ACCSCIENCE PUBLISHING

International Journal of AI for Materials and Design

ORIGINAL RESEARCH ARTICLE

Evaluating multimodal foundation models for engineering drawing assessment and spatial reasoning

Javier Munguia[1]*, Lim Jen Nee Jones[2], Yudhi Ariadi[1], and Jing Fu[3]

[1]School of Engineering, University of Warwick, Coventry, United Kingdom

[2]School of Engineering, Monash University Malaysia, Petaling Jaya, Selangor, Malaysia

[3]Department of Mechanical and Aerospace Engineering, Monash University, Clayton, Australia

Abstract

This work investigates two different aspects of contemporary multimodal vision language models (VLMs) regarding their analysis and understanding of engineering technical drawings: (i) their capacity to grade drawings based on their text, image and graphic analysis, and (ii) their engineering spatial reasoning capabilities. Using a collection of one hundred authentic engineering drawing assessments and a novel drawing-to-isometric correspondence task, we examined how two widespread commercially available VLM models, namely Gemini (Google DeepMind) and Copilot (Microsoft), perform across multiple repeats of the same task. Both experiments were benchmarked to human experts who fulfilled the same task as a baseline comparison. Results from the drawing assessment and marking experiment showed a notable capacity of both VLMs to grade submissions, both within 10% error relative to human marking results. The intra-marking consistency and the ranking ability varied across models, with Copilot displaying higher reliability levels closer to human level. The isometric-matching task, however, demonstrated that current VLMs struggle with highly visual abstraction tasks, as their model-matching capacity was well below human marker's performance. We conclude that, although VLMs exhibit high performance in analytical tasks involving optical recognition, text and data mining, they appear to still lack sufficient abstraction capacity consistent with human mental processes such as mental rotation, graphic memory retention and shape matching. A cross-analysis further revealed that spatial reasoning performance and grading accuracy are moderately associated for Gemini, but independent for Copilot, suggesting fundamentally different underlying processing strategies between the two models.

Keywords: Engineering 2D drawings; Artificial intelligence; Spatial reasoning

***Corresponding author:**
Javier Munguia
(Javier.munguia@warwick.ac.uk)

Citation: Munguia J, Jones LJN, Ariadi Y, Fu J. Evaluating multimodal foundation models for engineering drawing assessment and spatial reasoning. *Int J AI Mater Design*. 2026;3(3):026340030. doi: 10.36922/IJAMD026340030

Received: August 17, 2026

Revised: September 5, 2026

Accepted: September 11, 2026

Published online: September 23, 2026

Copyright: © 2026 Author(s). This is an Open-Access article distributed under the terms of the Creative Commons Attribution License, permitting distribution, and reproduction in any medium, provided the original work is properly cited.

Publisher's Note: AccScience Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

Engineering drafting is a multifaceted activity; on one hand, it requires expertise and command of common engineering communication cues such as tolerances, dimensioning, drawing conventions, projections, as well as the detailing of sufficient

information to make them readable and valid for production.[1,2] On the other hand, engineering drawings also reflect the engineer's spatial reasoning and overall design intent.[3] Because of this, two-dimensional (2D) drawing drafting remains one of the core activities actively assessed at the higher education level, as they have the ability to summarize the student's understanding not only of the formal design process itself but across wider domains such as materials selection, manufacturing ability, standards and documentation.

Technical drawing documentation is usually part of larger design portfolios which students usually submit as part of individual or group-based projects, generating dozens of individual and assembly drawings per project, making it one of the most time-consuming tasks in engineering design education. For this reason, there is an increased interest in exploring AI's ability to support the marking/grading process either as a 'blind second marker' support element or as a fully automated marking agent.[4,5]

Recently, artificial intelligence (AI) has demonstrated real capacity to scrutinize highly detailed technical 2D drawings, with various AI agents currently operating commercially[6–9] providing a detailed 2D technical drawing analysis, review, and feedback service. Some sources consider the depth of analysis and speed for generating highly technical feedback to be on par with expert technical reviews, claiming a degree of human-AI equivalency;[5] however, there are no conclusive scientific studies supporting these particular claims.

Although AI's principal advantage relies in the combined and simultaneous operation of machine-learning, text-recognition, character-detection and image analysis[10], modern agentic AI has been reported to underperform whenever its spatial reasoning ability is tested in highly demanding scenarios, for instance, when dealing with multiple files simultaneously.[11,12] AI is known to struggle in contexts where profound design-intent, perception and holistic engineering judgement are required.[13] However, the speed and scale capacity at which AI can process and provide feedback based on purely geometric, text and image data, still make AI-assisted review an attractive proposition for universities and colleges.

Spatial reasoning, in the context of engineering design, is understood as the cognitive ability to manipulate and transform relationships between objects and their specific representations.[14,15] In an engineering drawing and drafting context, spatial reasoning is the ability to understand the engineer's design intent when assessing a particular drawing or three-dimensional shape.[16,17] Current agentic-generative AI engines have quickly transitioned from purely text-based processes (large language models – LLMs) into vision learning models (VLMs) with the capacity to understand and answer highly specialized questions about technical drawings based on graphical inputs.[18,19]

In this work, we seek to answer a different question: Do AI models genuinely reconstruct 3D geometry from their 2D orthographic representation, or are they merely recognizing visual patterns? The capacity of VLMs/LLMs to 'understand' three-dimensional reasoning has not been fully understood or established; while some studies point that current AI's visualization capabilities can support or augment cognitive demands associated with spatial analysis tasks[20], there are no studies that directly compare human spatial reasoning traits with those of AI systems.

Given the growing interest in AI-assisted marking and drawing assessment in engineering education, this work evaluates whether current widely available VLM platforms possess sufficient spatial reasoning ability which is at a par with their other technical capabilities such as data mining, image or character recognition. Unlike existing studies that focus on feature extraction or qualitative drawing comprehension, we evaluate VLMs as formal grading agents using a validated academic rubric, while introducing a novel 2D-to-3D isometric correspondence benchmark designed to isolate spatial reasoning from text recognition.

1.1. Work hypothesis and research questions

Can contemporary VLMs understand and reconstruct sufficient three-dimensional representations from orthographic engineering drawings to correctly establish the correspondence with their isometric representation, and is their spatial understanding capability influencing their 2D assessment?

In this work, we intend to answer the following:

RQ 1: Establish whether VLMs are capable of consistently understanding and grade complex engineering drawings.

RQ 2: Establish whether AI can interpret and conceptualize orthographic drawings and match them to standard isometric views.

RQ 3: establish whether AI's marking ability is supported or limited by their spatial ability.

The following section reviews prior research on AI and agentic-generative AI as engineering drawing assessor, with an emphasis on grading performance and spatial reasoning.

2. Literature review

The AI-based review and analysis of engineering drawings have been well documented; Picard *et al.*[21]

evaluated contemporary agentic AI for engineering drawing understanding, sketch similarity and engineering education concluding that VLMs are capable of analyzing engineering visual artifacts beyond natural images. They did, however, highlight the high prompt-dependency of their responses. They also named 'spatial reasoning' as a potential research avenue to evaluate future multi-modal VLMs. Khan *et al.*[22] focused on Geometric Dimensional and Tolerancing (GD&T) for feature extraction as a benchmark to assess LMV performance over 400 samples, concluding that, while AI excelled in the automatic extraction for engineering interpretation, a degree of domain-specific fine-tuning was necessary for higher precision.

