sonnet	claude-sonnet-4-5-20250929	Claude API
Claude-4.5-Haiku	claude-haiku-4-5-20251001	Claude API
LLaVA-OneVision-1.5-8B-Instruct	lmms-lab/LLaVA-OneVision-1.5-8B-Instruct	HuggingFace, local inference
LLaVA-v1.6-Mistral-7B	llava-hf/llava-v1.6-mistral-7b-hf	HuggingFace, local inference
LLaVA-v1.6-34B	llava-hf/llava-v1.6-34b-hf	HuggingFace, local inference
Llama-3.2-11B-Vision-Instruct	meta-llama/Llama-3.2-11B-Vision-Instruct	HuggingFace, local inference
Llama-4-Scout-17B-16E-Instruct	meta-llama/Llama-4-Scout-17B-16E-Instruct	HuggingFace, local inference
Qwen3-VL-8B-Instruct	Qwen/Qwen3-VL-8B-Instruct	HuggingFace, local inference
Qwen3-VL-32B-Instruct	Qwen/Qwen3-VL-32B-Instruct	HuggingFace, local inference
Qwen3-VL-30B-A3B-Instruct	Qwen/Qwen3-VL-30B-A3B-Instruct	HuggingFace, local inference
InternVL3.5-8B	OpenGVLab/InternVL3.5-8B-HF	HuggingFace, local inference
InternVL3.5-38B	OpenGVLab/InternVL3.5-38B-HF	HuggingFace, local inference
InternVL3.5-30B-A3B	OpenGVLab/InternVL3.5-30B-A3B-HF	HuggingFace, local inference
Phi-4-multimodal-instruct	microsoft/Phi-4-multimodal-instruct	HuggingFace, local inference

F.3. Human baseline

As part of an effort to ensure that the tests being conducted on various SOTA MLLMs are reasonable and not superficial or impractical, a human baselines for the same questions were created to compare their performance against the latest models.

Similar to how various models may have variant parameter sizes or access to different knowledge, a similar delineation was made to the human baseline tests based on their industry experience in years. We recruited 52 participants within the Civil Engineering or Construction Management industry with 3 levels of expertise in the following manner: Undergraduate (Years of experience = 0 years); Young Professionals such as graduate students or those who have just started working in the industry ($0 \leq \text{Years of experience} \leq 2$ years); Professionals ($3 \leq \text{Years of Experience}$). This delineation will also aim to help create human benchmark checkpoints that are comparable with model performance.

As many of the participants were students and working professionals, the questionnaire for the baselines was shortened to 20 questions per person, but the questions were asked in the exact same format as to the MLLMs, with one image + multiple choices for the closed-ended questions, and one image + free response for the open-ended questions. The following rules were used to set up the human benchmark test:

- Questions were provided in a google form.
- Participants did not view any questions prior to completing the form.

- Any internet, LLM or other resource access were prohibited, to preserve testing hygiene, where only the context window of the drawing image, as well as their prior experiences in the industry, are utilized.
- Time limit was set up to be 20 minutes for 20 questions, to ensure all participants had the same control variable, regardless of their experiences.

At the end of the benchmark test, all the participants were asked to rank the difficulty of the QA set, as well as the realistic nature of the QA set to the practical industry application. The results collected by the participants are shown below.

Table 5. DrawingVQA Human Benchmark Test Feedback Results (%).

	1	2	3	4	5
How realistic were the questions?	0.0	0.0	28.3	32.1	39.6
How difficult were the questions?	3.8	17.0	50.9	28.3	0.0

Table 6. Participant votes for the dataset that best requires reasoning skills on real-world AEC industry context and engineering practices.

Dataset	Source	Votes	Percentage (%)
DrawingVQA	Construction drawings from real-world projects	40	76.9
MMMU	FE exam or college exam	7	13.5
CEQuest	Textbooks and guidelines	3	5.8
None	-	2	3.8

F.4. Dense and MoE models

To understand the trade-off between inference efficiency and reasoning accuracy on IFC construction drawings, we compared Dense architectures against Mixture-of-Experts (MoE) variants within the same model families. MoE models are designed to reduce computational cost by activating only a subset of parameters per token.

Table 7 summarizes the performance across the Qwen, InternVL, Llama, and LLaVA families. We observe a distinct trend where *Dense architectures generally outperform their MoE counterparts on the DrawingVQA benchmark.*

- **Qwen Family:** The Qwen3-VL-8B (Dense) achieved **53.3%** accuracy, surpassing the Qwen3-VL-30B-A3B (MoE) which scored **47.8%**, despite the MoE model having nearly 4× the total parameters.
- **InternVL Family:** A similar pattern emerges, where the Dense 8B model (50.0%) significantly outperforms the MoE 30B variant (41.3%).
- **Scaling Behavior:** While the Llama-4-Scout MoE achieved a competitive 53.3%, it required a massive total parameter budget of 109B to match the performance of the significantly smaller Qwen-8B Dense model.

These results suggest that for high-fidelity visual tasks requiring global context interpretation—such as reading complex construction drawings—the sparsity of MoE models may be a limiting factor compared to the dense connectivity of standard Transformers.

F.5. No-Image Ablations

To quantify the extent to which models rely on visual information versus language priors, we conducted a no-image ablation study, a standard diagnostic in VQA evaluations [2, 13]. In this setting, the MLLMs are provided with the textual question alone, without the accompanying visual context. Following [2], we prepended the prompt with the following sentence:

If an image is mentioned, ignore this information and try your best to answer the question.

A rigorous VQA benchmark is “vision-centric”, ensuring that questions cannot be solved solely through common sense reasoning or textual artifacts. DRAWINGVQA was explicitly curated to reflect realistic engineering problems encountered by professionals on job sites, which inherently require visual verification.

As presented in Table 8, the significant performance drop (Δ) observed across high-performing models provides strong empirical evidence that DRAWINGVQA is vision-dependent. Qualitative analysis reveals that when text-only models answer correctly, they often rely on hallucinations (inventing scenarios), bias toward specific answers (e.g., defaulting to “True” on

Table 7. Performance comparison between Dense and Mixture-of-Experts (MoE) models across different parameter scales. † These numbers are from each model’s report and official website.

Model	Arch.	Total Params.†	Activated Params.†	Acc. (%)
<i>Qwen Family</i> [9]				
Qwen3-VL-8B-Instruct	Dense	8B	8B	53.3
Qwen3-VL-32B-Instruct	Dense	32B	32B	58.7
Qwen3-VL-30B-A3B-Instruct	MoE	30B	3B	47.8
<i>InternVL Family</i> [10]				
InternVL3.5-8B	Dense	8B	8B	50.0
InternVL3.5-38B	Dense	38B	38B	50.0
InternVL3.5-30B-A3B	MoE	30B	3B	41.3
<i>Llama Family</i> [6]				
Llama-3.2-11B-Vision-Instruct	Dense	11B	11B	39.1
Llama-4-Scout-17B-16E-Instruct	MoE	109B	17B	53.3
<i>LLaVa Family</i> [5]				
LLaVA-v1.6-Mistral-7B	Dense	7B	7B	28.3
LLaVA-v1.6-34B	Dense	34B	34B	42.4

binary questions), or exploit domain priors (e.g., guessing typical quantities of drawing notes). However, the substantial performance gap (around or even below that random guess) confirms that reasoning on this benchmark necessitates visual grounding.

