Quantitative Evaluation and Domain Adaptation of Vision–Language Models for Mixed-Reality Interpretation of Indoor Environmental Computational Fluid Dynamics Visualizations

Soushi Futamura and Tomohiro Fukuda *

Posted Date: 31 January 2026

doi: 10.20944/preprints202601.2306.v1

Keywords: vision-language models (VLMs); mixed reality (MR); computational fluid dynamics (CFD); domain adaptation; thermal environment

Article

Quantitative Evaluation and Domain Adaptation of Vision–Language Models for Mixed-Reality Interpretation of Indoor Environmental Computational Fluid Dynamics Visualizations

Soushi Futamura and Tomohiro Fukuda

Divisions of Sustainable Energy and Environmental Engineering, Graduate School of Engineering, The University of Osaka, 2-1 Yamadaoka, Suita, Osaka 565-0871, Japan

* Correspondence: fukuda.tomohiro.see.eng@osaka-u.ac.jp

Abstract

In built environmental design, incorporating building user participation and verifying indoor thermal performance at early design stages have become increasingly important. Although Computational Fluid Dynamics (CFD) analysis is widely used to predict indoor thermal environments, its results are difficult for non-expert stakeholders to interpret, even when visualized using Mixed Reality (MR). Interpreting CFD visualizations in MR requires quantitative reasoning that explicitly cross-references visual features with legend information, rather than relying on prior color–value associations learned from natural images. This study investigates the capability of Vision–Language Models (VLMs) to interpret MR visualizations of CFD results and respond to user queries. We focus on indoor temperature distributions and airflow velocities visualized in MR. A novel dataset was constructed, consisting of MR images with CFD results superimposed onto real indoor spaces, paired with domain-specific question–answer annotations requiring legend-based reasoning. Using this dataset, a general-purpose VLM (Qwen2.5-VL) was fine-tuned. Experimental results show that the baseline model achieved less than 30% accuracy, whereas fine-tuning improved accuracy to over 60% across all categories while largely preserving general reasoning performance. These results demonstrate that domain adaptation enables VLMs to quantitatively interpret physical information embedded in MR visualizations, supporting non-expert understanding in built environmental design.

Keywords: vision-language models (VLMs); mixed reality (MR); computational fluid dynamics (CFD); domain adaptation; thermal environment

1. Introduction

In recent years, consensus building among diverse stakeholders, including building occupants, has become increasingly important in architectural and built environmental design to enhance occupant satisfaction [1]. People spend approximately 90% of their time indoors [2], and this proportion has further increased since the COVID-19 pandemic [3]. As a result, indoor environmental conditions play a critical role in occupants' productivity, health, and overall well-being [4]. Accordingly, involving occupants in the design process has become essential for achieving higher levels of satisfaction with indoor environments.

Despite its importance, effective consensus building remains challenging. Non-expert stakeholders, such as building occupants, often struggle to comprehend information that requires specialized expertise, including the interpretation of two-dimensional (2D) drawings and technical design documents [5]. This difficulty creates a substantial gap in understanding between experts (e.g.,

 doi:10.20944/preprints202601.2306.v1

architects and engineers) and non-experts, which frequently results in mismatched expectations and reduced post-occupancy satisfaction [6].

To mitigate this gap, various methods for visually presenting design information have been explored [7]. Among them, Mixed Reality (MR) has attracted increasing attention [8,9]. By superimposing digital design models onto real-world spaces, MR enables users to intuitively understand spatial relationships, scale, and geometry directly on site. Following the definition by Milgram and Kishino, MR refers to a continuum in which real and virtual environments are combined [10]. In this study, Augmented Reality (AR) is treated as a subset of MR, as the precise mixing ratio between real and virtual content is not explicitly specified.

However, consensus building in environmental design requires consideration not only of visible spatial elements—such as geometry and furniture layout—but also of invisible physical factors, including temperature, airflow, and acoustics. Among these, thermal comfort is a particularly influential determinant of occupant productivity and comfort [11]. To support design evaluation, prior studies have proposed visualizing otherwise invisible physical phenomena by conducting Computational Fluid Dynamics (CFD) simulations and overlaying the results onto indoor spaces using MR, typically in the form of contour plots or vector fields [12]. Nevertheless, users are susceptible to perceptual biases when interpreting color maps [13]. As a result, simply visualizing CFD results in this manner does not allow non-experts to intuitively understand indoor environmental performance or comfort levels. Consequently, disparities in interpretation ability between experts and non-experts persist, hindering effective information sharing and consensus building in environmental design.

One potential approach to addressing this issue is the application of Vision–Language Models (VLMs), which can jointly process visual and linguistic information. While VLMs have demonstrated strong performance in various tasks, most existing studies focus on natural images and everyday visual content [14,15]. In contrast, limited research has examined whether VLMs can interpret scientific visualization data that represent physical quantities, such as CFD simulation results commonly used in architectural and environmental design [16]. Moreover, when VLMs are applied to MR images, real-world background imagery may introduce visual noise that interferes with the recognition of virtual objects. To date, few studies have systematically evaluated whether VLMs can accurately interpret virtual information embedded within MR scenes [17].

Therefore, assessing whether VLMs can correctly recognize and reason about color-encoded physical information—such as contour plots superimposed onto real indoor environments—is essential for evaluating the feasibility of integrating MR and VLMs in future environmental design workflows.

In this study, we construct a novel dataset comprising MR images that visualize indoor CFD analysis results, specifically contour plots and isolines, along with corresponding question–answer pairs. Using this dataset, we evaluate the reasoning and interpretation capabilities of VLMs with respect to scientific visualization data and examine the effectiveness of domain adaptation through fine-tuning. By enabling VLMs to verbalize and explain visualized physical information, this research aims to support non-expert stakeholders in understanding future indoor environments, thereby facilitating smoother consensus building and improving satisfaction with indoor environmental design.

2. Literature Review

2.1. Development and Current Status of Vision–Language Models

In recent years, Large Language Models (LLMs) have advanced rapidly in the field of artificial intelligence, with state-of-the-art models exhibiting language understanding and generation capabilities comparable to those of humans [18]. However, because LLMs process only textual information, their applicability to real-world tasks is inherently limited. Many practical scenarios rely heavily on visual information, such as spatial relationships and geometric configurations, which

 doi:10.20944/preprints202601.2306.v1

cannot be fully conveyed through text alone [19]. This limitation is particularly pronounced in architectural and environmental design, where complex indoor layouts and spatial arrangements are difficult to express precisely using language only, and where LLMs struggle to infer accurate spatial structures from textual descriptions [20]. Consequently, extending AI applications to environmental design requires models with advanced visual and spatial understanding capabilities.

To overcome these limitations, VLMs, which jointly process visual and linguistic information, have been actively developed [21,22]. VLMs typically integrate the language processing capabilities of LLMs with vision encoders—such as Vision Transformers [23]—enabling them to perform image-based tasks including image captioning, visual question answering, and multimodal reasoning.