Zhou *et al.*[23] developed a vision-augmented framework for automatically comparing engineering drawings, mostly to compare performance with a predefined benchmark; the team identified 'geometric reasoning' as one of the core LLM limitations, alongside precise feature localization as well as interpretation of complex semantics.

Edwards *et al.*[5] coined the term 'AI as judge in design', highlighting the potential of modern VLMs in assessing not only generic features but in understanding design intent and spatial recognition. This study pioneered in investigating AI in relation to human performance on attributes such as creativity, uniqueness and drawing quality. However, their core claim of AI surpassing the ability of human experts on sketch analysis has not been confirmed in literature.

The agentic-AI analysis capability landscape is rapidly changing. Doris *et al.*[24] explored state-of-the-art multimodal large language models (MLLMs) in 2024, concluding that across the different agents being studied, there was a widespread lack of understanding of graphical technical documentation. Many of these hurdles are no longer true for contemporary agentic AI models; therefore, it is reasonable to speculate on their potential ability for spatial recognition.

3. Methods

A collection of 100 two-dimensional (2D) Mechanical Engineering drawings has been used to assess AI's capacity for grading each individual drawing and the ability to interpret tri-dimensional isometric images. The drawings are part of Warwick University's Advanced Mechanical Engineering Design course (ES3C2)[25], in the context of a racecar design group assignment. The dataset comprises 100 third-year undergraduate Mechanical Engineering submissions assessed under a formal 0%–100% marking rubric (**Table 1**). Drawings were submitted as part of a design portfolio and archived in PDF format, preserving native vector geometry, line types, GD&T annotations and title block text without pixelation. Selection criteria required each drawing to be a complete, independently assessable submission. The complete dataset, including 2D drawings and corresponding 3D isometric images, is publicly available as an open-source database in the Zenodo repository.[26] Each drawing is different and depicts a specific mechanical component belonging to an automotive suspension, containing different versions of race-car uprights, hubs, Constant-Velocity-Joints (CVJ), brackets, wishbones and caps. All experiments were conducted using independent prompts. File metadata, component identifiers and other semantically revealing textual information were removed to minimize non-geometric cues and encourage reasoning over graphical content. Each drawing has been anonymized, with all sensible identifiers removed and metadata disabled via Kofax Power PDF (Tungsten Automation, U.S.). Additionally, a separate collection of one hundred 3D isometric images has been compiled, each one corresponding to one orthographic 2D drawing. The specific preparation of each dataset is detailed in the relevant experimental setup below.

Two consumer-type agentic VLMs have been chosen for this study, as they both represent the current Institutional recommendation by the two universities participating in this study: Microsoft Copilot GPT-4o-based multimodal model (University of Warwick, UK) and Gemini 3.1 Deep Think mode by Google DeepMind (Monash University, Australia). Both VLMs were accessed exclusively via their consumer web interfaces (not API): Microsoft Copilot (GPT-4o-based, web interface, accessed May–August 2026) and Gemini 3.1 Deep Think mode by Google DeepMind (web interface, accessed May–August 2026). All experiments were conducted under default interface settings. In contrast to other studies where contextual information and prior sample results are provided to VLMs to shape their outputs via a Short-Shot strategy, all prompts have been supplied without preliminary background or domain-specific information. This study is aimed at characterizing VLMs' grading capacity and graphical information analysis in the space without providing sample benchmarks.

3.1. Experiment 1: 2D engineering drawing assessment

The PDF file format was adopted to preserve geometry, dimensions, annotations, GD&T symbols, as well as line-types and text, while avoiding the pixelation or distortion that image files can exhibit. Native 2D drawing formats such as DXF/DWG are currently not supported by most VLMs, meaning they lack the parsing ability of dedicated computer-aided design (CAD) software. The drawing assessment analysis was designed to determine whether

VLMs can assess, understand and grade engineering drawings using the same rubric used by human markers (Table 1).

3.1.1. Prompt design and processing capacity

Several initial trials were attempted for batch-processing multiple 2D drawing files simultaneously on both VLMs; initially, one single PDF file with the complete 100 drawing dataset was attached to the prompt; however, it became rapidly evident that all VLMs processing capacity decreased and automatically ceased to produce meaningful results after 40 drawings; therefore the number of drawings included within one single multi-page PDF file was gradually decreased until it was determined that thirty three drawings provided a good balance for multi-file batch-processing and VLM analysis speed.

Three different independent marking instances were run on each VLM, each one using a new independent session with no access to previous outputs to prevent previous responses from influencing each consecutive analysis, thus avoiding conversational 'carry-over' while creating a 'fresh inference context'.

3.2. Human grading and VLM agreement analysis

3.2.1. Human markers and Spearman's rank correlation

For each drawing, two historic marks from independent graders were available from the original dataset: the course leader/assignment setter (Human Marker 1, >15 years of engineering drawing assessment experience) and a second independent marker (Human Marker 2), both familiar with the BS 8888-aligned marking rubric through established departmental marking practice. Marks were produced independently with no cross-reference between markers, consistent with standard double-blind academic marking procedures. A third senior academic served as an assignment moderator for the portfolio assignment. To establish the human-human agreement, the Spearman correlation (ρ) was calculated between both datasets ρ(H1,H2) where H1: primary first marker and H2: second marker, with Spearman's correlation defined as **Equation 1**.

$$\rho = 1 - \frac{6\sum d_i^2}{n\left(n^2 - 1\right)} \tag{1}$$

where:

ρ = Spearman rank correlation coefficient,

d_i = observation differences (i) by both reviewers,

n = sample number.

Table 1. Assignment marking rubric for 2D technical drawings

Weighting	Marking rubric criteria
10%	1. Uses the correct template, clean, tidy, uncluttered.
15%	2. Information box, with the correct projection symbol in relation to the projection used. A different person must sign the designer, revision and approval.
10%	3. Reasonable scale. Not predetermined, but ideally a good size using the whole template space.
15%	4. Information text: including material (specifying alloy, condition, *etc.*), manufacturing process, finishing and post-processing.
5%	5. Functional surface and tolerancing information.
15%	6. Sufficient dimensions for key features without duplicates.
15%	7. Center lines, center marks, hidden lines.
5%	8. Datums and cross sections are clearly specified.
5%	9. Isometric 3D view included
5%	10. Basic compliance with BS 8888

3.2.2. Intraclass correlation coefficient

The intraclass correlation coefficient (ICC) was introduced to capture any potential differences between markers as well as the variance within each dataset to establish the agreement between the two. The two-way random-effects model was selected specifically to account for the paired and repeated structure of the data, partitioning variance attributable to individual drawings from variance attributable to assessor disagreement. This is particularly relevant as it considers the variance attributed to differences between assessed 2D drawings, as opposed to marker disagreements (**Equation 2**).

$$ICC = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_W^2} \tag{2}$$

where:

σ_B^2 = variance across 2D drawings,

σ_W^2 = variance due to error/disagreement amongst assessors.

As the variances across datasets are initially unknown, they are estimated from a two-way random-effects ANOVA. Specifically for this study, the two-way random-effects

model with absolute agreement for single measurements, denoted ICC(2,1) was employed. This model assumes that both the assessed drawings and the assessors represent random samples from larger populations and therefore allows the findings to be generalized beyond the specific assessors included in the study (**Equation 3**).

$$ICC(2,1) = \frac{MS_B - MS_E}{MS_B + (k-1)MS_E + \frac{k(MS_R - MS_E)}{n}} \quad (3)$$

where:

MS_B = the mean square between drawings,

MS_R = the mean square between assessors,

MS_E = the residual (error) mean square,

n = the number of assessed drawings (100),

k = the number of 'human' assessors (2).