Notably, refusal rates offer insight into model safety and grounding capabilities. Models such as Gemini-2.5-flash and LLaVA-OneVision exhibited high refusal rates (55.4% and 81.5%, respectively), correctly identifying that the questions were unanswerable without the visual drawing context.

Table 8. Text-Only Ablation Study. Value is a percentage.

Model	Multimodal Acc. (Image+Text)	Text-Only Acc. (Ablation)	Drop (Δ)	Refuse to answer (Require image)
GPT-4o	48.9	31.5	17.4 (↓)	15.2
o3	58.7	29.5	29.2 (↓)	1.1
Gemini-2.5-pro	71.7	37.0	34.7 (↓)	13.0
Gemini-2.5-flash	66.3	9.8	56.5 (↓)	55.4
Claude-4.5-Sonnet	57.6	35.9	21.7 (↓)	2.2
Claude-4.5-Haiku	54.3	18.5	35.8 (↓)	56.5
LLaVA-OneVision-1.5-8B-Instruct [1]	40.2	8.7	31.5 (↓)	81.5
LLaVA-v1.6-Mistral-7B [5]	28.3	25.0	3.3(↓)	2.2
Llama-3.2-11B-Vision-Instruct [6]	39.1	30.4	8.7(↓)	3.3
Qwen3-VL-8B-Instruct [9]	53.3	34.8	18.5 (↓)	9.9
InternVL3.5-30B-A3B[10]	41.3	37.0	4.3 (↓)	3.3
Phi-4-multimodal-instruct [8]	40.2	31.5	8.7 (↓)	17.4
Random (Baseline)	27.2	-	-	-

F.6. Breakdown results on Dual Category Mapping

To identify specific cognitive bottlenecks in construction drawing interpretation, we introduce our dual-category mapping to the error instances of the top-performing model, Gemini-2.5-Pro. Figure 1 illustrates the distribution of failure causes across seven domain-specific tasks. The analysis reveals that failure modes are highly context-dependent. General administrative tasks (“Admin.”) require *OCR* and *Visual Perception* capabilities from MLLMs, indicating the model gets answers by reading explicit information in drawings. In contrast, technical tasks such as “Domain Element Identification” and “Dimensional Understanding (Cross-Referencing)” exhibit a shift toward *Reasoning* (spatial, visual, and alignment), *OCR*, and *Visual Perception* capabilities required. “QTO” (Quantity Take-off) shows sensitivity to *OCR* and *Visual alignment* errors, highlighting the challenge of accurately grounding text within complex graphical geometries. On the other hand, “Compliance” requires knowledge beyond explicit information within the drawings. This heterogeneity indicates that a single optimization strategy is insufficient; distinct construction tasks require targeted improvements in specific multimodal capabilities.

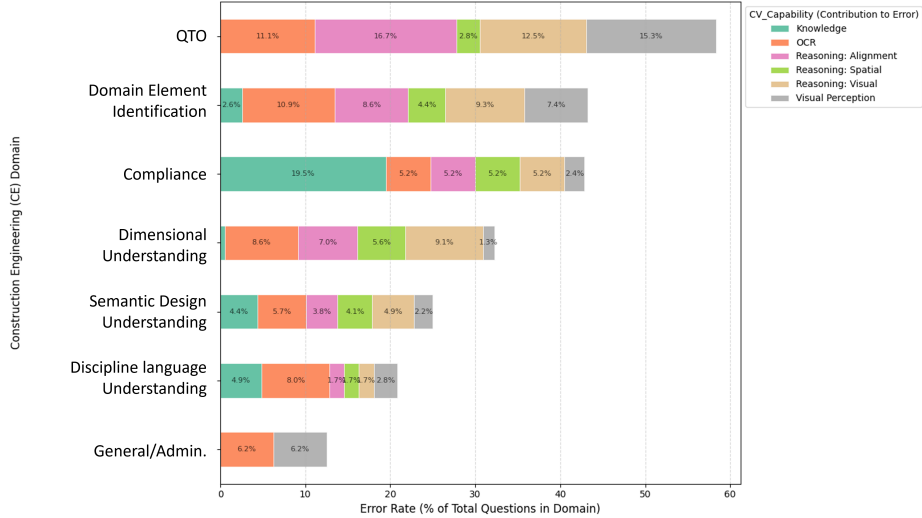


Figure 1. Dual-Category Error Breakdown.

F.7. More experiments on Image resolution

We investigate the impact of image resolution on the performance of MLLMs within the context of engineering drawings. In our dataset construction, we distinguish between Region of Interest (ROI) images and Full Drawing Sheets. While ROI images are typically user-generated crops that do not require standardization, the Full Drawing Sheet serves as the global context and must be rasterized at a resolution that balances legibility with computational efficiency.

To determine the optimal setting, we generated variants of the full drawing sheets at pixel densities ranging from 25 PPI to 125 PPI. Table 9 details the resulting pixel dimensions. Note that the **100 PPI** setting yields a median resolution of 4800×3600 pixels. This resolution substantially exceeds standard 4K UHD resolution (3840×2160), ensuring that fine-grained details—such as dimension text, line weights, and hatch patterns—are preserved without the artifacts common in lower-resolution rasterization.

On top of this, we conducted an ablation study across varying resolutions to assess robustness. As shown in Table 10, model performance generally correlates with increased resolution. However, the accuracy gains plateau beyond 100 PPI, with some models (e.g., Gemini-2.5-flash, Qwen3) showing peak performance at 100 PPI rather than 125 PPI. The 100 PPI setting achieved the highest average accuracy of **55.7%** across all tested MLLMs. Based on these results, we selected 100 PPI as the default resolution for the DrawingVQA benchmark, as it offers the best trade-off between high-fidelity visual detail and model performance.

Table 9. Summary of Image Dimensions for Different PPI Variants on a Full Drawing Sheet.

Images	Width (px)				Height (px)			
	Avg	Min	Median	Max	Avg	Min	Median	Max
125 PPI	5625.1	5077	6000	6000	4178.7	3750	4500	4500
100 PPI	4500.1	4062	4800	4800	3343	3000	3600	3600
75 PPI	3375.0	3046	3600	3600	2507.3	2250	2700	2700
50 PPI	2250.1	2031	2400	2400	1671.5	1500	1800	1800
25 PPI	1125.1	1016	1200	1200	835.8	750	900	900

F.8. More experiments on on Option Permutation and Prefix Sensitivity

Prior research indicates that Large Language Models (LLMs) often exhibit sensitivity to the ordering of choices and the assignment of option labels (e.g., A, B, C, D) in multiple-choice settings, a phenomenon known as position bias or selection bias [12, 14].

To evaluate the robustness of the models on DRAWINGVQA, we introduced an ablation setting named “Prefix v2”. In this setting, we shuffle the order of the answer options and randomize the assignment of alphabetical prefixes. As shown in Table 11, this randomization does not result in a significant difference (Avg. 49.4% and 49.5%).