As VLMs have matured, their application scope has expanded beyond natural images, such as everyday scenes and landscapes, to the interpretation of graphical and diagrammatic data, including charts, plots, and technical illustrations [24–26]. Effective interpretation of such graphical data requires the ability to process high-resolution images and achieve high optical character recognition (OCR) accuracy. Recent models, such as Qwen2.5-VL [27], address these requirements by dynamically splitting high-resolution input images into manageable segments, allowing them to process images at their original resolution. As a result, these models demonstrate superior OCR performance and enhanced capabilities for interpreting complex graphical information.

Nevertheless, existing benchmarks for graphical data understanding [24–26] are largely limited to graphs with clearly defined axes or discrete data points. As a result, the ability of VLMs to interpret physical quantities encoded through continuous color variations—such as those found in scientific visualizations—has not been systematically evaluated.

In addition, applying general-purpose VLMs to specialized technical domains remains challenging. Because these models are predominantly trained on natural images and generic diagrams, they often lack sufficient domain-specific knowledge required for accurate interpretation in specialized fields. One commonly adopted approach to addressing this limitation is fine-tuning, which enables domain adaptation by further training a pre-trained general-purpose model using domain-specific datasets [28].

2.2. Adaptation of VLMs to Visualized Physical Information and MR Image

To utilize VLMs for explaining CFD analysis results, models must be capable of interpreting images that visualize inherently invisible physical quantities, such as temperature distributions and airflow velocities. One of the few studies evaluating VLMs in this context is that of Moshtaghi et al. [16], who constructed a paired dataset of RGB and thermal images to assess the ability of VLMs to interpret thermal information. Their results showed that general-purpose models exhibited insufficient performance, achieving a maximum accuracy of only 17.24%, and the evaluation focused primarily on qualitative judgments rather than quantitative interpretation.

Beyond thermal imagery, several studies have examined VLM performance on other types of physical visualizations. Examples include ClimaQA [29], which targets weather maps, and research evaluating the interpretation of stress distribution maps [30]. However, these studies similarly do not assess the quantitative interpretation capabilities of VLMs with respect to physical quantities encoded in visual form.

In the context of indoor environmental evaluation, accurate interpretation of physical quantities—rather than merely identifying qualitative trends—is essential for explaining factors such as thermal comfort. As noted above, however, systematic evaluations of the quantitative interpretation capabilities of VLMs for such physical information remain largely unexplored.

Moreover, research integrating MR and VLMs is still limited. Most existing studies [31,32] propose systems in which real-world images are provided as input to VLMs, and the inference results are overlaid onto the real-world view. Consequently, few studies have explicitly examined whether VLMs can correctly interpret MR images in which virtual objects are superimposed onto real-world scenes.

 doi:10.20944/preprints202601.2306.v1

Duan et al. [17] investigated the understanding of MR scenes by VLMs and reported that VLMs are capable of recognizing objects displayed virtually via MR. However, the virtual objects considered in their study were tangible entities, such as animals and furniture. To date, no studies have evaluated VLM performance on MR images that visualize inherently invisible physical phenomena—such as airflow or heat—through virtual overlays.

2.3. Applications of AI and VLMs in Built Environmental Design

In built environmental design, conducting prior CFD analyses of indoor environments and sharing the results among stakeholders is widely recognized as essential. However, non-expert stakeholders often struggle to understand these analysis results due to their technical complexity. Existing AI-based studies related to CFD analysis [33] primarily focus on accelerating simulations or enabling real-time design iterations using techniques such as convolutional neural networks (CNNs). In contrast, relatively little attention has been paid to improving the interpretability of CFD results for non-expert users.

Zhu et al. [34] proposed a system that visualizes CFD analysis results at multiple spatial scales within an MR environment. While this approach enhances spatial understanding, simply displaying physical quantities such as temperature and airflow remains insufficient for enabling non-experts to assess indoor comfort in a concrete and intuitive manner.

Separately, several studies have explored the application of VLMs to environmental evaluation tasks. For example, Zhang et al. [35] proposed a system that evaluates thermal comfort in urban spaces based on landscape images using VLMs, and reported a positive correlation between model predictions and expert assessments. Similarly, the effectiveness of VLMs for perceptual evaluation of urban safety has been demonstrated [36]. However, because these approaches rely on landscape images as input, they cannot account for environmental factors such as temperature and airflow, which are not directly observable in such imagery.

2.4. Contribution

In response to the challenges identified above, this study aims to evaluate the quantitative interpretation accuracy of general-purpose VLMs with respect to physical quantities, as well as the extent to which this accuracy can be improved through domain adaptation. To this end, we construct a dataset specifically designed to assess quantitative interpretation based on legend information in scientific visualizations.

Furthermore, this study evaluates the feasibility of integrating MR and VLMs by examining the recognition and interpretation of MR images in which CFD analysis results are superimposed onto real indoor spaces. By focusing on the visualization of inherently invisible physical phenomena through MR, this work addresses a research gap not covered by existing studies.

The outcomes of this research provide a foundational technology for supporting non-expert stakeholders, such as building occupants, in understanding CFD analysis results used during the design process. By facilitating comprehension of future indoor environments, this approach aims to promote smoother consensus building and improve satisfaction with indoor environmental design. In addition, this study extends the application scope of VLMs to the interpretation of invisible physical information, laying the groundwork for future VLM-based environmental evaluation systems.

3. Dataset Construction

3.1. Dataset Overview

In this study, to evaluate and enhance the ability of VLMs to interpret images visualizing physical quantities, we constructed a dataset consisting of MR images and associated question–

 doi:10.20944/preprints202601.2306.v1

answer pairs. In the MR images, indoor CFD analysis results are overlaid onto real-world indoor scenes.

Unlike existing heatmap datasets [16], which primarily target qualitative judgments such as identifying which object or region is hotter, the proposed dataset includes questions that require the extraction of quantitative values based on legend information. Using this dataset, we aim to verify whether VLMs can quantitatively infer physical quantities from visualized scientific data.

Each dataset entry comprises an MR image, legend information provided in either image or text form, a question, a ground-truth answer, and an explanation intended to support the reasoning process. The MR images include contour plots representing temperature distributions and isoline maps representing airflow velocity. Legend information consists of color bar images corresponding to the temperature contours and textual descriptions defining the numerical meaning of the isoline colors. An example data sample is shown in Figure 1.

