CReFT-CAD: Boosting Orthographic Projection Reasoning for CAD via Reinforcement Fine-Tuning

Anonymous Author(s)
Affiliation
Address
email

A The complete reasoning guidance for each input format

For the Test Image-Only format:

- Identify all annotated numbers in the orthographic projection, including dimensions, spacing, and height.
- Interpret their meanings based on position, orientation, and surrounding context.
- Directly assign values to some parameters, count graphical elements for quantity parameters, and compute spacing parameters based on position.

For the +Reference Image format:

- Use the reference template to understand the correspondence between primitive and parameters, and identify annotated numbers in the target image.
- Assign parameters based on position and template primitive, counting graphical primitive for quantities.
- Calculate parameters based on positional relationships or template definitions, and output the full set of parameter values.

For the +Answered Image format:

- Learn the correspondence between structures and parameters from the example image.
- Identify and interpret annotated numbers in the target image, considering their position and similarity to the example.
- Assign values to parameters, count graphical primitive for quantities, calculate parameters, and apply necessary constraints to produce final results.

For the +Attribute Explanation format:

- Clarify parameter definitions and their geometric meanings, including component types and their properties.
- Identify annotated numbers in the orthographic projection and infer their corresponding parameters.
- Assign values to parameters, compute quantities, derive spacing values from positional relationships, and output the complete parameter set with necessary constraints.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

Table 1: Performance comparison of various VLMs on orthographic projection reasoning tasks.

PROMPTS	Without Reasoning Guidance				Reasoning Guidance			
	Test Img	+Reference Img	+Answered Pair	+Attribute Explanation	Test Img	+Reference Img	+Answered Pair	+Attribute Explanation
Phi-3.5-Vision	4.22	8.95	8.05	6.26	11.32	14.40	16.16	15.79
LLaVA-OneVision	9.16	16.06	15.65	14.84	9.10	16.81	16.21	20.53
DeepSeek-VL	8.16	22.68	20.03	12.65	14.02	20.31	24.62	25.45
InternVL2.5	15.79	18.82	23.30	17.16	15.63	22.35	24.47	23.73
InternVL3	15.46	17.90	22.44	17.91	15.81	21.98	23.58	17.05
Qwen2.5-Omni	23.43	28.89	28.74	26.50	26.12	30.93	29.40	35.71
Qwen2.5-VL	24.54	30.76	30.47	25.86	24.54	32.78	33.64	38.88
Ours	**80.86**	**82.99**	**83.24**	**82.67**	**81.35**	**83.11**	**82.87**	**84.03**

Table 2: Performance comparison of training-free model, SFT model,and our GRPO-based model on orthographic projection reasoning tasks.

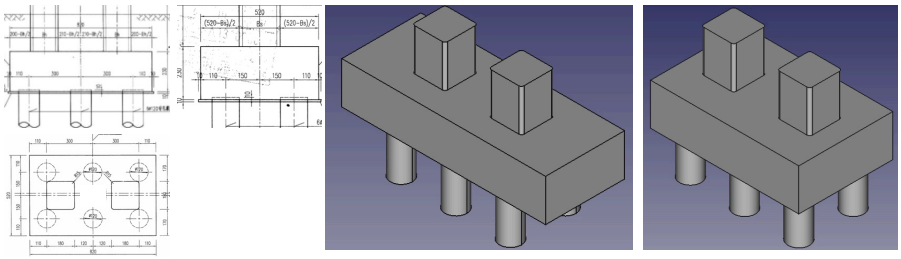
PROMPTS	Without Reasoning Guidance				Reasoning Guidance			
	Test Img	+Reference Img	+Answered Pair	+Attribute Explanation	Test Img	+Reference Img	+Answered Pair	+Attribute Explanation
Qwen2.5-VL	24.54	30.76	30.47	25.86	24.54	32.78	33.64	38.88
Qwen2.5-VL(SFT)	76.33	79.50	77.12	78.64	76.42	76.78	77.20	80.30
Ours	**80.86**	**82.99**	**83.24**	**82.67**	**81.35**	**83.11**	**82.87**	**84.03**