The absolute agreement amongst both markers, ICC(2,k), was additionally reported to estimate the reliability of the average mark obtained from two assessors, with variables defined as above (**Equation 3**). The same procedure was repeated for each VLM to assess their repeatability and reliability of results after three separate independent runs were executed, thus comparing three different marking datasets (**Equation 4**).

$$ICC(2,k) = \frac{MS_B - MS_E}{MS_B + \frac{MS_J - MS_E}{k}} \quad (4)$$

3.2.3. Intra-model variability

To quantify the consistency of independent VLM's assessments, each drawing was evaluated multiple times under identical conditions. Intra-model variability was quantified using the mean absolute deviation (MAD), as per **Equation 5**.

$$MAD = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - \bar{x}\right| \quad (5)$$

where:

x_i = score obtained from each independent assessment,

n = number of evaluations,

$\bar{x}$ = mean predicted score.

Lower MAD values indicate greater assessment of repeatability. The maximum score deviation was additionally recorded as the assessment range (**Equation 6**), providing an indication of the largest variation between independent model responses. All statistical analysis calculations were performed in MATLAB R2025b (MathWorks).

$$R = x_{max} - x_{min} \quad (6)$$

3.3. Experiment 2: Spatial reasoning analysis

This test was conducted to evaluate the VLMs' capacity to inspect orthographic 2D drawings and match them with their corresponding 3D-isometric figures (Figure 1). All experiments were conducted using independent prompts with three different runs, executed on independent sessions to prevent conversational influence from historic chats. The metadata, component identifiers and other semantically revealing textual information were removed from each 2D drawing to minimize non-geometric cues and encourage reasoning over graphical content. The main dataset modification for this experiment was the removal of any 3D isometric projection in the drawing, to prevent VLM agents from exploiting their image-based recognition rather than promoting shape and geometric inference. A collection of one-hundred isometric tri-dimensional images was generated in PNG image format, which included one single view of each component, alongside an identifier code to randomize them.

3.3.1. Part A. Human experts' spatial reasoning analysis

The spatial analysis test was conducted by three human experts to serve as the baseline for comparison. For both datasets, 2D drawings and isometric projections were provided to three academic experts who conducted the 2D-3D matchmaking independently. All experts were asked to nominate their top 3 matches, with the ability to name the same image multiple times to prevent error accumulation on subsequent tasks. All three academics reported their rankings on an Excel file which was then analyzed for consistency. The three complete datasets are available on an open-source research repository for research reproducibility.[26]

3.3.2. Part B. VLMS spatial reasoning analysis

Both VLMS undertook the 2D-drawing/3D-image sorting task via prompts, which were used in independent chat sessions; this involved logging off, clearing cache, browsing history and conversation chat history to prevent VLMs from exploiting their ability to retrieve solutions from previous conversation history.

Several methods were explored to feed the 3D-ISO images to both VLMs, including individual PDFs, multi-page PDFs, and individual images in JPG and PNG formats. During this series of tests, it became evident both VLMs memory clogged after multiple-pages or multiple files were uploaded; therefore, a thumbnail clustering activity was adopted which provided VLMs with a sufficiently

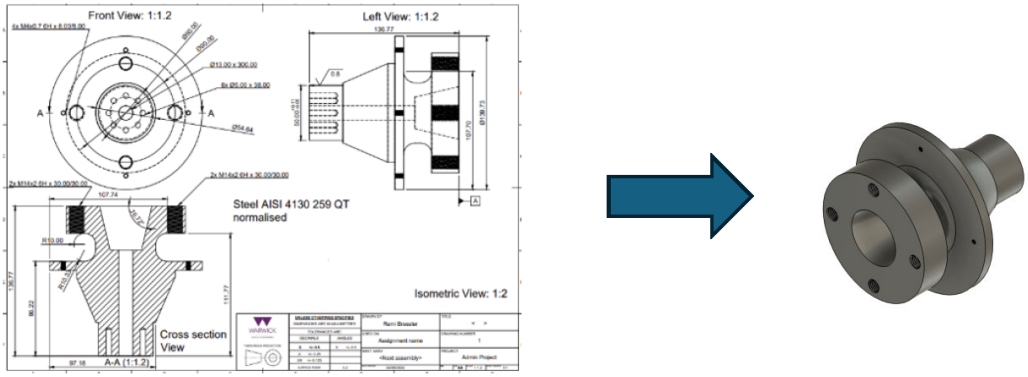


Figure 1. A sample 2D drawing with the isometric projection removed (left) and its corresponding 3D isometric image for matchmaking (right)

clear isometric image of each independent component along with their 4-digit ID (Figure 2). This strategy served as "memory cards" which were easily processed by both VLMs and efficiently reduced the probability of processing memory clogs.

Three rounds of component identification were performed on each VLM, matching all 2D-drawings to their 3D-isometric counterparts. As opposed to 'human experts', who were asked to provide only their single best guess (Top-1 response), VLMs were asked to provide their Top-1, Top-2, and Top-3 responses. Performance was recorded both as the proportion of top-1 responses that were correct and as the mean reciprocal rank (MRR), which incorporates the second- and third-ranked responses into the final score using a rank-specific correction multiplier, as defined in **Equation 7**. The same prompt structure and sequence was used across both VLMs, and the results were exported as CSV data for comparison (Figure 3, Appendix).

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i} \quad \text{(VII)}$$

where:

- $|Q|$ = total number of instances,
- $Rank_i$ = the position of (1-indexed) of the first relevant item for the i-th query,
- Top-1 correct option: Rank 1 = 1.0 (1/1),
- Top-2 correct option = Rank 2= 0.5 (1/2),
- Top-3 correct option = Rank 3 = 0.33 (1/3),
- Not in top 3 choices = Rank null = 0.

A supplementary cross-analysis was conducted to investigate whether VLMs' spatial reasoning performance is associated with their per-drawing grading accuracy relative to human markers. Two variables were defined for each drawing: (i) the mean absolute grading error averaged across three independent runs, calculated relative to the average of both human markers; and (ii) a binary spatial reasoning outcome under two conditions: Top-1 and Top-3, determined across three independent runs. The Mann–Whitney U test was used to compare grading error distributions between correctly and incorrectly matched drawings.

4. Results

4.1. Expert grading agreement analysis

4.1.1. Spearman's rank correlation

The initial Spearman rank correlation for both human markers indicates a strong positive association ($\rho = 0.867$, $p < 0.001$), indicating that both human markers agree on the marking level assigned amongst the whole dataset (Figure 4).

4.1.2. Intraclass correlation coefficient

With agreement amongst ranks established, we applied an ICC analysis to investigate the consistency of both markers on each individual drawing. The ICC(2,1) yielded a coefficient of 0.826, demonstrating a good agreement between markers. The average reliability was ICC(2,k) = 0.905, indicating high reliability when average marks from both markers are considered.