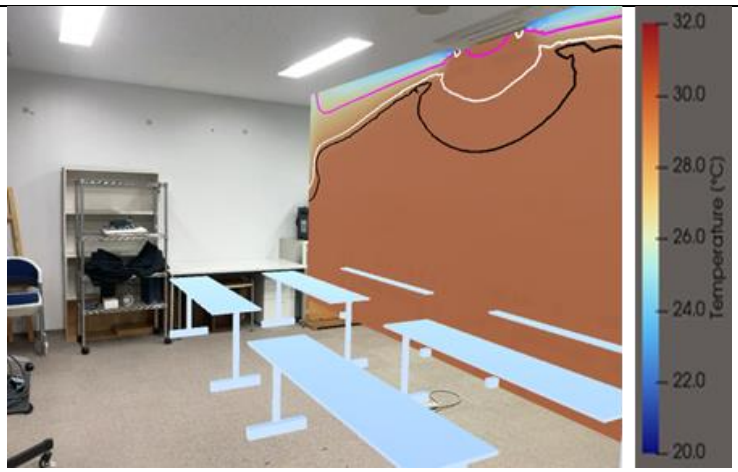
MR Image + Legend image	
Legend texts	Magenta line: Represents 1.0 m/s, indicating the discharged airflow from the air condiioner. White line: Represents 0.25 m/s, indicating the threshold for draft risk, where airflow may cause thermal discomfort to occupants. Black line: Represents 0.1 m/s, indicating the boundary of air stagnation. The area inside this line indicates stagnant air.
Question	What is the average temperature in areas where the airflow velocity is less than 0.1 m/s?
Answer (Ground Truth)	29.93
Explanation text	The area enclosed by the black line is primarily indicated by deep warm colors, such as red. By cross-referencing these colors with the color distribution in the legend image, the representative temperature of this region is identified.

Figure 1. Example of a data sample included in the constructed dataset.

The dataset was constructed through the following steps:

1. Creation of three-dimensional models of the target space and execution of CFD analyses;
2. Generation of MR images with superimposed CFD analysis results;
3. Creation of question–answer pairs and explanatory texts based on the MR images.

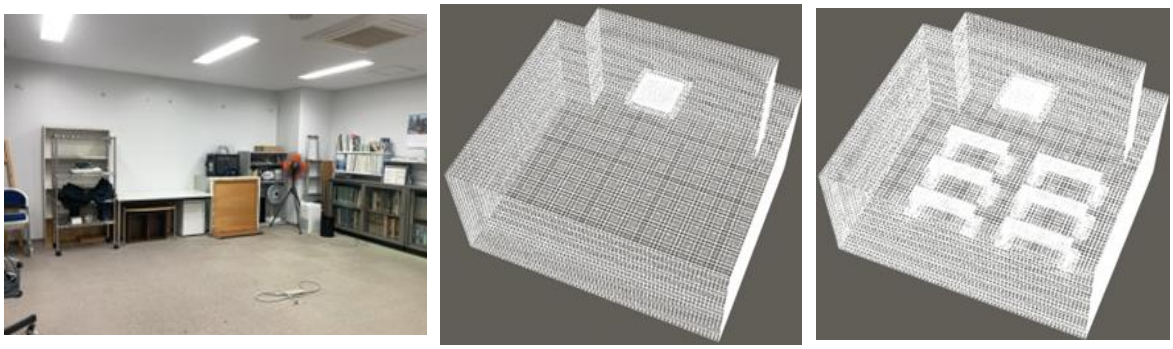
3.2. CFD Analysis

This subsection describes the CFD analysis methodology used to generate indoor environmental data for MR visualization. First, the analysis domain was defined and a three-dimensional model of the target indoor space was constructed. Mesh generation was then performed, followed by transient CFD simulations using an appropriate solver.

 doi:10.20944/preprints202601.2306.v1

Room 410 of the M3 Building at The University of Osaka (W 6.940 × D 6.555 × H 3.200 m) was selected as the target space. A three-dimensional model of the room was created using FreeCAD. To introduce variability into the dataset, two layout configurations were prepared: an empty room without furniture (Pattern A) and a furnished room equipped with desks (Pattern B).

Meshes were generated using cfmesh, with a base cell size of 100 mm. Local mesh refinement was applied in regions of geometric or physical importance: a cell size of 50 mm around the desks due to their geometric complexity, and 25 mm around the air conditioner inlets and outlets. Figure 2 shows the interior view of Room 410 and the generated meshes. Although existing objects such as shelves appear in the interior photograph, they were excluded from the CFD model; conversely, the desks in Pattern B were introduced as virtual objects.



Interior View Pattern A Pattern B

Figure 2. Interior view of Room 410 and the corresponding computational meshes for each layout pattern.

CFD simulations were conducted using the open-source software OpenFOAM. The *buoyantPimpleFoam* solver, which supports unsteady flow analysis with buoyancy effects, was employed, and the standard k–ε turbulence model was adopted. This configuration enabled the output of temperature contour plots and airflow isoline maps at arbitrary time steps.

To diversify the dataset, multiple analysis cases were created by varying the air conditioner operation mode (cooling or heating) and outlet angle (0° or 60°). The outlet angle was defined relative to the ceiling surface, with positive values indicating a downward direction. CFD simulations were conducted for all combinations of these conditions. Detailed analysis parameters are summarized in Tables 1 and 2.

Table 1. Boundary conditions applied to the air conditioner in the CFD analysis.

Item	Cooling	Heating
Initial Temperature (℃)	30	10
Outlet Temperature (℃)	22	30
Outlet Speed (m/s)	3.2	3.2
Direction (°)	0, 60	0, 60

Table 2. Boundary conditions of the construction materials used in the CFD analysis.

Item		Wall	Ceiling	Floor
Thermal Transmittance (W / (m² ・K))		1.38	1.8	0
Cooling	Initial wall temperature (℃)	30	30	30
	Outdoor temperature (℃)	35	35	-
Heating	Initial wall temperature (℃)	10	10	10
	Outdoor temperature (℃)	5	5	-

 doi:10.20944/preprints202601.2306.v1

3.3. MR Image Genetarion Method

This subsection describes the procedure used to generate MR images for the dataset. First, CFD analysis results were visualized, after which the visualized outputs were imported into a game engine and overlaid onto real-world images to create MR scenes.

CFD results obtained as described in Subsection 3.2 were visualized using ParaView. Data were extracted at selected time steps and included temperature contour plots and airflow velocity isolines on three vertical cross-sections—near the air conditioner (1), at the room center (2), and near a wall (3)—as well as one horizontal cross-section at a height of 1.1 m above the floor (4), corresponding to the occupied zone. The locations of these cross-sections are illustrated in Figure 3.

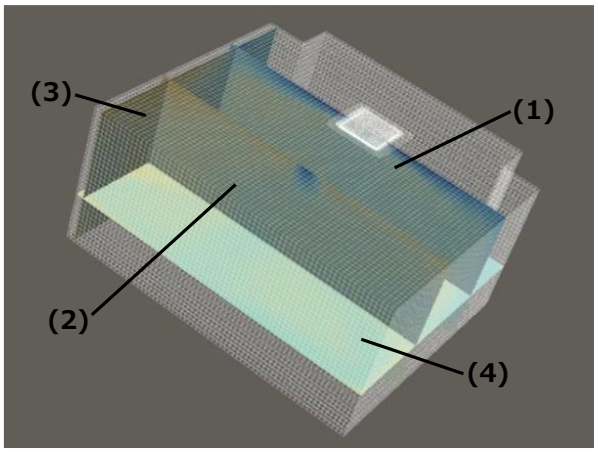


Figure 3. Locations of the extracted cross-sections used for CFD visualization.