Table 10. Ablation study on image resolution (Accuracy %). Best results are highlighted in bold.

Model	125 PPI	100 PPI	75 PPI	50 PPI	25 PPI
Gemini-2.5-pro	71.7	71.7	68.5	69.6	64.1
Gemini-2.5-flash	58.7	66.3	65.2	57.6	56.5
o3	60.0	58.7	55.4	56.5	57.6
gpt-4o	52.2	48.9	51.1	52.2	42.4
Claude-4.5-Sonnet	55.4	57.6	56.5	51.1	55.4
Claude-4.5-Haiku	53.2	48.9	50.0	46.7	54.4
Qwen3-VL-8B-Instruct	52.2	53.3	47.8	45.7	46.7
Phi-4-multimodal-instruct	39.1	40.2	37.0	32.6	29.4
<i>Average</i>	55.3	55.7	53.9	51.5	50.8

We attribute this result to the design of the original DRAWINGVQA. The original dataset features carefully curated options and their sequences that are intentionally designed to challenge humans’ reasoning capabilities.

Table 11. Ablation Study on Random Option Prefixes. Values represent accuracy percentages. “Prefix v2” denotes results with shuffled option orders and randomized prefixes.

Model	DRAWINGVQA (As-is)	DRAWINGVQA (Prefix v2)	Δ
GPT-4o	48.9	52.2	+3.3
o3	58.7	59.8	+1.1
Gemini-2.5-pro	71.7	71.7	0.0
Gemini-2.5-flash	66.3	65.2	-1.1
Claude-4.5-Sonnet	57.6	53.3	-4.3
Claude-4.5-Haiku	54.3	57.6	+3.3
LLaVA-OneVision-1.5-8B-Instruct [1]	40.2	31.5	-8.7
LLaVA-v1.6-Mistral-7B [5]	28.3	32.6	+4.3
Llama-3.2-11B-Vision-Instruct [6]	39.1	41.3	+2.2
Qwen3-VL-8B-Instruct [9]	53.3	54.3	+1.0
Qwen3-VL-30B-A3B-Instruct [9]	41.3	48.9	+7.6
InternVL3.5-8B [10]	50.0	44.6	-5.4
InternVL3.5-30B-A3B [10]	41.3	45.7	+4.4
Phi-4-multimodal-instruct [8]	40.2	34.8	-5.4
Average	49.4	49.5	0.1

F.9. Ablation Study on Input Modality: The PDF Hybrid Setting

In standard construction industry workflows, stakeholders typically exchange drawings as vectorized PDFs (e.g., exported from BIM/CAD platforms) rather than rasterized, scanned images. To align our evaluation with this real-world practice, we introduced a PDF Hybrid experimental setting. In this configuration, whenever a question references a full drawing sheet, the model is provided with the native vectorized PDF file instead of a raster image. Conversely, if a question targets a specific Region of Interest (ROI), the input remains a cropped raster screenshot. For questions containing both a full sheet and an ROI, the input includes both the raster crop with the vectorized full-sheet PDF. This hybrid approach mirrors practical scenarios where engineers might upload a full PDF sheet but query a specific visual detail. We limited this study to proprietary MLLMs (GPT, Gemini, Claude series), as most current open-weights models do not natively support PDF document ingestion.

Table 12 presents the comparative performance. Theoretically, vectorized PDFs should offer superior fidelity by providing explicit text layers and sharp vector paths, bypassing the resolution constraints of rasterization. However, our results indicate that this benefit is inconsistent and minor. While GPT-4o and o3 got performance gain (+5.4%), the Gemini and Claude families either stagnated or suffered performance degradation (e.g., Gemini-2.5-flash dropped by 15.2%). This suggests that MLLMs struggle to spatially ground the explicit information contained in vector PDFs.

In queries such as “How many columns are located along vertical grid line Q?”, MLLMs in the PDF setting frequently fail despite the text “Q” and adjacent lines being machine-readable. The models appear to perceive the text and lines as disjointed lists of data rather than a spatially coherent map, failing to align the identifier “Q” with the specific vertical vector line and its intersecting column elements.

Furthermore, the hybrid modality (pairing a vector PDF full-sheet with a raster ROI crop) introduces a *modality gap*. The

models struggle to perform visual referencing when the context (full sheet) is represented in vector space while the detail (crop) exists in pixel space.

This confirms that DRAWINGVQA is a robust, visual-centric benchmark; simply providing machine-readable text via PDFs does not bypass the core requirement for complex spatial and visual reasoning. Future research should focus on bridging this gap, potentially by developing architectures that can better align vector primitives with raster visual features.

Table 12. Ablation Study on PDF Hybrid setting.

Model	DRAWINGVQA (Raster)	DRAWINGVQA (PDF Hybrid)	Δ
GPT-4o	48.9	54.3	+5.4
o3	58.7	64.1	+5.4
Gemini-2.5-pro	71.7	66.3	-5.4
Gemini-2.5-flash	66.3	51.1	-15.2
Claude-4.5-Sonnet	57.6	56.5	-1.1
Claude-4.5-Haiku	54.3	54.3	0.0

G. Error Analysis

We reviewed the Chain-of-Thought (CoT) responses on the best models that were available from our benchmark study, which was Gemini-2.5-pro and Qwen3-VL-8B-Instruct. The following figures illustrates the reasoning the model made to evaluate the options that were given to make the final determination. The figures aim to illustrate where in their Chain-of-Thought went wrong. For all the following figures, light green highlights and comments are added to indicate when models have performed well. The highlight and comments are added in red when model makes an error or false claims.