4.2. VLMS grading agreement analysis

The results from three independent marking runs were assessed for each VLM: Gemini and Copilot. As a first step, an intra-agreement analysis was performed for consistency, followed by a VLM/Human grader comparison. The concentration of marks is shown in Figure 5 with VLM1 (Gemini) exhibiting a higher density of marks along the highest band (>70, <90). Figure 6 shows the respective box plots for grading runs 1–3 using Gemini; the similar medians and interquartile ranges suggest a similar behavior across repeated evaluations; however, several outliers were

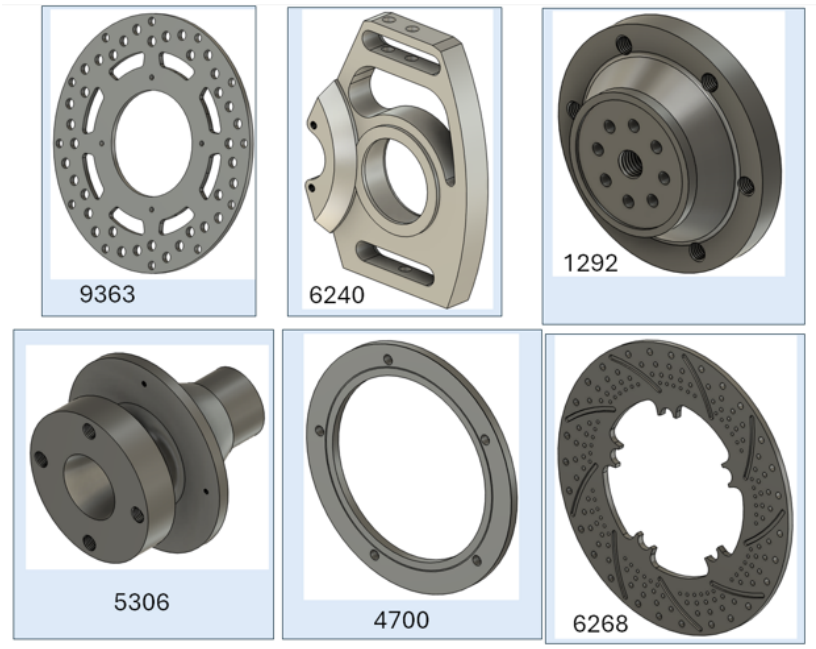


Figure 2. Sample 3D isometric thumbnails provided to vision language models (VLMs) as PNG image files

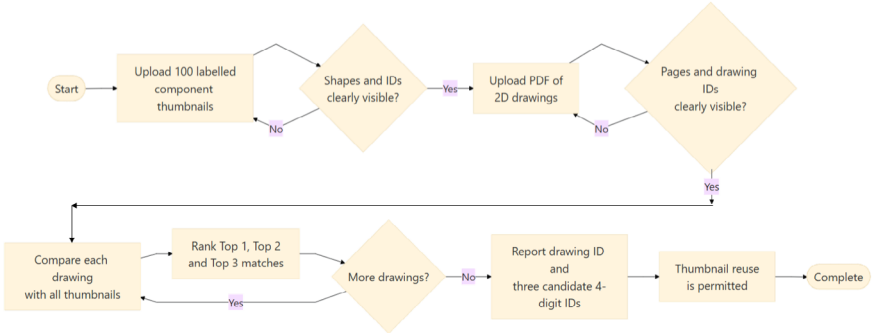


Figure 3. Flow chart of deployment for image sorting analysis

noted especially in run 1, suggesting either difficulties in processing certain specific drawings or an inconsistent behavior across different runs.

4.2.1. Spearman coefficient and ICCs

Table 2 shows a summary of the most significant statistical results in this study; the Spearman coefficient shows whether human assessors and VLMs ranked the drawings similarly. Gemini exhibited moderate agreement across repeated grading runs (mean Spearman ρ = 0.436), meaning it lacks ranking consistency across multiple assessment runs. On the other hand, both Human markers and Copilot exhibited highly stable ranking performance.

Reliability analysis of a single rating via the ICC(2,1) analysis, suggests any results under 0.50 have low reliability for single assessments, with results > 0.75 demonstrating good reliability. The ICC(2,k) index measures reliability across the 3 ranking runs, indicating that for all three assessors (humans and VLMs) averaging multiple runs improved reliability compared with a single grading instance.

4.3. VLM and human marking agreement and comparison

The boxplot analysis (Figure 7) suggests noticeable differences in the central tendency of marks assigned by the three assessors. Gemini generally awarded higher scores than the human markers, with its interquartile range positioned predominantly above the human median. Conversely, Copilot assigned lower scores than the human assessors and Gemini. Despite these differences, the score

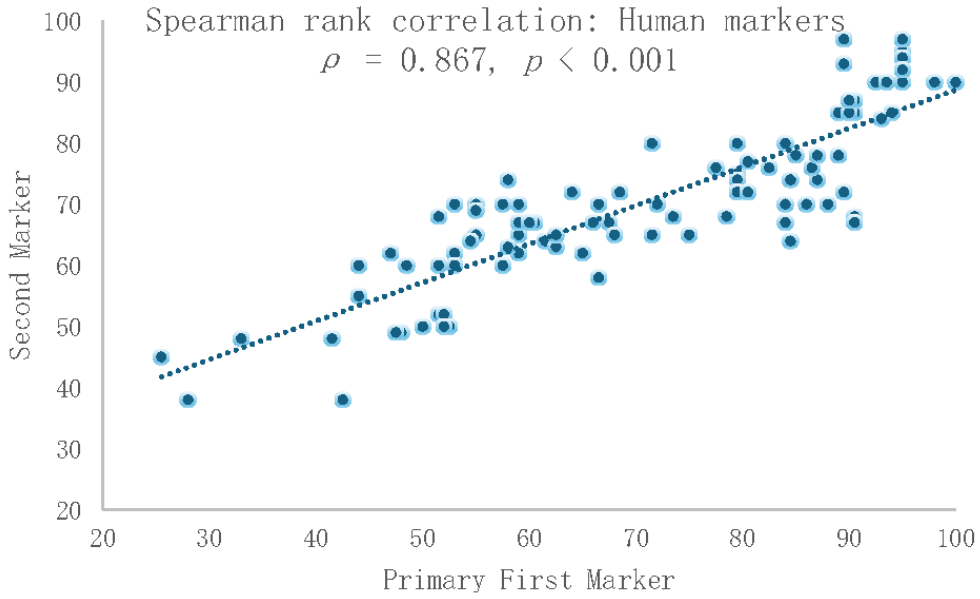


Figure 4. Spearman rank correlation between human markers, with each point representing one engineering drawing

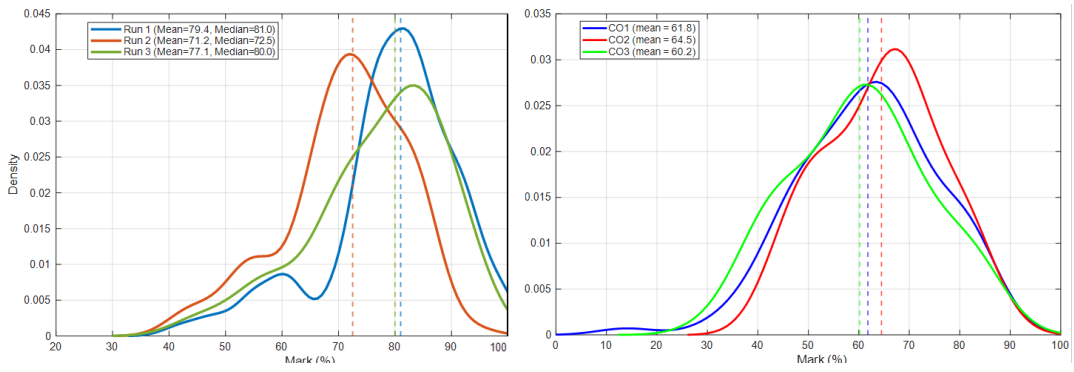


Figure 5. Distribution of marks generated across three repeated runs for both vision language models (VLMs): Gemini (left) and Copilot (right)