To increase dataset diversity, both the temperature range and the colormap type were varied across analysis cases during visualization. The visualized geometries were exported in *.ply* format along with the corresponding colormap images. The colormaps used were *Fast* (a rainbow-based scheme), *Fast (Reversed)* (its inverted variant), and *Viridis* (a perceptually uniform purple–yellow scheme). These colormaps are shown in Figure 4.

(a) Fast (b) Fast (Reversed) (c) Viridis

Figure 4. Colormap types and examples of the corresponding temperature contour plots used in the dataset: (a) Fast, (b) Fast (Reversed), and (c) Viridis.

Airflow velocity was visualized using isolines rather than vector arrows. Although vector arrows are commonly used in fluid analysis, their apparent length varies with viewpoint in MR environments, making quantitative interpretation difficult. Moreover, coloring arrows by velocity would introduce color overlap with temperature contours when both are displayed simultaneously. Therefore, isolines were employed to represent airflow velocity using colored curves, facilitating clearer identification of velocity ranges.

MR image generation was performed using the game engine Unity 2022.3.3f1. Visualization data exported from ParaView in *.ply* format were converted to *.fbx* format to enable import into Unity while preserving color information. Background images, contour plots, and isoline maps were then

 doi:10.20944/preprints202601.2306.v1

integrated. A total of 18 photographs of Room 410, captured from various viewpoints using a smartphone, were prepared as background images (examples are shown in Figure 5). Spatial alignment was achieved using the Visual Positioning System provided by Immersal [37], after which CFD results were superimposed to generate MR images.

Figure 5. Examples of real-world background images used for generating MR images in the dataset.

To diversify MR presentation conditions, random transparency levels were assigned to the superimposed analysis objects. For each MR image, only a single cross-section was displayed, as simultaneous visualization of multiple sections leads to occlusion and hinders accurate color identification. Two display patterns were generated: images showing only temperature contours, and images showing both temperature contours and airflow isolines. Multiple MR images were created from a single background image by varying the superimposed cross-section and the time step of the CFD results.

3.4. Creation of Question-Answer Pairs

The question–answer pairs in the dataset were categorized into the following three types:

1. Temperature interpretation;
2. Airflow interpretation;
3. Integrated interpretation of temperature and airflow.

For the temperature interpretation category, questions were designed to evaluate the ability to extract quantitative information from contour plots in MR images. Typical examples include questions such as “What is the maximum temperature?”. Ground-truth answers were obtained directly from the numerical CFD results. In addition, explanation texts included bounding box coordinates identifying the relevant image regions, and the prompts instructed the model to output both the temperature value and the corresponding bounding box.

For airflow-related questions, the objective was to assess interpretation of airflow velocity visualized using discrete isolines. Because precise velocity values cannot be determined at positions between isolines, questions were divided into two types:

- Type 1: Questions asking for the color of isolines present in the image;
- Type 2: Questions asking for the airflow velocity range at specified coordinates.

Type 1 questions evaluate whether a VLM can correctly recognize isoline colors in MR images with real-world background noise. Type 2 questions assess whether the model can infer velocity ranges by identifying isoline colors around the specified coordinates and comparing them with legend information. Ground-truth answers for airflow-related questions were manually created.

For integrated temperature–airflow interpretation, both qualitative and quantitative questions were designed to evaluate the ability to synthesize multiple types of physical information. Qualitative questions asked about differences in temperature distributions across regions with varying airflow velocities, and ground-truth answers were manually generated based on visual inspection. Quantitative questions required identifying specific temperatures within regions defined by airflow velocity ranges. Ground-truth answers for these questions were generated using an automated script that extracts RGB values from target regions and maps them to temperature values using the color bar. Although this approach does not rely on raw CFD numerical outputs, it ensures consistency between the visual input presented to the VLM and the expected answers.

 doi:10.20944/preprints202601.2306.v1

In addition to MR images and question–answer pairs, the dataset includes legend information and explanation texts. Legend information comprises temperature color bar images and textual descriptions explaining the numerical meaning of airflow isolines. Specifically, the legend text defines the following relationships:

- Magenta line: 1.0 m/s, representing discharged airflow from the air conditioner;
- White line: 0.25 m/s, representing the draft risk threshold at which airflow may cause thermal discomfort;
- Black line: 0.1 m/s, representing the boundary of air stagnation, with enclosed regions indicating stagnant air.

Explanation texts were added to support learning of the reasoning process. Templates corresponding to each question category were defined, and explanation texts were automatically generated using a rule-based approach.

3.5. Dataset Details

The dataset constructed using the procedures described above consists of 266 MR images and 785 question–answer pairs. A detailed breakdown by category is provided in Table 3. Each data entry includes an MR image, a legend image, legend text, a question, a ground-truth answer, and an explanation.

Table 3. Category structure of the constructed dataset.

Category	Example Question	Sample
1. Temperature Interpretation	What is the highest temperature in the image?	336
2. Airflow Interpretation	What is the airflow velocity at the specified coordinates?	307
3. Integrated Interpretation of Temperature and Airflow	What is the temperature within the specified flow velocity range?	142
All	-	785

For the experiments described in subsequent sections, the dataset was divided into training, validation, and evaluation subsets with an approximate ratio of 6:2:2. Stratified sampling was employed to ensure that the distribution of question categories in each subset closely matched that of the full dataset.

4. Evaluation

This section evaluates the effectiveness of the dataset constructed in Section 3 and examines the impact of domain adaptation through fine-tuning. Specifically, the performance of the VLM on the evaluation dataset is assessed both before and after fine-tuning. In addition, the effect of fine-tuning on the model's general reasoning capabilities is investigated to identify potential trade-offs between domain-specific accuracy and generalization.

4.1. Evaluation Setup

In this study, Qwen2.5-VL-7B-Instruct [26] was employed as the baseline model. This model is known for its high recognition accuracy of fine-grained textual and geometric details in images. Because the MR images used in this study include small numerical labels within temperature color bars, high OCR accuracy is essential; therefore, this model was selected as an appropriate baseline.

Low-Rank Adaptation (LoRA) [38] was adopted for fine-tuning. LoRA enables efficient domain adaptation while significantly reducing computational cost by updating only a small subset of model parameters. The training subset of the dataset was used for model optimization, while the validation subset was used to monitor loss convergence and prevent overfitting during training. The main

 doi:10.20944/preprints202601.2306.v1

hyperparameter settings are summarized in Table 4, and the computational environment is detailed in Table 5.