G.1. Visual Perception errors

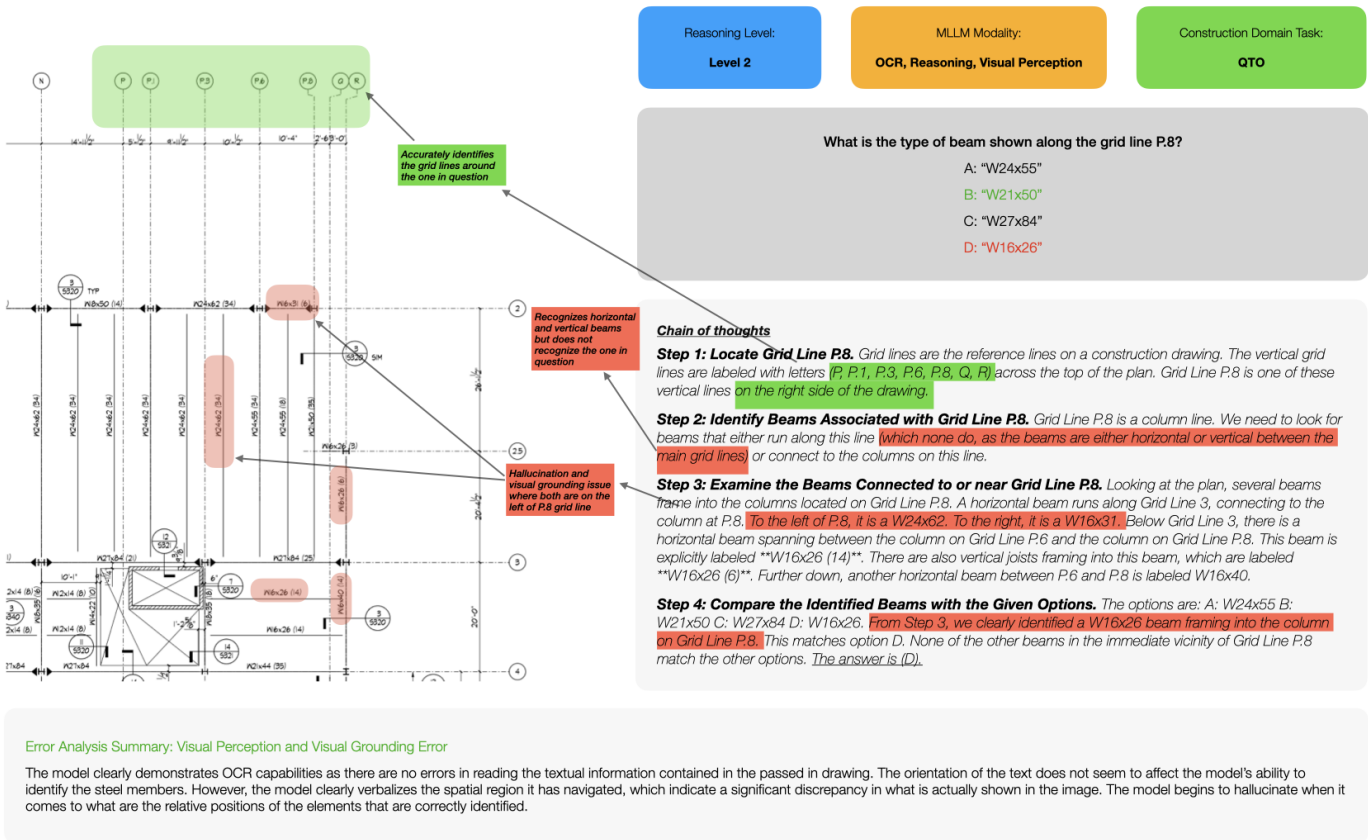


Figure 2. Visual Perception and Visual Grounding Errors for Structural Beam Identification based on Grid Lines (by Gemini-2.5-pro).

G.2. Visual and Text Perception errors

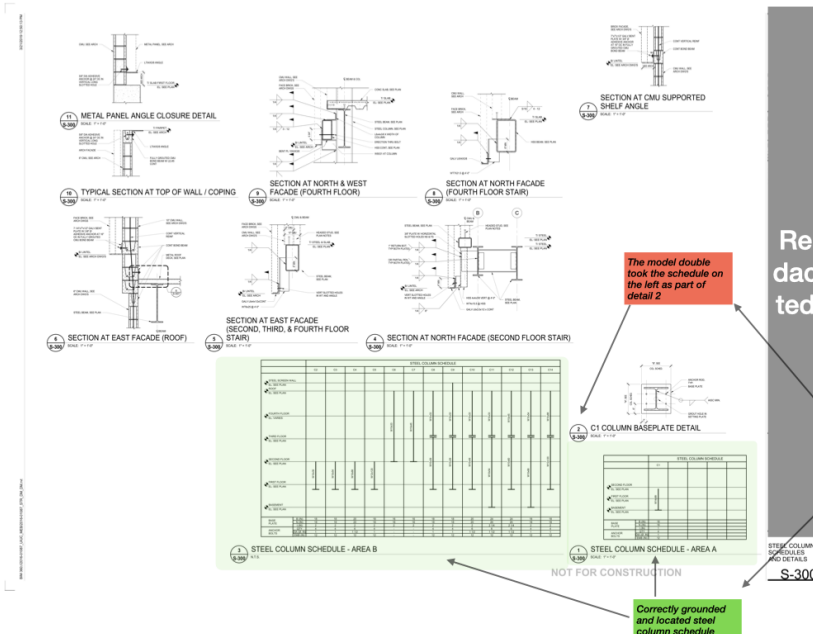
Reasoning Level:
Level 2

MLLM Modality:
Visual Perception, OCR

Construction Domain Task:
General, Language Identification

How many steel column schedules are available in this drawing?

A: "There are none"
B: "2"
C: "3"
D: "10"



Chain of thoughts

Step 1: Examine the architectural drawing and look for any elements titled "STEEL COLUMN SCHEDULE."

Step 2: Look at the bottom portion of the drawing sheet, where several tables and details are located.

Step 3: Identify the first schedule. In the bottom right, there is a detail labeled "1" titled "STEEL COLUMN SCHEDULE - AREA A." This is one schedule.

Step 4: Identify the second schedule. To the left of the first schedule, there is a larger table labeled "3" titled "STEEL COLUMN SCHEDULE - AREA B." This is the second schedule.

Step 5: Look for any additional schedules. Above the "STEEL COLUMN SCHEDULE - AREA A," next to the "C1 COLUMN BASEPLATE DETAIL" diagram (labeled "2"), there is a smaller table with the header "STEEL COLUMN SCHEDULE." Even though it appears within the baseplate detail callout, it is still a distinct column schedule for column C1. This is the third schedule.

Step 6: Count the identified schedules. There is one for Area A, one for Area B, and one for column C1, giving a total of three steel column schedules. Answer: C.

Error Analysis Summary: Visual Perception and Visual Grounding Error

The model is able to identify the actual schedules pretty easily as it is outlined in step 3. The model's usage of positional language like "left of the first schedule" demonstrates their visual grounding capability to a certain extent. However, step 5 indicates that there is still ambiguity in how the model delineates details, as the table which was counted as part of step 4 was basically double counted in step 5. This leads to the model having a false understanding of how many objects, in this case steel column schedule table, are there in this drawing.

Figure 3. A Mix of Visual and Text Perception Errors for Column Schedule Tabular Identifications (by Gemini-2.5-pro).

G.3. Text Perception (OCR) errors

Reasoning Level:
Level 2

MLLM Modality:
OCR, Reasoning, Visual Perception

Construction Domain Task:
QTO

What is the type of beam shown along the grid line P.8?

A: "W24x55"
B: "W21x50"
C: "W27x84"
D: "W16x26"

Incorrectly separates grid line P.8 as 2 separate grid lines

Hallucination and visual grounding issue

Hallucination and text issue

Visual grounding error where it is not looking at grid line P.8

Chain of thoughts

To determine the type of beam shown along grid line P.8, I need to examine the provided architectural drawing.

- The drawing is a "SECOND FLOOR FRAMING PLAN" for "AREA 'B'".
- The grid lines are labeled along the bottom and left side of the plan. I need to locate grid line P and grid line 8.
- Looking at the plan, grid line P is the vertical line running from the bottom to the top on the right side, and grid line 8 is the horizontal line running from left to right near the top of the plan.
- The intersection of grid line P and grid line 8 is at the point labeled "P.8".
- I need to find the beam designation at this intersection. Looking at the beam labels, I can see a beam labeled "P24x55" running along the top of the plan, which is along grid line P.
- The beam labeled "P24x55" is located at the intersection of grid line P and grid line 8.
- The notation "P24x55" indicates a W24x55 beam (W stands for wide flange, 24 is the depth in inches and 55 is the weight per linear foot in pounds).