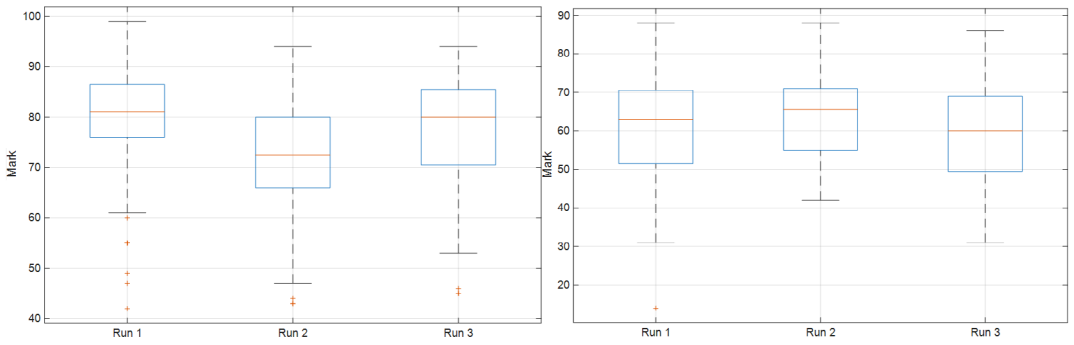


Figure 6. Boxplots show score distributions across three repeated grading runs for Gemini (left) and Copilot (right). Similar medians and interquartile ranges indicate broadly comparable scoring behavior across evaluations, although variability among individual drawings was observed.

Table 2. Summary of ICC and RMSE

Comparison	Spearman ρ	ICC(2,1)	95% CI lower/upper	ICC(2,k)	95% CI lower /upper	RMSE
Human-Human	0.867	0.826	0.76 / 0.874	0.905	0.863 / 0.932	9.421
Gemini repeatability	0.436	0.363	0.262 / 0.45	0.631	0.51 /0.710	13.62
Copilot repeatability	0.890	0.848	0.78 / 0.89	0.944	0.917 /0.96	7.18

Abbreviations: ICC: Intraclass correlation coefficient; RMSE: Root mean square error.

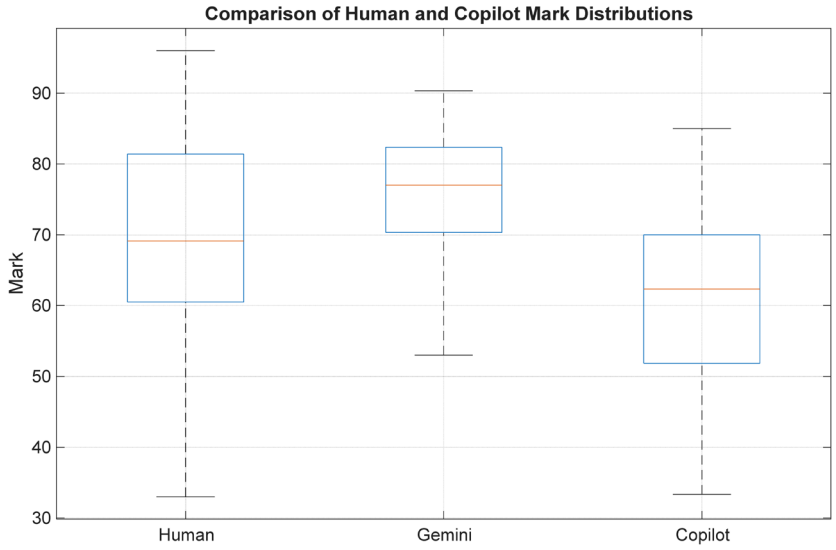


Figure 7. Box-plot results from overall marking across three runs

distributions exhibited substantial overlap, indicating that both VLMs operated within a broadly comparable grading range.

Figure 8 compares the distributions of marks assigned by human experts and both VLMs. The Gemini distribution exhibited a higher density peak and a narrower spread, indicating that scores were concentrated within a smaller range despite a mean score comparable to that of the human markers. In contrast, Copilot produced a distribution more closely resembling that of the human graders, although the slightly elevated density peak suggests a modest reduction in score variability. Overall, Copilot appeared to better replicate the distributional characteristics of human marking than Gemini.

4.4 Human marking direct comparison to VLMs

Both VLMs demonstrated good rank-order agreement compared with human grading with Gemini's Spearman ρ = 0.731 and Copilot's ρ = 0.776, respectively (Table 3). This means drawings that received relatively high marks from human markers also tended to receive high marks from VLMs; therefore, they exhibited similar rankings. However, when it comes to absolute numerical agreement, results were slightly lower for both agents, with ICC(2,1) values between 0.607 (Gemini) to 0.654 (Copilot). This suggests that both VLMs were relatively successful at identifying high and low performers; however, they struggle to allocate similar marks when directly compared to human marking levels.

Results for ICC(2,k) suggest that for both VLMs, averaging the dataset scores improved reliability with both results above the 0.75 threshold. Copilot demonstrated a moderately stronger agreement with human grading, exhibiting both high rank-order association (Spearman ρ = 0.776) and average agreement ICC(2, k) = 0.791, indicating an ability to closely replicate human grading behavior.

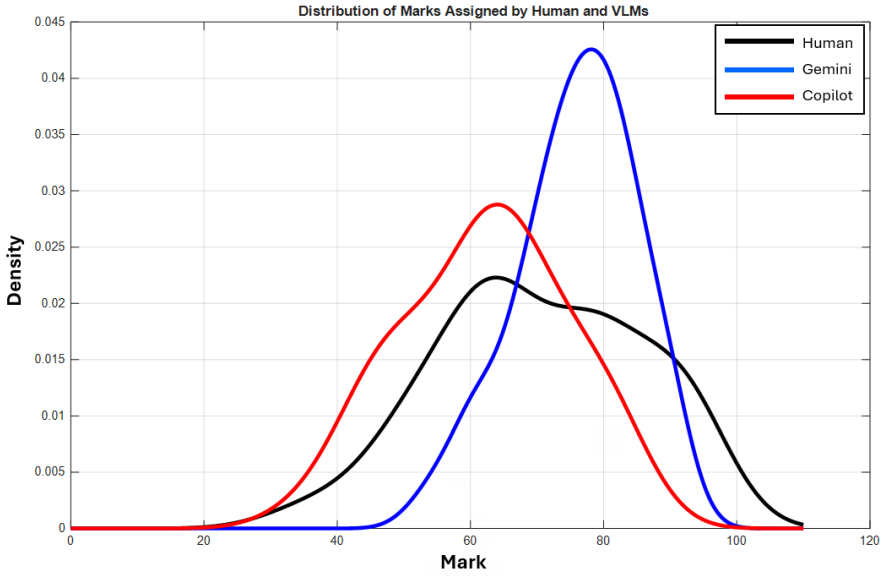


Figure 8. Density function plot showing marking trends for all assessors

4.4.1. Absolute error analysis

The different levels of error for both: intra-dataset, and compared with humans, are reported in Table 4. Intra-marker disagreement between human assessors across the full dataset was 7.61% which is within the acceptable boundary range for internal marking practice (0%–15%), as in normal academic practice, a larger disagreement > 15% usually triggers a third person to act as a moderator on a per-case occurrence. Both VLMs, Gemini and Copilot, achieved average marking errors of approximately 10% relative to both human markers. While this does not constitute a claim of human-equivalent grading performance, it demonstrates that both VLMs operate within a marking deviation range that would not trigger moderation in standard academic practice, suggesting practical utility as a supplementary assessment tool.

