Joint CSCE Construction Specialty & CRC Conference 2025
*Conférence conjointe spécialisée en construction de la SCGC et CRC-2025*

Montreal, Quebec
July 28-31, 2025 / *28-31 juillet 2025*

# Large Language Model and Synthetic Dataset based Occupant Clothing Insulation Recognition

Rui Li[1], Haibo Feng[1]

[1] Department of Wood Science, Faculty of Forestry, The University of British Columbia, Vancouver, BC, Canada

**ABSTRACT:** With individuals spending over 85% of their time indoors, maintaining Thermal Comfort (TC) is vital for well-being and productivity. The Predicted Mean Vote (PMV), a key metric for assessing thermal sensation, integrates four environmental parameters and two personal factors: Clothing Insulation (CI) and Metabolic Rate (MR). While environmental factors are measurable through sensors or simulations, personal factors are often oversimplified, reducing the accuracy of TC assessments. Recent advancements in Machine Learning (ML) and sensor technologies offer innovative methods for estimating these personal factors. However, current approaches rely on datasets collected labor-intensively and privacy-invasively, limiting their scalability and quality. Moreover, despite recent impressive achievements across many domains of Large Language Models (LLMs), the potential of zero-shot LLMs for personal factor recognition remains unexplored. To address these limitations, this study introduces a high-quality synthetic dataset developed using Unreal Engine (UE). The dataset includes 2,250 monocular RGB videos paired with expert-extracted keyframe images, featuring single occupant engaged in diverse activities. Each video-image pair is meticulously annotated with clothing types and corresponding overall CI values, adhering to ASHRAE standards. Leveraging this dataset, a state-of-the-art (SOTA) LLM is employed for zero-shot CI recognition, achieving an overall success rate of 75.6% in clothing recognition and a Mean Absolute Error (MAE) of 0.03 in CI predictions. These results validate the dataset's quality and demonstrate the potential of LLMs in capturing occupant-specific personal factors. This research highlights a pathway for more precise and scalable TC evaluations, paving the way for smarter and more occupant-centric indoor environments.

## 1. INTRODUCTION

Given that people now spend approximately 85–90% of their time within the built environment (Jenkins et al. 1992; Klepeis et al. 2001), improving Indoor Environmental Quality (IEQ) is crucial for ensuring occupant health and productivity (De Giuli et al. 2012). Among IEQ parameters, Thermal Comfort (TC) is widely recognized as the most significant and easily defined factor (Arif et al. 2016). However, the majority of existing Heating, Ventilation, and Air Conditioning (HVAC) systems in buildings primarily focus on energy efficiency, treating occupant comfort as a secondary constraint (Li and Zou 2025; Zou et al. 2020). This approach often results in occupant dissatisfaction, prompting manual overrides that inadvertently increase energy consumption (Gunay et al. 2021). To address these challenges, Occupant Centric Control (OCC) has gained significant attention in recent years for creating more adaptive and satisfying indoor environments (Soleimanijavid et al. 2024). To achieve OCC, precise modeling of TC to serve as the objective function for control systems is needed.

TC models are generally categorized into two types: index-based models and data-driven personalized models. The most widely used index-based model is the Predicted Mean Vote (PMV), developed by Fanger

(Fanger 1970). PMV estimates the average TC ratings on a standardized seven-point scale for a large group of individuals. It is derived by integrating physiological principles, such as human heat balance, with empirical data from thermal sensation studies, resulting in a robust mathematical model. Due to its reliability and standardization, PMV has been widely adopted by industry standards such as ASHRAE 55 (De Dear and Brager 2002) and ISO 7730 (ISO 2005). In contrast, data-driven personalized models have emerged in recent years, driven by advancements in Machine Learning (ML) and sensor technologies. These models utilize diverse inputs, such as heart rate (Zhu et al. 2018), wrist and facial temperature (Sim et al. 2016; Tian et al. 2023), to predict individual TC. However, these models face significant challenges: they lack a solid mathematical and experimental foundation, are often case-specific based on small-scale datasets, making them less representative. Additionally, the finer granularity of personalized TC models often mismatches the centralized control systems in most existing HVAC designs (Aryal and Becerik-Gerber 2020), limiting their practical applicability in real-world scenarios. Given these limitations, PMV remains the preferred model and is widely applied in industry.