Therefore, the type of beam shown along grid line P.8 is W24x55. The answer is (A).

Error Analysis Summary: Visual Text Perception, Visual Grounding, and Reasoning Error

While this sort of error was not prevalent, this was observed in some models that had poorer OCR capabilities. This example is when a model is explicitly conflating 2 texts in the same questions, where "W24x55" beam annotation is mistakenly perceived as "P24x55" instead. This may be a mixture of OCR or simply text perception issue when it comes to looking at array of texts in the same image. In addition to this, there is also a simple grounding error where the grid line "P.8" is mistakenly picked up as the grid line for "W24x55".

Figure 4. Text Perception (OCR) Errors on Beam Annotations (by Qwen3-VL-8B-Instruct).

G.4. Knowledge errors

Reasoning Level:
Level 1

MLLM Modality:
Visual Perception, Knowledge

Construction Domain Task:
Element Identification, QTO

How many section detail tags are seen in this drawing?

A: "0"
B: "1"
C: "2"
D: "3"
E: "None of the above"

DECK NOTE:
3" DEEP x18 GA COMP. METAL
DECK + 4 1/2" N.W.T. CONC.
TOTAL THICKNESS 7 1/2".
L5" = L5x5x1/2, (TYP.)
1/7.5" = 2 1/2" BELOW 1/DECK

TYP. FOR "L5"
1/4" x 3-12
WELD TO EMBED. PL.,
SEE 4/S2.25

1 STAIR NO.2 ROOF FRAMING PLAN
SCALE: 1/4" = 1'-0" 1:48
STAIR NO. 2 ROOF 1/SLAB EL. + 801'-4"

PLAN NORTH

Clearly indicates its latent knowledge to potentially understand the query

Some minute discrepancy in understanding where "detail callouts" mean the same as "section detail tag"

No pentagonal tag exists

Hallucinates that 1/S4.01 is a section detail tag but this is rather just a callout for detail, not necessarily a "section" detail.

Chain of thoughts

Step 1: Understand the question. The question asks to count the number of "section detail tags" in the provided architectural drawing.

Step 2: Define "section detail tag." In architectural and engineering drawings, a section tag (or section marker) is a symbol that indicates where a cross-section of the building or an assembly is taken. It typically consists of a line (the cutting plane line) showing where the cut is made, and an arrow or symbol at the end indicating the direction of view for the resulting section drawing. The tag will also have a reference number, often including the detail number and the sheet number where the section drawing can be found.

Step 3: Scan the drawing for these tags. Look for symbols that match the description of a section tag. Near the top of the stair opening, there is a horizontal dashed line with a triangular arrowhead symbol pointing upward. This tag is labeled "1/S4.01." Near the middle of the stair opening, there is another horizontal dashed line with a triangular arrowhead symbol pointing downward. This tag is labeled "2/S4.01."

Step 4: Distinguish section tags from other callouts. There are other symbols on the drawing that refer to details, but these are typically detail callouts, not section tags. Examples include: The circular tag labeled "1/S2.26", The pentagonal tag labeled "3/S2.25", Two pentagonal tag labeled "4/S2.25". These are detail references, not section cuts. The question "section detail tags" is most accurately interpreted as the section cut markers found in Step 3.

Step 5: Count the identified section tags. Based on the scan in Step 3, there are two distinct section tags on this drawing: 1/S4.01 and 2/S4.01.

Step 6: Compare the count with the given options. A: 0; B: 1; C: 2; D: 3; E: None of the above. The count is 2, which corresponds to option C. Answer: C.

Error Analysis Summary: Knowledge and perception error

The model demonstrates a good grasp of what is the composition of a "section tag". However, in the analysis, the model focused on callout annotations as "tags", and made the erroneous case that 4 actual tags that exists in this drawing are not classified as tags, but rather as "callouts". There is also a case to be made that the question may have been vague, however, it does not dismiss the fact that those "callouts" as mentioned in step 4 should be classified as tags in addition to what was discussed in step 3. This is a simple yet technical example that models may have learnt to recognize symbols in a very one sided way, where it may not have the multi-faceted understanding humans may implicitly understand where certain things can be classified or identified in more than one way.

Figure 5. Knowledge and perception error on reading section detail tags (by Gemini-2.5-pro).