When assessing the inner-dataset variability for both agents across their respective three-independent marks per drawing, both VLMs demonstrated strong model consistency, with run-to-run variability of 6.26% for Gemini and 2.91% for Copilot.

4.5. Engineering spatial reasoning

The results from the three independent human markers are summarized in Table 5, with all three assessors scoring a >90% success rate, with each academic reporting a different strategy to undertake the sorting process. It must be noted that all three human markers submitted Top 1 matches only and declined to name 2nd and 3rd options. Experts A and B reported that after a few sorting rounds, the task turned into an unconscious 'sorting' of isometric images by families (uprights, discs, bolts and shafts) with a secondary instinct to look out for detail features like hole patterns and diameter changes. Another subconscious strategy reported by human experts was to take mental notes of pattern features, such as holes arranged in square or pentagonal shapes. Expert A reported an approximate 120 min, and Expert B reported 45 min in completing the task. Having 100 candidate options means that random

Table 3. Agreement analysis between human markers and vision language models (VLMs)

Comparison	Spearman ρ	ICC(2,1)	95% CI lower/upper	ICC(2,k)	95% CI lower/upper	RMSE
Human–Gemini	0.731	0.607	0.51/0.692	0.756	0.756/0.818	11.56
Human–Copilot	0.776	0.654	0.55 /0.73	0.791	0.714/0.846	12.49

Abbreviation: ICC: Intraclass correlation coefficient; RMSE: Root mean square error.

chance performance is 1%, providing a clear baseline against which human (91%–100%) and VLM (17%–38%) performance can be directly contextualized.

Expert C's strategy was to group 2D drawing components before sorting using A5 paper printouts. Any components where the best match was uncertain were marked as 'medium confidence'. The total reported time was 70 min. All experts coincided in their analyses to begin with a specific 2D drawing at a time and then scroll through the 3D-isometric options that best matched that particular shape.

Table 4. Error evaluation across assessors

Comparison	Mean absolute deviation (MAD) (%)
Human–Human	7.61
Gemini–Human	10.9
Copilot–Human	10.7
Gemini intra-run	6.26
Copilot intra-run	2.91

4.5.1. Spatial reasoning VLM and human markers

Human correspondence performance ranged from 91% to 100%, whereas Gemini achieved 57.7% Top 3 accuracy and Copilot achieved 34.0%. Table 5 also shows the results for both VLMs based on: the Top-1 choice being correct, the correct option being within the top 3 options, and finally the MRR score which applies a penalty-multiplier to all choices made as Tiers 2 and 3. In both cases, it is evident that VLMs were outperformed by all human markers for this particular task.

Table 5. Results from the retrieval and matching task

Evaluator	Top-1 correct match	Top-3 correct match	MRR score
Human 1	100	-	-
Human 2	91	-	-
Human 3	98	-	-
VLM1: Gemini	37.71	57.71	46.41
VLM2: Copilot	17.70	34.02	24.80

4.6. Relationship between spatial reasoning and grading performance

The cross-analysis results are summarized in Table 6 and Figure 9. For Gemini, a statistical difference was established between spatial reasoning performance and grading accuracy under both Top 1 (p = 0.025) and Top 3 (p = 0.009) conditions. This means that drawings correctly matched in the spatial reasoning task exhibited consistently lower grading errors relative to human markers compared to incorrectly matched drawings. This suggests that Gemini's grading performance is partially coupled to its spatial reasoning ability. On the other hand, no significant association was found for Copilot under either condition (Top 1: p = 0.063; Top 3: p = 0.162), with grading errors showing no consistent directional trend relative to spatial matching outcome. This indicates that Copilot's grading performance operates largely independently of its spatial reasoning ability, consistent with the hypothesis that it relies predominantly on OCR and pattern recognition rather than geometric understanding.

5. Discussion

On the initial grading experiment, Copilot demonstrated strong repeatability across repeated assessments (Spearman ρ = 0.890; ICC(2,1) = 0.848), exhibiting a reliable behavior comparable to the human assessors (ρ = 0.867; ICC(2,1) = 0.826). Gemini exhibited substantially lower rank repeatability (ρ = 0.436; ICC(2,1) = 0.363), although averaging multiple runs improved reliability (ICC(2,k) = 0.631). It was demonstrated that a single analysis run may not yield consistent results when using a specific VLM; however, having multiple instances increases the overall reliability. It must be noted that, unlike traditional computer vision pipelines which process text and graphic content through separate modular stages, VLMs perform joint multimodal inference; therefore, artificially isolating individual modality contributions was not part of this study as it would fundamentally alter the operational nature of the system being evaluated and would not reflect authentic deployment conditions.

5.1. Error analysis significance

The average marking error between human markers and VLMs contrasts sharply with both VLMs' poor performance on spatial engineering ability, but also with their low to average agreement following the ICC(2,1) analysis (Table 3). One critical aspect on the VLM and human marking similarity may be reflected on the marking rubric itself, as the results may suggest that the current marking criteria heavily relies on visible drawing features which are easily detectable by VLM agents as straightforward visual cues (missing centerlines, missing

Table 6. Grading error comparison by spatial reasoning outcome for both vision language models (VLMs)

VLM	Condition	Correctly matched error (%)	Incorrectly matched error (%)	p-value
Gemini	Top 1	9.75	12.77	0.025*
Gemini	Top 3	10.1	14.12	0.009**
Copilot	Top 1	12.56	9.79	0.063
Copilot	Top 3	11.57	9.77	0.162

Notes: *$p < 0.05$; **$p < 0.01$.

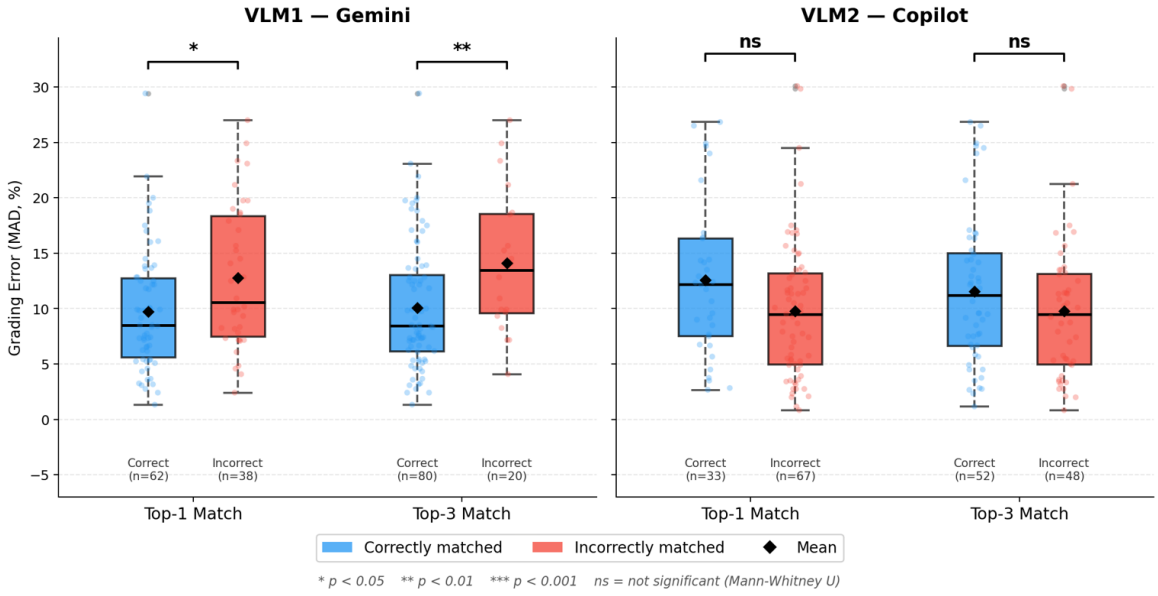


Figure 9. Relationship between 2D-3D spatial matching accuracy and grading error per drawing for both VLMs
Abbreviations: MAD: Mean absolute deviation; VLMs: Vision language models.

tolerances, template information, *etc.*). This suggests that rubric-based engineering drawing assessment may rely more heavily on pattern recognition and standards compliance than on the deeper geometric correspondence capabilities required for spatial reasoning tasks.