Table 4. Primary hyperparameters used for fine-tuning.

Learning Rate	0.0002
Batch Size	1
Gradient Accumulation Steps	16
Optimizer	AdamW
LoRa Rank	16

Table 5. Computational environment used for fine-tuning and evaluation.

CPU	Intel Core i9-14900K
GPU	NVIDIA GeForce RTX 4090
RAM	64GB
OS	Windows 11 Education 25H2

4.2. Evaluation Methodology

This subsection describes the verification experiments conducted in this study and the evaluation metrics used to assess model performance.

4.2.1. Verification Experiments

To verify both the effectiveness of the constructed dataset and the impact of domain adaptation through fine-tuning, the following three models were compared:

- Baseline Model: Qwen2.5-VL-7B-Instruct without fine-tuning;
- Epoch–5 Model: The baseline model fine-tuned on the constructed dataset for five epochs;
- Epoch–10 Model: The baseline model fine-tuned on the constructed dataset for ten epochs.

For each model, category-wise performance was evaluated on the evaluation subset to analyze the effects of domain adaptation and performance changes across training epochs.

In addition to domain-specific performance, generalization capability was evaluated using MMBench [39], a widely used benchmark for assessing perception and reasoning in VLMs. By comparing MMBench scores before and after fine-tuning, the impact of domain adaptation on the model's inherent general reasoning ability was quantitatively assessed.

4.2.2. Evaluation Metrics

To improve the reliability of model predictions, multiple inference runs were conducted for each sample in the evaluation subset, and the final prediction was determined through statistical aggregation. Specifically, three inference runs were performed per question. For qualitative questions, the most frequently occurring answer (mode) was selected, while for quantitative questions, the median value was used to mitigate the influence of outliers.

Based on these aggregated predictions, evaluation metrics were applied according to the question type.

For qualitative questions, accuracy was computed by automatically checking for exact matches between predicted answers and ground-truth labels. Because the baseline model frequently failed to adhere to the specified output format, its qualitative responses were manually reviewed to determine correctness.

For quantitative questions, the Mean Absolute Error (MAE) was computed using Equation (1) to measure the absolute difference between predicted and ground-truth values. In addition, because the range of the temperature color bar varied across samples, accuracy was also evaluated in a relative

 doi:10.20944/preprints202601.2306.v1

manner. Specifically, predictions were considered correct if the absolute error fell within 10% of the corresponding color bar range, as defined by Equations (2) and (3).

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|\hat{x}_i - x_i| \tag{1}$$

$$f(x_i) = \begin{cases} 1, & if\ |\hat{x}_i - x_i| \le R_i \times 0.10 \\ 0, & otherwise \end{cases} \tag{2}$$

$$Accuracy = \frac{1}{N}\sum_{i=1}^{N} f(x_i) \times 100 \tag{3}$$

Here, N denotes the total number of data samples, $\hat{x}_i$ is the predicted value, x_i is the ground truth value, R_i represents the color bar range of the i-th sample.

4.3. Evaluation Results

This subsection presents the results obtained from the evaluation procedures described in Subsection 4.2. Table 6 summarizes category-wise accuracy for each model on the evaluation subset, and a comparative visualization is provided in Figure 6. For Category 1 (temperature interpretation), Figures 7(a) and 7(b) show the accuracy and MAE, respectively, across different colormap types. Figure 8 illustrates the accuracy of each model for individual question types within Category 2 (airflow interpretation).

Table 6. Category-wise accuracy for each model on the evaluation subset.

Model	1. Temperature		2. Airflow	3. Integrated	
	Accuracy [%] ↑	MAE [-] ↓	Accuracy [%] ↑	Accuracy [%] ↑	MAE [-] ↓
Baseline Model	10.4	4.00	28.8	5.00	4.14
Epoch – 5	61.2	0.92	72.7	70.0	0.87
Epoch - 10	68.7	0.93	75.4	70.0	1.01

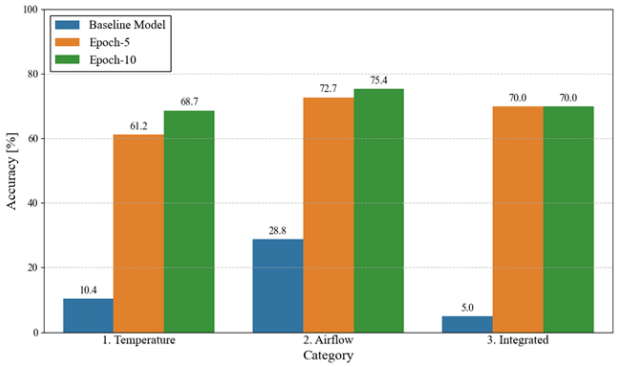
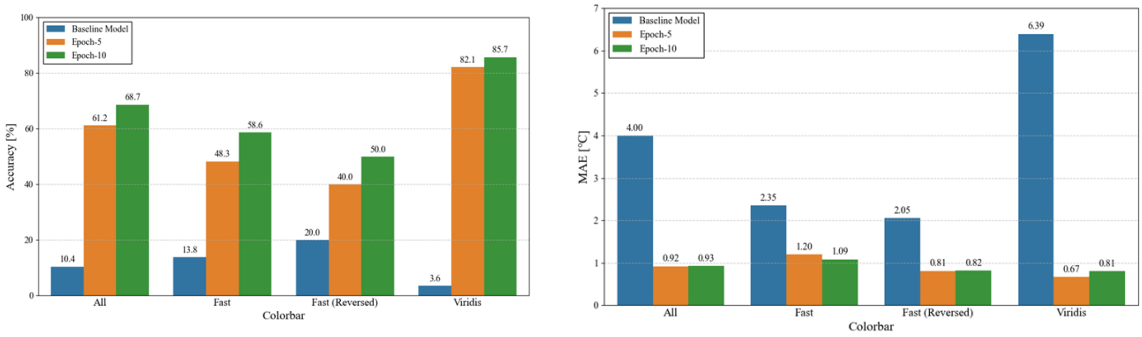


Figure 6. Category-wise accuracy for each model on the evaluation subset.

 doi:10.20944/preprints202601.2306.v1



Accuracy [%] MAE [-]

Figure 7. Accuracy and MAE for Category 1 (Temperature interpretation) across different colormap types.