G.5. Reasoning errors

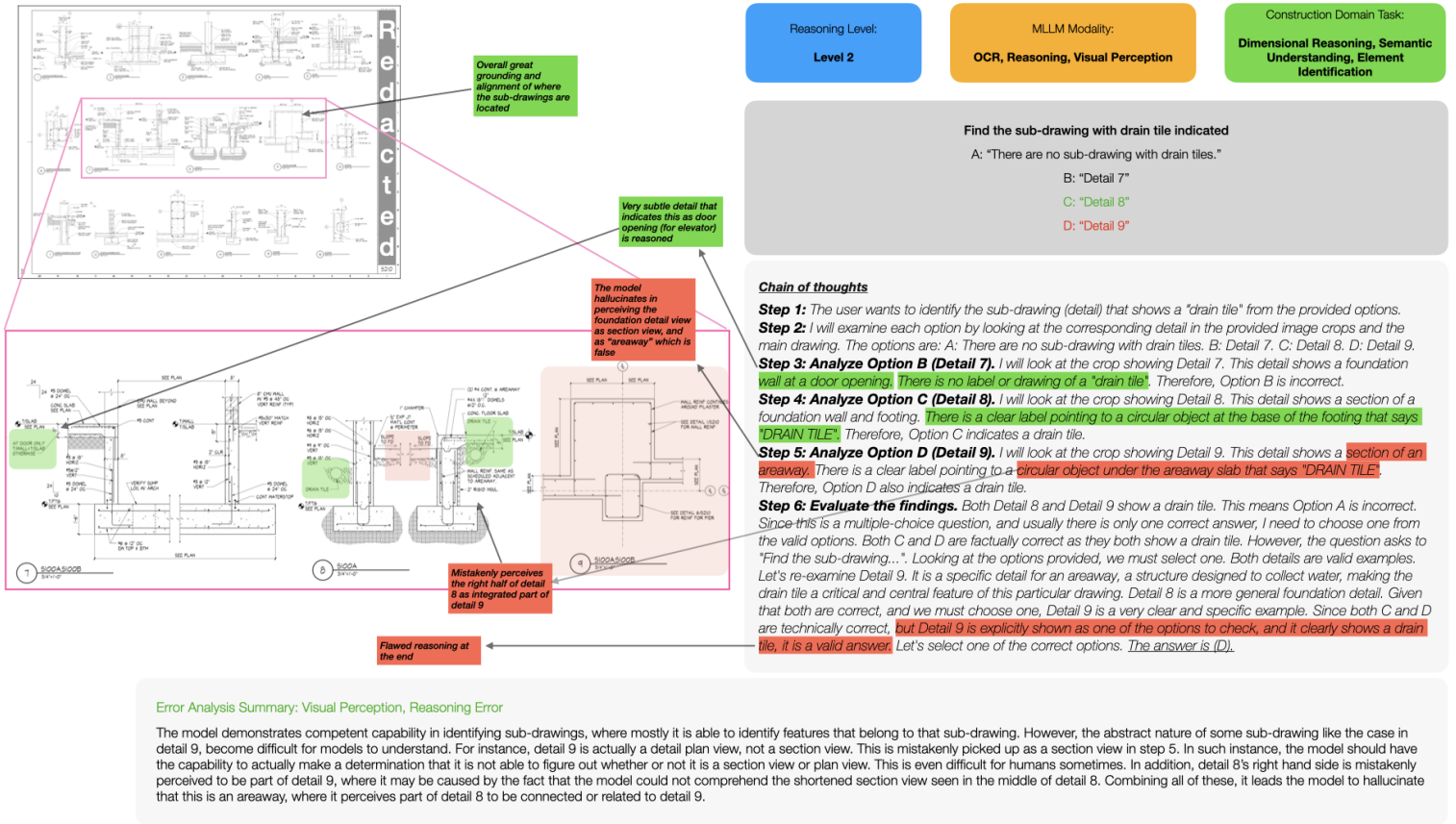
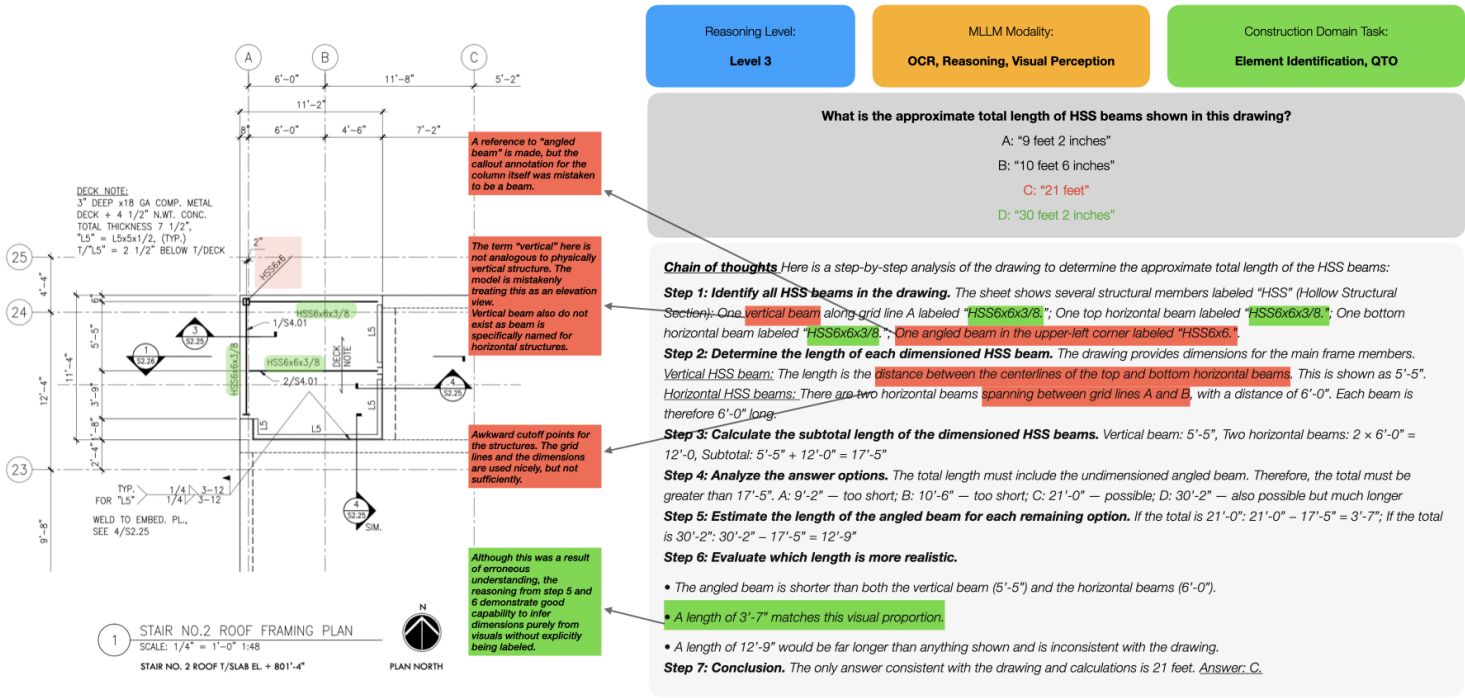


Figure 6. Visual Perception and Reasoning Error on Recognizing Drain Tiles (by Gemini-2.5-pro).



Error Analysis Summary: Knowledge, Reasoning and Visual Perception Error

The model makes a good attempt at deciphering where the beams are, then utilizing the dimension information via grid lines. However, the model may appear to misunderstand that the view seen is a plan (bird's eye view), not elevation view, due to the usage of the word "vertical beam". In addition, a column structure annotation tag's line was mistaken as a "beam" structure, or more specifically, "angled beam". This presents a continuing challenge for models to consistently understanding framing plans, which may be lacking in their latent knowledge. Upon identification of these structures, an attempt was also made to estimate the length, which follows a very compelling process, however, does not complete it in full where it only considers portions of it. This may be due to the fact that the beam structures do not completely touch the end line (i.e. for grid line A, the structure goes 5'-5" and 3'-9" but does not connect to end line). Despite these limitations, the model still tried to make sense of the "angled beam" by proportioning its length to see what is the most feasible, which presents a very sophisticated reasoning process.

Figure 7. Knowledge, Reasoning and Perception Error while calculating length of elements (by Gemini-2.5-pro).

G.6. Instruction Adherence and Output Formatting

We observed a distinct disparity in instruction following between proprietary and open-sourced models. Commercial MLLMs (GPT, Gemini, Claude) demonstrated robust adherence to formatting constraints with no syntax errors. In contrast, a few smaller open-sourced models and certain MoE variants exhibited occasional anomalies:

- **Prefix Omission:** Outputting raw text without the required option label, despite the explicit prompt instruction to use the format “The answer is (A)” (see Section F.1).
- **Language and Refusal Hallucinations:** Reverting to non-target languages (e.g., generating Chinese characters such as 无法确定) or producing generic refusal phrases like “I don’t know” instead of selecting a valid option.

G.7. Parametric Bias vs. Visual Grounding

Beyond formatting adherence, we investigated the models’ ability to prioritize explicit visual evidence over general domain knowledge (parametric priors). MLLMs often exhibit a strong bias toward standard engineering patterns, which can lead to hallucinations when a specific drawing deviates from typical conventions.

To evaluate this, we test an adversarial example where the visual detail contradicts common construction norms. Figure 9 illustrates a representative case involving a slab anchor detail. In standard structural engineering practice, such anchors are typically welded “all-around”. However, in this specific test case, the standard “all-around” circle symbol was intentionally removed from the welding notation. The original example is shown in Figure 8.

Despite the explicit visual absence of this symbol, models selected the “all-around weld” option. This indicates that the models are over-relying on their training priors regarding how slab anchors are *usually* detailed, rather than grounding their reasoning in the objective visual syntax provided in the drawing. This prior knowledge bias remains a significant challenge for automated checking systems, which must detect non-standard or erroneous deviations rather than assuming standard compliance.