The PMV model is determined by six key factors: four environmental factors—air temperature, mean radiant temperature, air velocity, and relative humidity—and two personal factors—Metabolic Rate (MR) and Clothing Insulation (CI) (Fanger 1970). While environmental factors are relatively easy to measure using real-world sensors or simulations such as Computational Fluid Dynamics (Cui et al. 2013; Naboni et al. 2017), accurately capturing personal factors presents a significant challenge. As a result, many studies simplify their analysis by assuming constant values for personal factors (Liang and Du 2005; Zou et al. 2020). Recent advancements in Computer Vision (CV) and biosensing technologies have introduced innovative approaches to address these challenges. Methods such as thermal or RGB cameras (Lee et al. 2016; Na et al. 2019) and smartphone-connected wristbands (Cvetković et al. 2018) have demonstrated promising results in estimating personal factors. However, data collection remains a critical bottleneck, as it is often labor-intensive, privacy-invasive, and conducted over short time periods. This leads to datasets that are small, low-quality, and unrepresentative of diverse real-world conditions. Moreover, most studies rely on collected data to train task-specific models, making it worse for scalability and generalization (Liu et al. 2022; Choi et al. 2022). Given the transformative capabilities of Large Language Models (LLMs) across various fields (OpenAI 2023; Team et al. 2024), a compelling question arises: Can zero-shot LLMs effectively address personal factor recognition tasks?

Therefore, to address these gaps and explore the posed question, we designed a novel pipeline that spans synthetic dataset generation to LLM-based CI recognition. Leveraging Unreal Engine (UE), a simulated multi-modal synthetic dataset comprising 2,250 single-occupant video clips paired with expert-extracted keyframes was formed. The dataset features diverse occupants with varying body shapes and skin tones, performing a range of actions while wearing clothing types specified in the ASHRAE clothing standard (De Dear and Brager 2002). Based on the standard, CI values for all video-image pairs were calculated and annotated by experts, establishing the dataset as a robust benchmark for CI estimation tasks. With this dataset, we implemented CI prediction using a SOTA LLM. This approach serves a dual purpose: validating the representativeness and quality of the dataset and assessing the effectiveness of zero-shot LLMs in performing the task. The proposed pipeline demonstrates the potential for advancing scalable and generalizable solutions in the field of TC modeling, paving the way for more advanced OCC in the future.

## 2. RELATED WORKS

The most commonly used method for determining CI involves using a thermal manikin, as outlined by authoritative standards from the ASHRAE and the ISO (De Dear and Brager 2002; ISO 2005). In this setup, CI values for standard clothing ensembles are derived empirically: a life-sized manikin equipped with sensors is dressed in the ensemble, and the heat loss required to maintain a stable temperature is recorded. Although thermal manikins provide highly accurate measurements, their high cost and operational complexity generally limit their use to controlled lab environments for establishing ground truth measurements, rather than widespread application in real-world building assessments.

Consequently, there is a need for more accessible and realistic solutions. Some methods rely on environmental conditions; for instance, some researchers developed models to predict CI using outdoor air

and indoor operative temperatures as inputs (Schiavon and Lee 2013; De Carli et al 2007). Other approaches use sensor-based measurements: Lee et al. (2016) and Lee et al. (2020) proposed methods for real-time clothing insulation estimation via infrared cameras, while Choi et al. (2021) employed deep learning techniques on RGB imagery for clothing recognition and insulation evaluation. Wearable devices were also explored by (Lee et al. 2019). where a data-driven model was developed to predict CI using skin and clothing surface temperatures.