B Further analysis regarding the three findings

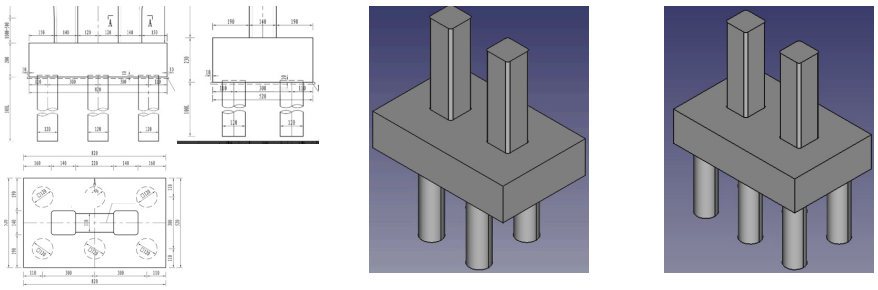
1) Remains a Tough Challenge for pretrained VLMs. Tab. 1 reports performance across four prompt formats without and with reasoning guidance. Overall, orthographic projection reasoning entails not only reading textual dimensions and visual features, but also matching annotations to geometry primitives, counting structural elements, and computing composite parameters. Consequently, none of the off-the-shelf models fully masters this composite task. Qwen2.5-VL Wang et al. [2024] achieves the highest accuracy—38.88% under reasoning guidance with the Test Image + Attribute Explanation prompt. This improvement stems primarily from its superior ability to parse geometry layers and read dimension labels. Crucially, its performance on reasoning-intensive parameters remains low. These findings underscore that orthographic projection CAD cannot be solved by prompting alone and requires dedicated fine-tuning strategy.

2) Prompt Format–Dependent Performance Gains. Appending an Attribute Explanation to the Test Image consistently boosts accuracy compared to the image-only baseline, with an average increase of 3 to 4 percentage points across models. This demonstrates that strong text encoders can leverage detailed, engineer-authored descriptions to guide complex geometric and numerical inferences. Similarly, providing a Reference Image or an Answered Pair results in comparable improvements, with a 4 to 5 percentage point increase in accuracy. These exemplars offer explicit visual-textual templates, simplifying the task of matching primitives to their corresponding semantic labels and dimensions. In contrast, using raw images forces models to address both perception and reasoning simultaneously, leading to the lowest performance. Collectively, these findings indicate that multimodal exemplars can enhance existing VLMs' ability to reason over orthographic projection. However, this further underscores that additional fine-tuning strategies are crucial to fully bridging the remaining performance gap.

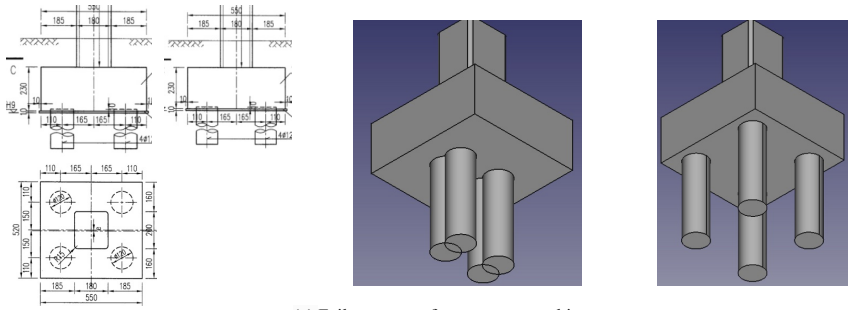
3)Introducing reasoning guidance yields consistent gains across all seven VLMs. Introducing reasoning guidance demonstrates that step-wise reasoning guidance helps models better decompose the multi-step orthographic projection tasks. In the absence of reasoning guidance, visual exemplar prompts (+Reference Image and +Answered Pair) deliver the largest relative improvements. However, when reasoning guidance is added, the Attribute Explanation prompt shows the greatest uplift. It stems from reasoning guidance is textual form, which synergizes most effectively with textual inputs. Hence, models with stronger text encoders (e.g., DeepSeek-V1) exhibit disproportionately larger boosts. These results underscore the necessity of integrating structured reasoning guidance to advance orthographic projection CAD reasoning.