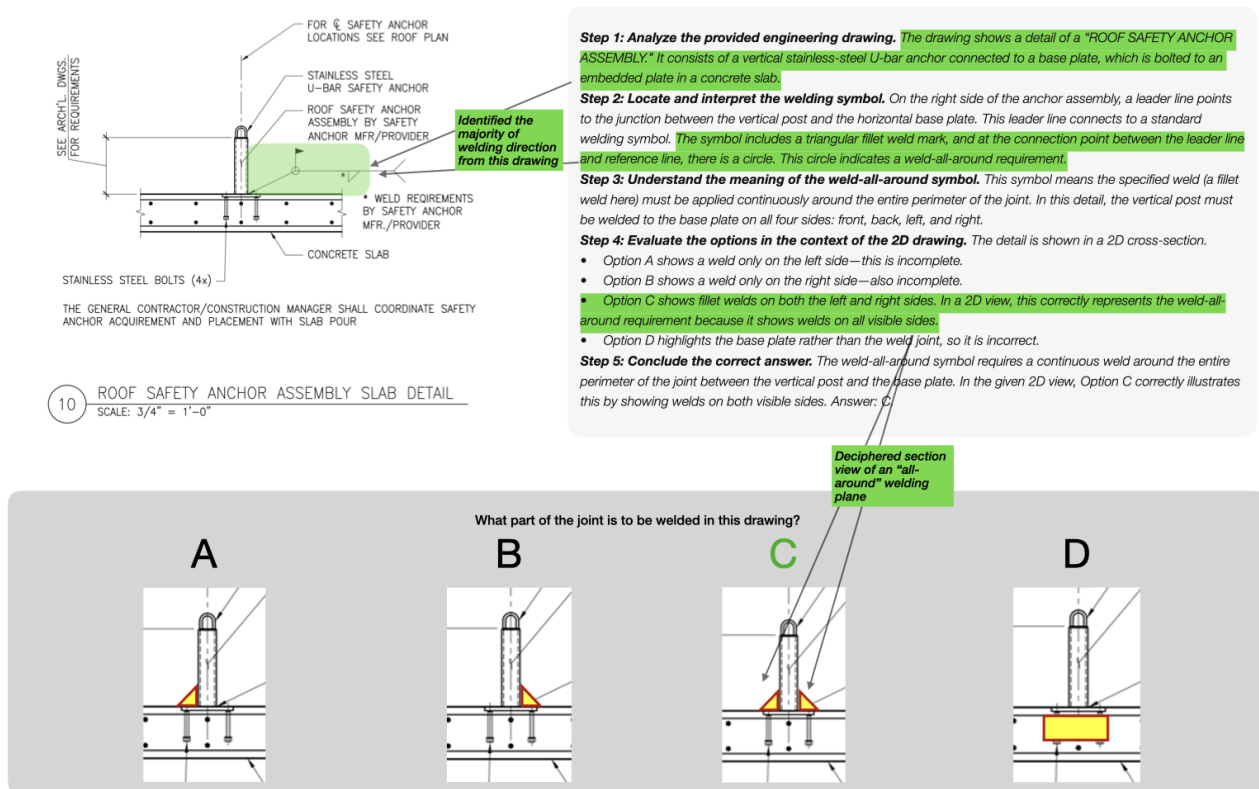
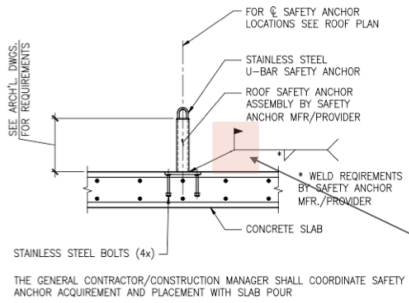


Figure 8. Original example on “all-around” welding example (Gemini-2.5-pro).



Model ignores the direction on the drawing and gets overpowered by its latent knowledge => Parametric bias

Identified the right welding direction from this drawing

Step 1: Analyze the provided image. The drawing is titled "ROOF SAFETY ANCHOR ASSEMBLY SLAB DETAIL" and shows a cross-section of a safety anchor embedded in a concrete slab.

Step 2: Locate the welding instruction. A note on the right side of the drawing reads "WELD REQUIREMENTS BY SAFETY ANCHOR MFR./PROVIDER." This indicates that a weld is required.

Step 3: Identify the weld location. An asterisk next to the note matches an asterisk on the drawing. This asterisk is placed next to a standard welding symbol—a triangle indicating a fillet weld. The leader line points directly to the joint where the vertical stainless steel U-bar anchor meets the horizontal plate of the assembly.

Step 4: Understand the weld symbol. The fillet weld symbol indicates that a weld should be applied at the corner where the vertical U-bar meets the base plate. For safety anchor assemblies, it is standard practice to weld on both sides of the bar unless otherwise noted, ensuring a strong and secure connection.

Step 5: Evaluate the given options.

- Option A: Weld on only the left side of the U-bar. Possible, but incomplete.
- Option B: Weld on only the right side. Also incomplete.
- Option C: Welds on both sides of the U-bar where it meets the plate. This correctly represents the required weld for this type of safety anchor.
- Option D: Highlights the entire base plate rather than the weld joint, which is incorrect.

Step 6: Conclude the correct answer. Option C correctly identifies the part of the joint that must be welded and accurately reflects the intent of the drawing. Answer: C.

10 ROOF SAFETY ANCHOR ASSEMBLY SLAB DETAIL
SCALE: 3/4" = 1'-0"

Hallucinates based on prior statement and forgets to look at the context window

What part of the joint is to be welded in this drawing?

A

B

C

D

Figure 9. Adversarial example to exclude “all-around” circle symbol to highlight parametric knowledge bias in MLLM (Gemini-2.5-pro).

G.8. Visual Attention

To investigate the interpretability of the model’s reasoning process, we extract the attention weights from the final layer of the Qwen3-VL-8B-Instruct decoder [4]. To ensure semantic coherence, we aggregate the attention maps for multi-token words (e.g., merging “P”, “.”, and “8” into a single “P.8” concept) by averaging their respective weights.

The visualization results in Figure 10 reveal a disconnect between the model’s textual output and its visual grounding. While the model attempts to localize general geometric features—showing some attention to linear drawing outlines when processing the token “grid line” or “P.8”—it fails to accurately ground specific beam entities (i.e., attention sink). Crucially, when processing the target location “P.8”, the attention map does not focus on the “P.8” label in the drawing; instead, the focus drifts to irrelevant regions such as grid line symbol F, G, J, P, P.1, G, or R. This visual misalignment indicates that the model is effectively hallucinating the entity’s location (i.e., spatial reasoning) rather than looking at the right pixel area. Consequently, this failure in visual grounding leads to a logic error: the model incorrectly predicts option, having “W24”, whereas the ground truth for the structural member at P.8 is “W21x50”.

We hypothesize that this phenomenon stems from a domain gap: MLLMs are predominantly trained on natural images and may lack the capability to interpret the sparse, symbolic, geometry representations found in 2D technical drawings. These findings bring attention to future work focused on improving cross-modal alignment and grounding specifically for engineering artifacts and 2D schematic views.

✘ Question: What is the type of beam shown along the grid line P.8?