Despite recent progress, existing datasets suffer from several limitations, including a lack of representativeness and labor-intensive, privacy-intrusive data collection processes. These datasets often involve a small number of participants, restrict clothing to a few fixed combinations, and require subjects to remain stationary during experiments for accurate camera capture—conditions that are far removed from real-world scenarios (Lee et al. 2020; Choi et al. 2021). Additionally, poor lighting conditions in real-world environments can further degrade data quality (Jung et al. 2025). This highlights a pressing need for a high-quality dataset that is both representative and capable of addressing these challenges without compromising privacy or requiring labor-intensive methods. Furthermore, to the best of our knowledge, no study has yet utilized LLMs for CI recognition; instead, existing approaches rely on training task-specific models on limited datasets, which significantly hampers scalability.

## 3. METHODOLOGY

This study presents a comprehensive framework for CI recognition using a synthetic dataset generated from UE-based video clips and state-of-the-art (SOTA) LLM. As shown in Figure 1, the left section illustrates the dataset creation process, including expert annotations that quantify the *clo* values (i.e., a unit of measurement that indicates how much thermal insulation a garment provides), as well as the extraction of keyframes for training. The annotations provide detailed descriptions of clothing combinations and their corresponding *clo* values based on ASHRAE Standard 55 (De Dear and Brager 2002). The right section outlines the CI recognition pipeline, which leverages a Transformer-based SOTA LLM for processing expert prompts and keyframe inputs. The validation process includes both binary classification (right/wrong) and quantitative error assessment using MAE and Root Mean Squared Error (RMSE). The following sections will thoroughly examine each part in detail.

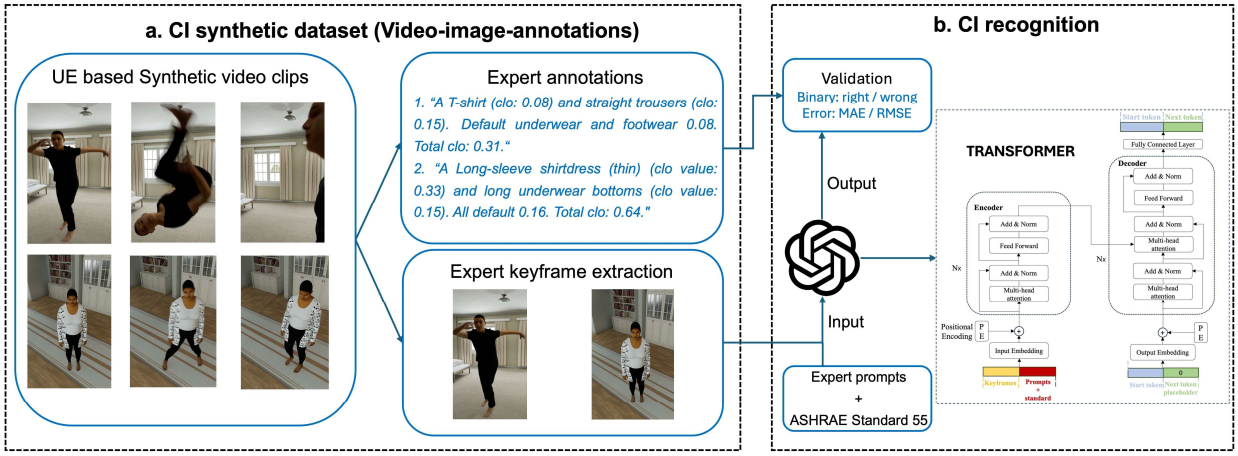

Figure 1: Overall CI synthetic dataset generation and CI recognition framework.

### 3.1 Synthetic Dataset Generation

A synthetic dataset offers several advantages over real-world datasets, including ensuring diverse training data, easy adaptability to new cameras and scenes, and cost-effective scalability. However, existing synthetic datasets often lack realism and diversity, limiting their effectiveness. To address this, Black et al. (2023) introduced BEDLAM, a high-quality synthetic dataset rendered in UE and designed to enhance 3D human pose and shape (HPS) estimation. Their findings demonstrated that, with sufficiently realistic training

data, even basic methods like HMR (Kanazawa et al., 2018) can achieve comparable performance to SOTA models such as CLIFF (Li et al., 2022). This success motivates us to further explore BEDLAM for applications beyond HPS, such as CI recognition.