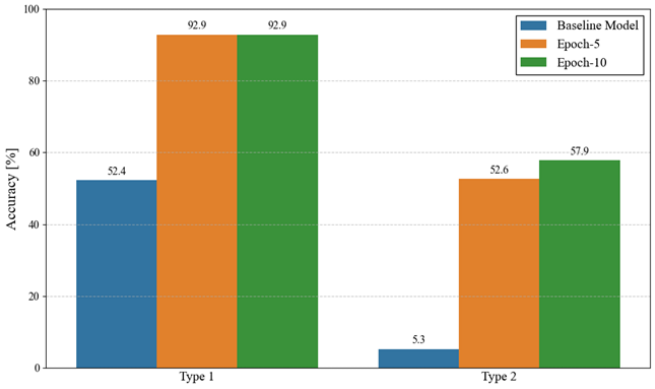
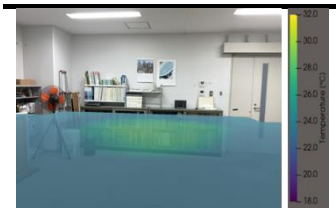


Figure 8. Accuracy of each model for Category 2 (Airflow interpretation) across different question types.

Across all categories, the fine-tuned models consistently outperformed the baseline model, demonstrating clear accuracy improvements and confirming the effectiveness of domain adaptation.

To provide a qualitative comparison between the baseline and fine-tuned models, Figure 9 presents representative samples from each category along with the corresponding model outputs. Furthermore, the evaluation of generalization performance summarized in Table 7 shows a decrease of approximately 4.4% in the MMBench score after fine-tuning, indicating a modest trade-off between domain-specific performance gains and general reasoning capability.

Input images	Question	Output of Baseline Model	Output of Epoch - 10	Ground Truth
	What is the maximum temperature?	18.0	21.29	21.55
	What is the range of airflow velocity at the specified coordinates [273, 479]?	Less than 0.1 m/s	Between 0.1 m/s and 0.25 m/s	Between 0.1 m/s and 0.25 m/s

 doi:10.20944/preprints202601.2306.v1

	What is the average temperature in areas where the airflow velocity is less than 0.1 m/s?	21.0	25.75	26.59

Figure 9. Representative data samples from each category and the corresponding outputs of the Baseline Model and the Epoch–10 fine-tuned model. If the MR image includes airflow isolines, the associated legend text described in Subsection 3.4 is also provided. Red dots indicate the coordinates referenced in the questions; these dots are shown for illustration only and are not included in the MR images input to the VLM.

Table 7. MMBench scores of each model used to evaluate general reasoning performance.

	Baseline Model	Epoch - 5	Epoch - 10
Accuracy [%] ↑	87.3	82.9	82.9

5. Discussion

5.1. Effectiveness of Domain Adaptation

As shown in Table 6, the Baseline Model exhibits low accuracy and high MAE across all categories. These results indicate that the model lacks the ability to interpret scientific visualizations representing physical quantities. In particular, the poor performance on quantitative questions suggests a failure to cross-reference color legends with contour plots in MR images to infer temperature values.

In contrast, the model fine-tuned on the constructed dataset achieved substantially higher accuracy and lower MAE across all categories. This improvement indicates that the VLM successfully acquired domain-specific reasoning and interpretation capabilities for scientific visualization. Notably, the marked performance gain in quantitative tasks demonstrates that the model learned to systematically associate legend information with visual features in MR images in order to infer numerical values.

5.2. Category-wise Analysis

5.2.1. Temperature Interpretation

As reported in Table 6, the Baseline Model achieved an accuracy of only 10% for temperature interpretation, indicating an inability to cross-reference image features with legend information. Figures 7 and 9 further show that performance was significantly degraded when the *Viridis* colormap was used, compared to *Fast*. This is likely because rainbow-based colormaps are widely used in thermography and weather visualizations, allowing the model to exploit prior knowledge learned during pretraining. In contrast, *Viridis* is less commonly encountered in general visual data, and therefore the model was unable to rely on prior associations, resulting in poor performance.

For the fine-tuned model, however, Figure 7 shows that accuracy with *Viridis* exceeded that achieved with *Fast* and *Fast (Reversed)*. This can be attributed to the fact that *Viridis* maintains a unique, monotonic correspondence between color and value, whereas *Fast* and *Fast (Reversed)* introduce ambiguity: identical colors may correspond to different numerical values depending on the legend orientation. Consequently, accurate interpretation using *Fast* requires rigorous cross-referencing between the MR image and the legend. The superior performance with *Viridis* therefore indicates that the fine-tuned model relied on explicit legend-based reasoning rather than prior assumptions, resulting in more robust temperature interpretation.

5.2.2. Airflow Interpretation

 doi:10.20944/preprints202601.2306.v1

As shown in Figure 8, the Baseline Model achieved an accuracy exceeding 50% for Type 1 questions, which ask about the colors present in the image. This suggests that the model retains basic perceptual abilities, such as recognizing isolines and their colors in MR visualizations. In contrast, for Type 2 questions—requiring inference of airflow velocity ranges at specific coordinates—the accuracy dropped to approximately 5%. This result indicates that the Baseline Model lacks the ability to interpret legend information and perform the reasoning necessary to estimate values between isolines.

The fine-tuned model, by contrast, achieved accuracies of approximately 93% for Type 1 questions and 53% for Type 2 questions. These results demonstrate that domain adaptation enabled the VLM to associate recognized visual features with legend information and convert them into physical quantities. The substantial improvement for Type 2 questions further indicates that the model acquired spatial reasoning capabilities, allowing it to identify regions between isolines and perform visual interpolation based on adjacent values.

5.2.3. Integrated Interpretation of Temperature and Airflow

As shown in Table 6, the Baseline Model achieved an accuracy of only 5.0% in Category 3. This low performance is primarily attributable to its inability to correctly infer airflow velocity ranges, as discussed in Subsubsection 5.2.2.

In contrast, the fine-tuned model achieved an accuracy of 70.0%. This result suggests that the training process enabled the model to sequentially perform multiple inference steps: first deriving airflow velocity ranges from legend information, and then inferring temperature values by cross-referencing contour plots with corresponding legends. This capability to integrate and reason over multiple physical quantities suggests that future VLM-based systems may be able to assess indoor thermal comfort using complex, multimodal information and provide accessible explanations to non-expert users.

5.3. Model Generalizarion Performance and Training Efficiency

While fine-tuning on domain-specific data improves task performance, it may also degrade a model's general reasoning capabilities. To assess this effect, we evaluated the models using MMBench. As shown in Table 7, the fine-tuned model exhibited a performance decrease of approximately 4.4%, with the score dropping from 87.3% to 82.9%. Given the substantial gains in quantitative interpretation accuracy achieved through fine-tuning, this reduction is relatively modest. The results therefore indicate that the model retains sufficient general reasoning ability despite domain adaptation.

Training efficiency was also examined. As shown in Table 6, accuracy converged rapidly, reaching a high level after only five epochs. No significant improvement was observed when training was extended to ten epochs. These results indicate that effective domain adaptation can be achieved with a small number of training iterations. This efficiency suggests that rapid fine-tuning tailored to specific indoor environments or analysis conditions will be feasible, enabling swift system adaptation in practical deployments.

5.4. Limitations and Future Work

This study has two primary limitations.