5.2. Spatial reasoning analysis

Spatial analysis in engineering design is a well-studied topic in engineering education and educational psychology where an expert or educator will typically execute multiple cognitive processes simultaneously, such as visual perception, pattern recognition, mental rotation, working memory, and decision making, across multiple others.[27,28] In this work, we propose a novel single-image sorting strategy that promotes problem solving without prescriptive constraints. As such, while some human experts preferred to perform 2D-3D isometric image matchmaking on the screen, others preferred to print out individual copies of each dataset for mental clarity. There is no specific correlation drawn between the time to solve the sorting task and the overall success rate; however, it was generally acknowledged to be a lengthy and mentally challenging operation requiring clustering ability, pattern identification, and memory.

Most 3D isometric images were intentionally rotated and presented in opposite views and angles compared to those included in 2D drawing projections; therefore, the task also involved an element of mental rotation, requiring human experts to cognitively transform 2D orthographic views into 3D representations and mentally rotate these objects to interpret their spatial structure.

5.3. Spatial reconstruction ability as an indicator of assessment ability?

We believe both experiments to be representative of two different qualities and capacities by VLMs. In the first instance, AI demonstrated a proven ability to understand and assess 2D drawings making use of a wide range of capabilities deployed in parallel, such as PDF parsing to identify vector graphics or embedded text, image rendering, Optical Character Recognition (OCR), object detection, language reasoning, and visual reasoning.[29] It must be noted that VLMs solve 2D-drawing navigation differently to traditional computer vision pipelines commonly reported in literature, as they do not detect every primitive or vector lines or nodes explicitly; instead, they perform a joint representation of all elements present in a single drawing while at the same time accessing their understanding of the whole 2D drawing collection.[30]

In the second instance, both VLMs were outperformed by the three human expert markers. Although these results are valid for the current experiment with this set of variables and curated information, it is not a reflection on the wider VLM capability for spatial reasoning, as several factors could have affected VLMs reasoning capacity, such as the lack of isometric projections on the 2D-drawings, which were removed to prevent the study from becoming a simple image-recognition and sorting task.

Both VLMs demonstrated a consistent ranking capacity, suggesting that they have the ability to repeatedly identify, cluster and organize graphical work according to the perceived correctness; however, their grading or marking ability is not consistent and can benefit from a robust and dedicated prompting engineering to fine-tune their grading behavior, trends and limits as demonstrated by previous literature.[31]

On the spatial reasoning and 2D/3D matching, leaving the name visible can let the model exploit text semantics, instead of geometry. Engineering-drawing and text-rich multimodal benchmarks explicitly separate OCR/text understanding from spatial reasoning because OCR is often the easiest capability and can dominate performance if left unchecked.

5.4. VLMs' capacity for large graphical datasets

Preliminary experiments revealed a substantial performance degradation for large-scale document matching when the complete dataset was used at once. During the early stages of data-input into both VLMs, it was observed that analysis stopped automatically after 40 pages, after which, both VLMs produced hallucinated results based on extrapolated and approximated information from the earlier pages. We concluded that the current state of development of commercial VLMs is not sufficiently reliable for large multi-page or multi-file datasets without being prone to hallucinations. Therefore, we recommend splitting tasks into well-controlled batches across different independent sessions.

The results indicate that expert human assessors possess significantly higher capability than current general-purpose VLMs in performing correspondence matching orthographic engineering drawings and isometric representations. The large performance disparity suggests that human expertise incorporates domain-specific spatial cognition mechanisms, including graphical interpretation, feature-based reasoning, and engineering drawing schema developed through experience, which are not yet reliably replicated by contemporary VLMs.

6. Conclusions

This research has been undertaken on two agentic general-purpose AI models, with no specialized engineering pre-training and no CAD-specific representation learning either via contextual prompting or short-shot of human-in-the-loop interventions. This is to assess the native ability of VLM agents to: (i) rank and grade engineering design submissions, and (ii) to assess its basic engineering visual-spatial corresponding ability. It must be highlighted that this test in isolation is not an indication of general spatial intelligence, which is assumed to be better performing when a learning protocol is established beforehand.

Our sample size of 100 drawings and image datasets demonstrate a clear trend; the preliminary findings reveal a substantial performance gap between expert human assessors and current general-purpose VLMs when performing correspondence matching between 2D orthographic drawings and 3D isometric representations. While expert assessors had a high performance of 96% average, both evaluated VLMs demonstrated considerably lower success rates (17% and 37%). This suggests that contemporary VLMs may not yet possess the domain-specific spatial reasoning capabilities required for reliable interpretation of engineering drawings. However, these findings should be interpreted cautiously, as the experiment evaluates a specific engineering visual reasoning task and compares experienced human experts against general-purpose AI systems without engineering-specific training or fine-tuning.

Finally, the major outcome of this work is not that human expert markers' performance is superior to AI in the field of engineering spatial recognition, but it raises a question of whether spatial reconstruction ability is a

real limiting factor for VLM-based engineering drawing assessment. We have also identified future work directions on the criterion-level analysis of VLM grading performance across individual rubric categories to investigate further the different cognitive requirements each assessment element poses on current VLM capacity.

Acknowledgments

None.

Funding

None.

Conflict of interest

The authors declare they have no competing interests regarding the core subject matter discussed in this manuscript.

Author contributions

Conceptualization: All authors
Investigation: All authors
Methodology: Javier Munguia
Supervision: Jing Fu
Writing–original draft: Javier Munguia
Writing–review & editing: Yudhi Ariadi, Lim Jen Nee Jones, Jing Fu

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

The original data presented in this study are openly accessible on Zenodo (doi: 10.5281/zenodo.21905698).

References

1. Xu Y, Zhang C, Xu Z, *et al.* Tolerance Information Extraction for Mechanical Engineering Drawings – A Digital Image Processing and Deep Learning-based Model. *CIRP J Manuf Sci Technol*. 2024;50:55-64.

   doi: 10.1016/J.CIRPJ.2024.01.013

2. Otey J, Company P, Contero M, Camba JD. Revisiting the design intent concept in the context of mechanical CAD education. *Comput Aided Des Appl*. 2018;15(1):47-60.

   doi: 10.1080/16864360.2017.1353733

3. Kadam K, Mishra S, Deep A, Iyer S. Enhancing engineering drawing skills via fostering mental rotation processes. *Eur J Eng Educ*. 2021;46(5):796-812.

   doi: 10.1080/03043797.2021.1920891

4. Xie W, Niu J, Xue CJ, Guan N. Grade Like a Human: Rethinking Automated Assessment with Large Language Models. In: Proceedings of the International Conference on Research in Adaptive and Convergent Systems. New York, NY, USA: ACM; 2025:1-8.