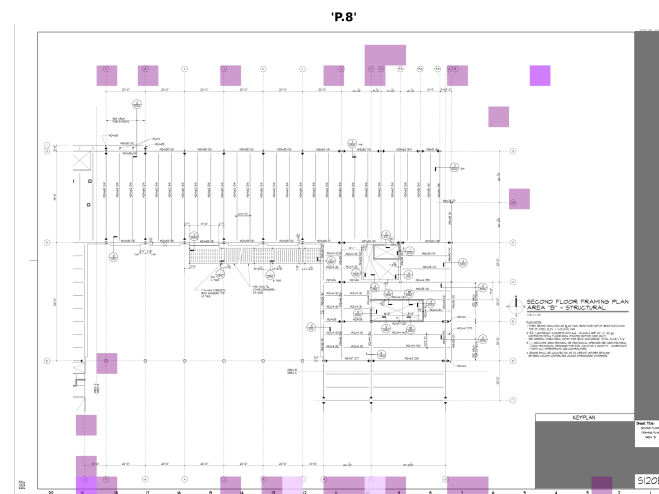
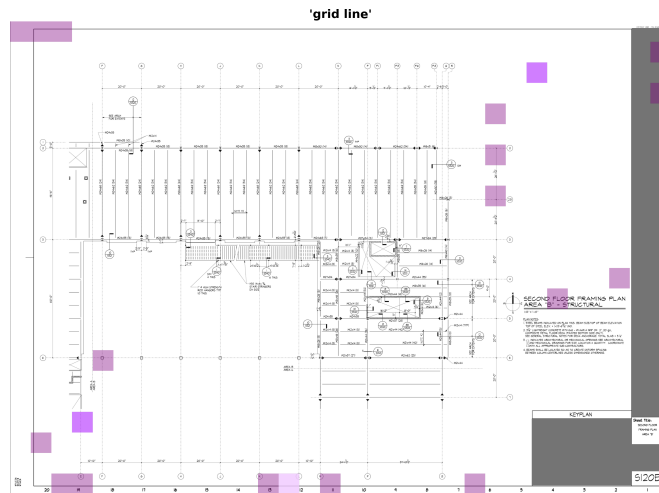
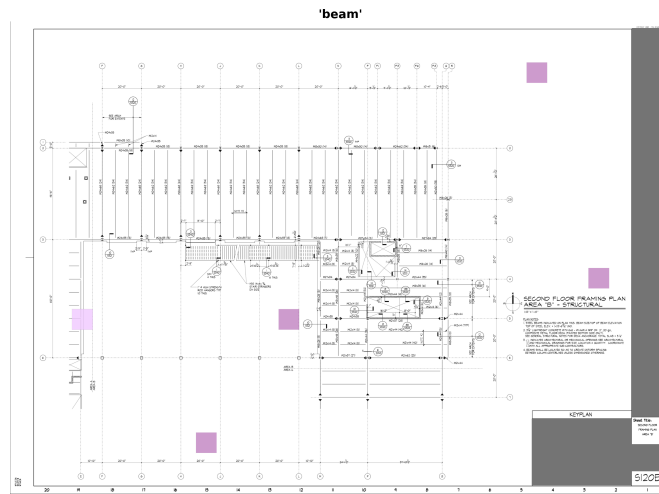


Figure 10. Visual attention maps of Qwen3-VL-8B-Instruct. The maps illustrate where the model “looks” when processing each text token.

References

- [1] Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Chunsheng Wu, Huajie Tan, Chunyuan Li, Jing Yang, Jie Yu, Xiyao Wang, Bin Qin, Yumeng Wang, Zizhen Yan, Ziyong Feng, Ziwei Liu, Bo Li, and Jiankang Deng. Llava-onevision-1.5: Fully open framework for democratized multimodal training, 2025. 2, 9, 11
- [2] James Burgess, Jeffrey J Nirschl, Laura Bravo-Sánchez, Alejandro Lozano, Sanket Rajan Gupte, Jesus G. Galaz-Montoya, Yuhui Zhang, Yuchang Su, Disha Bhowmik, Zachary Coman, Sarina M Hasan, Alexandra Johannesson, William D. Leineweber, Malvika G Nair, Ridhi Yarlagadda, Connor Zuraski, Wah Chiu, Sarah Cohen, Jan N. Hansen, Manuel D Leonetti, Chad Liu, Emma Lundberg, and Serena Yeung-Levy. Microvqa: A multimodal reasoning benchmark for microscopy-based scientific research. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19552–19564, 2025. 8
- [3] Anna C. Doris, Daniele Grandi, Ryan Tomich, Md Ferdous Alam, Mohammadmehdi Ataei, Hyunmin Cheong, and Faez Ahmed. Designqa: A multimodal benchmark for evaluating large language models’ understanding of engineering documentation, 2024. 3
- [4] Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. See what you are told: Visual attention sink in large multimodal models. *arXiv preprint arXiv:2503.03321*, 2025. 21
- [5] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 2, 9, 11
- [6] Meta. Llama-3.2-11b-vision-instruct. Technical report, 2024. 2, 9, 11
- [7] Meta AI. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. <https://ai.meta.com/blog/llama-4-multimodal-intelligence/>, 2025. Accessed: 2025-11-18. 2
- [8] Microsoft. :, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yiling Chen, Qi Dai, Xiyang Dai, Ruchao Fan, Mei Gao, Min Gao, Amit Garg, Abhishek Goswami, Junheng Hao, Amr Hendy, Yuxuan Hu, Xin Jin, Mahmoud Khademi, Dongwoo Kim, Young Jin Kim, Gina Lee, Jinyu Li, Yunsheng Li, Chen Liang, Xihui Lin, Zeqi Lin, Mengchen Liu, Yang Liu, Gilsinia Lopez, Chong Luo, Piyush Madan, Vadim Mazalov, Arindam Mitra, Ali Mousavi, Anh Nguyen, Jing Pan, Daniel Perez-Becker, Jacob Platin, Thomas Portet, Kai Qiu, Bo Ren, Liliang Ren, Sambuddha Roy, Ning Shang, Yelong Shen, Saksham Singhal, Subhojit Som, Xia Song, Tetyana Sych, Praneetha Vaddamanu, Shuohang Wang, Yiming Wang, Zhenghao Wang, Haibin Wu, Haoran Xu, Weijian Xu, Yifan Yang, Ziyi Yang, Donghan Yu, Ishmam Zabir, Jianwen Zhang, Li Lyna Zhang, Yunan Zhang, and Xiren Zhou. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras, 2025. 2, 9, 11
- [9] Qwen Team. Qwen3 technical report, 2025. 2, 9, 11
- [10] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 2, 9, 11
- [11] Ruoxin Xiong, Yanyu Wang, Suat Gunhan, Yimin Zhu, and Charles Berryman. Can ai master construction management (cm)? benchmarking state-of-the-art large language models on cm certification exams, 2025. 2
- [12] Zhen Yang, Ping Jian, and Chengzhi Li. Option symbol matters: Investigating and mitigating multiple-choice option symbol bias of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1902–1917, 2025. 10
- [13] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhui Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9556–9567, 2024. 8
- [14] Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors, 2024. 10