The first concerns the diversity of spatial conditions. All CFD analysis results in the constructed dataset were derived from a single room geometry. Consequently, the model's performance in spaces with different geometries or complex furniture arrangements has not been validated. Future work should therefore incorporate CFD results from a wider variety of spatial configurations to improve and evaluate the model's generalizability.

The second limitation relates to practical applicability. The primary objective of this study was to assess whether a VLM can accurately interpret the current state of an indoor environment, such as

 doi:10.20944/preprints202601.2306.v1

reading temperatures from images. However, real-world design processes require more than descriptive understanding; they demand explanations of causal mechanisms and proposals for design improvements. Future work will therefore focus on constructing datasets that include explanatory texts describing cause–effect relationships and specific improvement strategies. Training on such data would enable the VLM not only to interpret visualized CFD results, but also to provide causal explanations and actionable recommendations, thereby supporting non-experts throughout the design process.

6. Conclusions

This study investigated the application of VLMs to support non-expert understanding of CFD analysis results, with the goal of facilitating consensus building in architectural and environmental design. A fundamental prerequisite for this application is the ability of VLMs to accurately interpret visualizations of physical quantities.

To evaluate this capability, we constructed a dataset consisting of MR images with superimposed CFD analysis results and corresponding question–answer pairs. These images visualize temperature contour plots and airflow velocity isolines within real-world indoor spaces. Using this dataset, we fine-tuned a general-purpose VLM and evaluated its interpretation and reasoning performance.

The dataset enabled systematic evaluation of the model's ability to cross-reference legend information with quantitative values, as well as its capacity to interpret inherently invisible physical phenomena—such as heat and airflow—when visualized using MR.

The main findings are summarized as follows:

- The Baseline Model achieved accuracies below 30% across all categories, indicating that general-purpose VLMs lack sufficient quantitative reasoning ability for interpreting visualized physical quantities such as temperature and airflow.
- Fine-tuning the model on the constructed dataset improved accuracy to over 60% across all categories, demonstrating that domain adaptation enables effective cross-referencing between image features and legend information.
- Training convergence was achieved within five epochs, with no significant improvement observed at ten epochs, indicating that effective domain adaptation can be achieved with minimal training effort.
- The fine-tuned model enables quantitative interpretation of CFD analysis results, allowing building users to understand future indoor environments during the design phase. This capability is expected to facilitate smoother consensus building and improve satisfaction with indoor environmental quality.

Supplementary Materials: The following supporting information can be downloaded at: Preprints.org. The materials are organized into the Data Folder: Dataset, Fine-tuned Models, Inference Results and Scripts for fine-tuning and inference.

Author Contributions: Conceptualization, S.F. and T.F.; methodology, S.F.; software, S.F.; validation, S.F.; formal analysis, S.F.; investigation, S.F.; resources, S.F.; data curation, S.F.; writing—original draft preparation, S.F.; writing—review and editing, S.F. and T.F.; visualization, S.F.; supervision, T.F.; project administration, S.F.; funding acquisition, T.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by JSPS KAKENHI, grant number 23K11724.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset, fine-tuned models, and inference results are presented in this study are available in Zenodo at https://doi.org/10.5281/zenodo.18320255.

Acknowledgments: During the preparation of this study, the authors used Gemini for the purpose of improving the readability and language of the text. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Robertson, T.; Simonsen, J. Challenges and opportunities in contemporary participatory design. *Design Issues* **2012**, *28*(3), 3-9. doi: https://doi.org/10.1162/DESI_a_00157
2. Matz, C. J.; Stieb, D. M.; Davis, K.; Egyed, M.; Rose, A.; Chou, B.; Brion, O. Effects of age, season, gender and urban-rural status on time-activity: Canadian Human Activity Pattern Survey 2 (CHAPS 2). *International journal of environmental research and public health* **2014**, *11*(2), 2108-2124. doi: https://doi.org/10.3390/ijerph110202108
3. Morris, E. A.; Speroni, S.; Taylor, B. D. Going nowhere faster: Did the covid-19 pandemic accelerate the trend toward staying home?. *Journal of the American Planning Association* **2025**, *91*(3), 361-379. doi: https://doi.org/10.1080/01944363.2024.2385327
4. Al Horr, Y.; Arif, M.; Kaushik, A.; Mazroei, A.; Katafygiotou, M.; Elsarrag, E. Occupant productivity and office indoor environment quality: A review of the literature. *Building and environment* **2016**, *105*, 369-389. doi: https://doi.org/10.1016/j.buildenv.2016.06.001
5. Calderon-Hernandez, C.; Paes, D.; Irizarry, J.; Brioso, X. Comparing virtual reality and 2-dimensional drawings for the visualization of a construction project. In *ASCE International Conference on Computing in Civil Engineering* 2019, Atlanta, Georgia, 17-19 June 2019; pp. 17-24. doi: https://doi.org/10.1061/9780784482421
6. Ivić, I.; Cerić, A. Risks caused by information asymmetry in construction projects: A systematic literature review. *Sustainability* **2023**, *15*(13), 9979. doi: https://doi.org/10.3390/su15139979
7. Lange, E. Integration of computerized visual simulation and visual assessment in environmental planning. *Landscape and urban planning* **1994**, *30*(1-2), 99-112. doi: https://doi.org/10.1016/0169-2046(94)90070-1
8. Grossman, R. L. The case for cloud computing. *IT professional* **2009**, *11*(2), 23-27. doi: 10.1109/MITP.2009.40
9. Garbett, J.; Hartley, T.; Heesom, D. A multi-user collaborative BIM-AR system to support design and construction. *Automation in Construction* **2021**, *122*, 103487. doi: https://doi.org/10.1016/j.autcon.2020.103487
10. Milgram, P.; Kishino, F. A Taxonomy of Mixed Reality Visual Displays. *IEICE Transactions on Information and Systems* **1994**, Vol. E77-D, No. 12, 1321-1329.
11. Huizenga, C.; Abbaszadeh, S.; Zagreus, L.; Arens, E. A. Air quality and thermal comfort in office buildings: results of a large indoor environmental quality survey. *Healthy Buildings* **2006**, 393–397.
12. Fukuda, T.; Yokoi, K.; Yabuki, N.; Motamedi, A. An indoor thermal environment design system for renovation using augmented reality. *Journal of Computational Design and Engineering* **2019**, *6*(2), 179-188. doi: https://doi.org/10.1016/j.jcde.2018.05.007
13. Sibrel, S. C.; Rathore, R.; Lessard, L.; Schloss, K. B. The relation between color and spatial structure for interpreting colormap data visualizations. *Journal of vision* **2020**, *20*(12), 7-7. doi: https://doi.org/10.1167/jov.20.12.7
14. Zhou, X.; Liu, M.; Yurtsever, E.; Zagar, B. L.; Zimmer, W.; Cao, H.; Knoll, A. C. Vision language models in autonomous driving: A survey and outlook. *IEEE Transactions on Intelligent Vehicles* **2024**, 1-20. doi: 10.1109/TIV.2024.3402136
15. Ma, Y.; Song, Z.; Zhuang, Y.; Hao, J.; King, I. A survey on vision-language-action models for embodied ai. *arXiv* **2024**, *arXiv:2405.14093*. doi: https://doi.org/10.48550/arXiv.2405.14093
16. Moshtaghi, M.; Khajavi, S. H.; Pajarinen, J. RGB-Th-Bench: A Dense benchmark for Visual-Thermal Understanding of Vision Language Models. *arXiv* **2025**, *arXiv:2503.19654*. doi: https://doi.org/10.48550/arXiv.2503.19654
17. Duan, L.; Xiu, Y.; Gorlatova, M. Advancing the Understanding and Evaluation of AR-Generated Scenes: When Vision-Language Models Shine and Stumble. In *2025 IEEE Conference on Virtual Reality and 3D User*