   doi: 10.1145/3769002.3769962

5. Edwards KM, Tehranchi F, Miller S, Ahmed F. AI Judges in Design: Toward Expert-Equivalent Design Evaluations With Vision-Language Models and In-Context Learning. *J Mech Des*. 2026;148(7).

   doi: 10.1115/1.4071835

6. NexCAD Ltd. NexCAD - AI Drawing Checker. Accessed August 3, 2026. https://nexcad.ai/

7. 3YOURMIND. AI-Powered Technical Drawing Analysis by 3YOURMIND. Accessed August 3, 2026. https://www.3yourmind.com/ai-powered-technical-drawing-analysis

8. Dessia Technologies. Dessia. Accessed August 3, 2026. https://www.dessia.io/case-studies/automated-2d-drawing-check-for-design-completeness

9. CoLab Software. AI Engineering Drawing Checker | CoLab. Accessed August 3, 2026. https://www.colabsoftware.com/product/ai-drawing-reviews?utm_source=google&utm_medium=paidsearch&utm_campaign=PM-GOOG-SRC-AO-ai-engineering-low-ctr-organic-keywords&utm_term=ai+for+engineering+design&utm_content=ai-for-engineering-design

10. Zhang J, Huang J, Jin S, Lu S. Vision-Language Models for Vision Tasks: A Survey. *IEEE Trans Pattern Anal Mach Intell*. 2024;46(8):5625-5644.

    doi: 10.1109/TPAMI.2024.3369699

11. Chen S, Zhu T, Zhou R, *et al.* Why Is Spatial Reasoning Hard for VLMs? An Attention Mechanism Perspective on Focus Areas. In: Proceedings of the 42nd International Conference on Machine Learning. Vancouver, Canada; 2025;267:9910-9932. Accessed July 17, 2026. https://proceedings.mlr.press/v267/chen25cr.html

12. Zhang Z, Hu F, Lee J, *et al.* Do Vision-Language Models Represent Space and How? Evaluating Spatial Frame of Reference Under Ambiguities. In: 13th International Conference on Learning Representations, ICLR 2025. Singapore; 2025:14036-14062. Accessed August 12, 2026. https://arxiv.org/pdf/2410.17385

13. How to Use AI for Manufacturing Drawing Reviews: A Step-by-Step Guide. Accessed August 3, 2026. https://www.colabsoftware.com/post/ai-for-manufacturing-drawing-checks-a-guide-for-engineering-teams

14. Merzdorf HE, Jaison D, Weaver MB, Linsey J, Hammond T, Douglas KA. Work In Progress: An Object Assembly Test of Sketching in Undergraduate Engineering. *Proceedings* -

Frontiers in Education Conference, FIE. Uppsala, Sweden; 2022;1-5.

doi: 10.1109/FIE56618.2022.9962634

15. Raju G, Sorby S. The Role of Spatial Skills and Sketching in Engineering Design Problem Solving. In: *2024 ASEE Annual Conference & Exposition Proceedings*. Portland, Oregon. 2024.

doi: 10.18260/1-2--48133

16. Ernst JV, Lane D, Clark AC. Pictorial Visual Rotation Ability of Engineering Design Graphics Students. *Eng Des Graph J.* 2015;79(1):1-13.

doi: 10.18260/edgj.v79i1.510

17. Han W, Xiang S, Liu C, Wang R, Feng C. SPARE3D: A Dataset for SPAtial REasoning on Three-View Line Drawings. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA; 2020:14678-14687.

doi: 10.1109/CVPR42600.2020.01470

18. Kunz L, Klostermeier M, Thanabalan K, Legler T, Ruskowski M, Autor K. TechMB: Exploring the Potential of Vision Language Models for Interpreting Technical Drawings. In: DS 140: Proceedings of the 36th Symposium Design for X (DFX2025). Hamburg, Germany: 2025:179-188.

doi: 10.35199/DFX2025.19

19. Wu S, Khasahmadi AH, Katz M, *et al.* CadVLM: Bridging Language and Vision in the Generation of Parametric CAD Sketches. In: *Lecture Notes in Computer Science*. Switzerland: Springer Nature; 2024:368-384.

doi: 10.1007/978-3-031-72897-6_21

20. Monjoree U, Yan W. AI's Spatial Intelligence: Evaluating AI's Understanding of Spatial Transformations in PSVT:R and Augmented Reality. *arXiv*. Preprint posted online 2024.

doi: 10.48550/arXiv.2411.06269

21. Picard C, Edwards KM, Doris AC, *et al.* From concept to manufacturing: evaluating vision-language models for engineering design. *Artif Intell Rev.* 2025;58(9).

doi: 10.1007/S10462-025-11290-Y

22. Khan MT, Chen L, Ng YH, Feng W, Tan NYJ, Moon SK. Fine-Tuning Vision-Language Model for Automated Engineering Drawing Information Extraction. *arXiv*. Preprint posted online 2024.

doi: 10.48550/ARXIV.2411.03707

23. Zhou J, Camba JD, Company P, Contero M. Drawing-Checker: A Vision RAG Framework for Automated Comparison of Engineering Drawings. *Procedia CIRP*. 2026;142:127-132.

doi: 10.1016/j.procir.2026.05.235

24. Doris AC, Grandi D, Tomich R, *et al.* DesignQA: A Multimodal Benchmark for Evaluating Large Language Models' Understanding of Engineering Documentation. *J Comput Inf Sci Eng.* 2025;25(2):021009.

doi: 10.1115/1.4067333

25. University of Bristol. ES3C2-15 Advanced Mechanical Engineering Design - Module Catalogue. Accessed August 3, 2026. https://courses.warwick.ac.uk/modules/2021/ES3C2-15

26. Munguia J. 2D Engineering drawing collection + Isometric representations. *Zenodo*. Preprint posted online August 2026.

doi: 10.5281/zenodo.21905698

27. Bartlett KA, Dorribo Camba J. The role of a graphical interpretation factor in the assessment of Spatial Visualization: A critical analysis. *Spat Cogn Comput*. 2021;23(1):1-30.

doi: 10.1080/13875868.2021.2019260

28. Bartlett KA. Engineering Design and Spatial Skills: A Systematic Literature Review. *J Mech Des.* 2026;148(4):040901.

doi: 10.1115/1.4070965

29. Hertanti A, Retnawati H, Wutsqa DU. The role of spatial experience in mental rotation. *J Phys Conf Ser*. 2019;1320(1):012043.

doi: 10.1088/1742-6596/1320/1/012043

30. Gan Z, Li L, Li C, Wang L, Liu Z, Gao J. Vision-Language Pre-training: Basics, Recent Advances, and Future Trends. *Found Trends Comput Graph Vis.* 2022;14(3-4):163-352.

doi: 10.1561/0600000105

31. Munguia J. Leveraging AI for 2D technical drawing analysis, feedback and assessment in higher education. *Int J Mech Eng Educ.* 2026.

doi: 10.1177/03064190251414394

Appendix

Prompt deployment for image sorting analysis
I'll provide you with a collection of images, each one containing individual thumbnails of mechanical components with an individual 4-digit-ID code. Please make sure you can visualize and access the shape and geometric details of all 100 thumbnails.
I'll now provide you with a collection of 2D drawings in a single PDF file, listed with an ID code located on the Top-left corner of each page. Can you confirm you have clear view, access and full spatial recognition of every single page?
Considering the total collection of 100 image thumbnails, please find the matching 4-digit ID for each specific 2D drawing. For each guess, please provide the Top 1, Top 2 and Top 3 options. Please note you may select the same Isometric multiple times if needed.