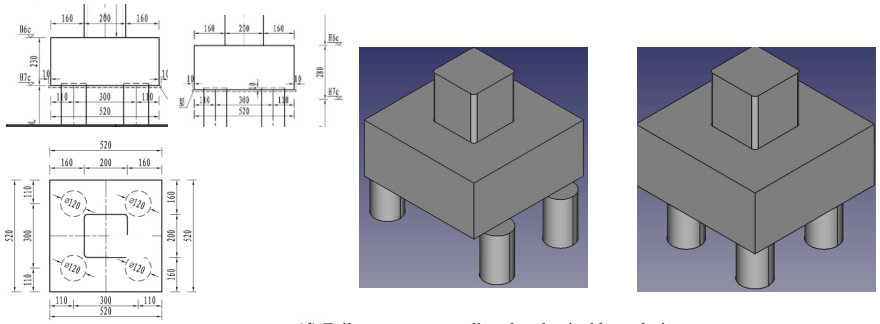
(a) Failure cases due to incorrect dimensions



(b) Failure cases due to incorrect counting



(c) Failure cases of component stacking



(d) Failure cases exceeding the physical boundaries

Figure 1: Four sets of failure cases. Each set consists of three parts: the left part shows the orthographic projection, the middle part presents the failed 3D model construction, and the right part illustrates the correctly constructed 3D model.

C More performance on in-domain testing set

Tab. 2 presents the performance in-domain testing set. The first row (Qwen2.5-VL) represents a training-free baseline, which exhibits consistently poor performance across all settings, indicating the inherent difficulty of this reasoning task without any fine-tuning or adaptation. The second row (Qwen2.5-VL with SFT) demonstrates a significant performance improvement, particularly when provided with reference images, answered pairs, and attribute explanations. However, despite

these gains, the model still suffers from notable drops in more complex reasoning tasks, especially those involving composite parameter computation. In contrast, our method consistently outperforms both baselines across all settings, achieving the highest accuracy in all prompt configurations, with particularly strong results under reasoning guidance. These results validate the effectiveness of our approach in handling complex geometric reasoning and highlight its robustness across both simple and compositional inference scenarios.

D Failure cases and analysis

As illustrated in Fig. 1, we present four distinct failure cases, each highlighting different types of errors that can occur during the 3D model reconstruction process due to incorrect 2D parameterization. Fig. 1(a) demonstrates the error induced by incorrect dimensional parameters. Specifically, due to errors in the parameterization of certain dimensions, the constructed 3D model exhibits significant scaling issues, leading to a distorted and non-accurate representation. This failure emphasizes the critical importance of precise dimensional input when translating from 2D projections to 3D models, as even minor errors in size parameters can drastically affect the final output. Fig. 1(b) illustrates a failure caused by incorrect counting, which results in an incorrect number of primitives being identified and subsequently integrated into the 3D model. The discrepancy in the number of primitives reflects the direct impact that improper counting and recognition of geometric elements can have on the success of the 3D reconstruction. In Fig. 1(c), we observe a stacking issue caused by errors in the calculation of composite parameters. The failure manifests as improper alignment of components in the final model, where parts of the 3D structure fail to align correctly with each other. Lastly, Fig. 1(d) shows a failure resulting from exceeding physical boundaries due to incorrect composite parameter calculations. In this case, the incorrect parameters lead to elements of the 3D model extending beyond the physical constraints or limits of the original design. Together, these four failure cases provide strong evidence that errors in the parameterization of 2D models directly lead to inaccuracies in 3D model reconstruction. Each case illustrates the cascading effect of parameter errors from the 2D representation to the final 3D model, further underscoring the necessity for robust and precise parameterization methods in 2D-to-3D model conversion processes.

References

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.