Interfaces Abstracts and Workshops (VRW), Saint-maro, France, 8-12 March 2025; pp. 156-161. doi: 10.1109/VRW66409.2025.00039

18. OpenAI. Gpt-4 technical report. *arXiv* **2023,** *arXiv:2303.08774*. doi: https://doi.org/10.48550/arXiv.2303.08774
19. Li, Z.; Wu, X.; Du, H.; Liu, F.; Nghiem, H.; Shi, G. A Survey of State of the Art Large Vision Language Models: Benchmark Evaluations and Challenges. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, Tennessee, America, 11-15 Jun 2025; pp. 1587-1606. doi: https://doi.org/10.1109/CVPRW67362.2025.00147
20. Zhong, X.; Meng, X.; Li, Y.; Fricker, P.; Liang, J.; Koh, I. An agentic vision-action framework for generative 3D architectural modeling from sketches. *International Journal of Architectural Computing* **2025**, *23*(3), 679-700. doi: https://doi.org/10.1177/14780771251352950
21. Liu, H.; Li, C.; Li, Y.; Lee, Y. J. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, Seattle, America, 17-21 Jun 2024; pp. 26296-26306. doi: https://doi.org/10.1109/CVPR52733.2024.02484
22. Zhu, D.; Chen, J.; Shen, X.; Li, X.; Elhoseiny, M. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv* **2025**, *arXiv:2304.10592*. doi: https://doi.org/10.48550/arXiv.2304.10592
23. Dosovitskiy, A. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, *arXiv:2010.11929*. doi: https://doi.org/10.48550/arXiv.2010.11929
24. Methani, N.; Ganguly, P.; Khapra, M. M.; Kumar, P. Plotqa: Reasoning over scientific plots. In *Proceedings of the ieee/cvf winter conference on applications of computer vision*, Colorado, America, 1-5 March 2020; pp. 1527-1536. doi: https://doi.org/10.1109/WACV45572.2020.9093523
25. Masry, A.; Do, X. L.; Tan, J. Q.; Joty, S.; Hoque, E. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the association for computational linguistics: ACL 2022*, Dublin, Ireland, May 2022; pp. 2263-2279. doi: 10.18653/v1/2022.findings-acl.177
26. Xia, R.; Ye, H.; Yan, X.; Liu, Q.; Zhou, H.; Chen, Z.; Shi, B.; Yan, J.; Zhang, B. Chartx & chartvlm: A versatile benchmark and foundation model for complicated chart reasoning. *IEEE Transactions on Image Processing* **2025**, 34, 7436-7447. doi: 10.1109/TIP.2025.3607618
27. Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. Qwen2. 5-vl technical report. *arXiv* **2025**, *arXiv:2502.13923*. doi: https://doi.org/10.48550/arXiv.2502.13923
28. Lu, J.; Batra, D.; Parikh, D.; Lee, S. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems* **2019**, *32*.
29. Manivannan, V. V.; Jafari, Y.; Eranky, S.; Ho, S.; Yu, R.; Watson-Parris, D.; Ma, Y.; Bergen, L.; Berg-Kirkpatrick, T. ClimaQA: An Automated Evaluation Framework for Climate Question Answering Models. *arXiv* **2024**, *arXiv:2410.16701*. doi: https://doi.org/10.48550/arXiv.2410.16701
30. Lautenschlager, S. True colours or red herrings?: colour maps for finite-element analysis in palaeontological studies to enhance interpretation and accessibility. *Royal Society Open Science* **2021**, *8*(11), 211357. doi: https://doi.org/10.1098/rsos.211357
31. Xu, F.; Nguyen, T.; Du, J. Augmented reality for maintenance tasks with ChatGPT for automated text-to-action. *Journal of Construction Engineering and Management* **2024**, *150*(4), 04024015. doi: https://doi.org/10.1061/JCEMD4.COENG-14142
32. Fan, H.; Zhang, H.; Ma, C.; Wu, T.; Fuh, J. Y. H.; Li, B. Enhancing metal additive manufacturing training with the advanced vision language model: A pathway to immersive augmented reality training for non-experts. *Journal of Manufacturing Systems* **2024**, *75*, 257-269. doi: https://doi.org/10.1016/j.jmsy.2024.06.007
33. Calzolari, G.; Liu, W. Deep learning to replace, improve, or aid CFD analysis in built environment applications: A review. *Building and Environment* **2021**, *206*, 108315. doi: https://doi.org/10.1016/j.buildenv.2021.108315
34. Zhu, Y.; Fukuda, T.; Yabuki, N. Integrating animated computational fluid dynamics into mixed reality for building-renovation design. *Technologies* **2019**, *8*(1), 4. doi: https://doi.org/10.3390/technologies8010004
35. Zhang, D.; Xiong, Z.; Zhu, X. Evaluation of Thermal Comfort in Urban Commercial Space with Vision–Language-Model-Based Agent Model. *Land* **2025**, *14*(4), 786. doi: https://doi.org/10.3390/land14040786

 doi:10.20944/preprints202601.2306.v1

36. Zhang, J.; Li, Y.; Fukuda, T.; Wang, B. Urban safety perception assessments via integrating multimodal large language models with street view images. *Cities* **2025**, *165*, 106122. doi: https://doi.org/10.1016/j.cities.2025.106122
37. Immersal SDK. Available online: https://immersal.com/ (16 January 2026)
38. Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. Lora: Low-rank adaptation of large language models. *ICLR* **2022**, *1*(2), 3. doi: https://doi.org/10.48550/arXiv.2106.09685
39. Liu, Y.; Duan, H.; Zhang, Y.; Li, B.; Zhang, S.; Zhao, W.; Yuan, Y.; Wang, J.; He, C.; Liu, Z.; Chen, K.; Lin, D. Mmbench: Is your multi-modal model an all-around player?. In *European conference on computer vision*, Milano, Italy, 29 September - 4 October 2024; pp. 216-233. doi: https://doi.org/10.1007/978-3-031-72658-3_13