BEDLAM consists of monocular RGB videos generated using human models with 271 diverse body shapes (109 men and 162 women) and 100 unique skin textures covering a broad range of skin tones. To enhance realism, it incorporates 27 distinct hairstyles and 111 clothing outfits with 1691 artist-designed textures, allowing for extensive visual variation. The dressed human bodies are then animated using 2,311 diverse motions, set in a variety of environments with different camera angles and movements. The full UE-generated dataset comprises 10K motion clips, totaling 380K RGB frames. For this study, we specifically focus on single-person CI recognition and selected 2,250 video clips featuring a single occupant. Of these, 250 clips include zoom-in camera views, introducing an additional challenge for recognition and serving as a comparative study.

To label CI, we referenced ASHRAE Standard 55 (De Dear and Brager, 2002), specifically Table 5.2.2.2B, which provides garment descriptions and corresponding clo values. The dataset includes eight clothing categories, with representative examples and average *clo* values listed in Table 1. Since underwear is not visible in synthetic models, we assigned default values: 0.04 *clo* for men's briefs and women's bras and panties. For upper-body garments, we set a T-shirt (0.08 *clo*) as the baseline for outfits covering the torso. Due to a technical limitation in BEDLAM, shoes cannot be simulated, so we assume all occupants wear socks and shoes, contributing 0.04 *clo* to total insulation. As shown in Figure 1a, during annotation, experts assigned clothing types and corresponding clo values while carefully selecting keyframes that clearly expose the occupants' clothing. Since some sampled motions involve intense activity or partial occlusions, ensuring garment visibility in the keyframes is crucial for accurate model inference (e.g., in Fig. 1a, a side flip may result in motion blur or the occupant being absent in the second and third frames, making recognition more challenging). Additionally, using images instead of full video sequences as input can enhance LLM inference speed while improving compatibility with further multimodal research applications.

Table 1: Garment insulation

| Garment category | Examples | Type counts | Average clo |
|---|---|---|---|
| Underwear | Panties, T-shirt | 8 | 0.10 |
| Footwear | Shoes, socks | 8 | 0.03 |
| Shirts and Blouses | Flannel shirt, scoop-neck blouse | 6 | 0.24 |
| Trousers and Coveralls | Walking shorts, sweatpants | 7 | 0.23 |
| Dress and Skirts | Skirt, shirtdress | 7 | 0.28 |
| Sweaters | Sleeveless vest | 4 | 0.24 |
| Suit jackets and Vests | Single-breasted jacket | 6 | 0.33 |
| Sleepwear and Robes | Short-sleeve pajamas | 9 | 0.41 |

### 3.2 Large Language Model based Clothing Insulation Recognition

With a well-labeled dataset, we proceed to implement CI recognition using LLMs. As shown in Figure 1b, built on Transformer architectures (Vaswani et al. 2017), LLMs excel at processing structured input data and making contextual inferences. Their self-attention mechanisms enable them to analyze relationships between different tokens, making them effective for zero-shot inference, where no additional fine-tuning is required. By leveraging LLMs, we aim to automate CI recognition by interpreting textual descriptions of clothing and calculating corresponding clo values.

To perform zero-shot CI recognition, we provide the LLM with two key inputs: (1) Table 5.2.2.2B mentioned in Section 3.1, which lists all garment insulation values, and (2) an expert-designed prompt that instructs the model to infer the clothing worn by an individual and compute the total clo value. Specifically, the prompt is structured as follows: *"Based on the garment insulation values in the table provided, can you infer what clothing the person is wearing (excluding footwear)? Please provide the clo value for each clothing item and calculate the total insulation value."*

The LLM generates a response in a structured format, detailing the recognized garments and their associated clo values. A typical example of an LLM response is:

- *"Top: The individual is wearing a short-sleeve T-shirt (clo value: 0.08) underneath a single-breasted jacket (thin) (clo value: 0.36).*
- *Bottom: The individual is wearing straight trousers (thin) (clo value: 0.15).*
- *Total Clo Value (Estimated): 0.59."*

Once the LLM has processed the entire dataset, we evaluate its performance against ground truth labels. As outlined in Section 3.1, we apply compensatory adjustments by adding 0.04 *clo* for both footwear and underwear, ensuring baseline consistency across all predictions. Additionally, if an expert label includes a T-shirt but the LLM fails to recognize it due to outfit occlusions, we supplement the prediction with an additional 0.08 *clo* to accurately reflect the expected insulation value. To validate the model's accuracy, we employ three key evaluation metrics. First, we use a binary classification approach, where 1 indicates a correct prediction and 0 indicates an incorrect one (i.e., success rate), providing an intuitive assessment of the model's correctness. Second, we compute the MAE, which measures the average deviation between predicted and ground truth clo values, offering insight into prediction precision. Lastly, we calculate the RMSE, which penalizes larger errors more heavily, helping to identify extremely wrong predictions. The equations of both MAE and RMSE are shown in Equation 1 and Equation 2, where $y_i$ represents the ground truth clo value, $\hat{y}_i$ denotes the predicted clo value, and $n$ is the total number of samples in the dataset. By combining all the metrics, we ensure that both qualitative correctness and quantitative precision are considered, achieving a comprehensive evaluation of the LLM's performance in CI recognition.

[1] $MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$

[2] $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$

## 4. RESULTS AND DISCUSSION

In our research, we leveraged GPT-4o, one of the SOTA LLMs, for zero-shot CI recognition. GPT-4o is particularly suited for this task due to its strong reasoning capabilities, contextual understanding, and ability to generalize across diverse datasets (Hurst et al. 2024). Key parameters of the model, such as token limit and inference speed, are detailed in Table 2.

Table 2: GPT-4o Key Parameters

| Parameter | Value | Description |
|---|---|---|
| Architecture | Transformer-based | Utilizes multi-head attention for contextual understanding. |
| Context Length | 128,000 tokens | Maximum tokens that can be processed in a single query. |
| Inference Speed | ~250-500 ms per response | Average response time per query. |
| Training Data Cutoff | 2024 | Last year of data used for model training. |
| Zero-shot Capability | Yes | Performs tasks without additional fine-tuning. |
| Fine-tuning Support | No | Not customizable through fine-tuning. |
| Temperature | 0.7 (Default) | Controls randomness in responses. |
| Multimodal Processing | Yes (Text, Image) | Capable of processing both textual and visual inputs. |

To evaluate GPT-4o's performance, we conducted quantitative validation using three above-mentioned key metrics. The results are summarized in Table 3. It can be observed that GPT-4o achieves an overall success rate of 75.6%, correctly predicting the clo category in 1,702 out of 2,250 cases. The MAE of 0.0298

suggests that, on average, the predicted clo values deviate by 0.0298 from the ground truth, while the RMSE of 0.0754 indicates that larger errors are relatively infrequent but still present. Overall, GPT-4o demonstrates strong performance in recognizing CI. Notably, considering the smallest average clo value in Table 1 (i.e., footwear at 0.03), the model's error is remarkably low, meaning that, on average, it mispredicts by less than the clo value of a single footwear item.

Table 3: Results

| Binary | MAE | RMSE | Upper/Lower/Full body |
|---|---|---|---|
| 1702/2250 = 75.6% | 0.0298 | 0.0754 | 337/55/156 |

Furthermore, we analyzed the pattern in error distribution: predictions for upper-body garments (e.g., shirts and sweaters) exhibit a higher failure rate compared to lower-body garments (e.g., trousers, and skirts). This discrepancy can be attributed to the greater variability in upper-body clothing, which includes a wider range of garment types (seen in Table 1), and certain clothing categories with similar clo values (e.g., short-sleeve knit sport shirts vs. short-sleeve dress shirts) also introduce ambiguity. Additionally, the complexity of layering, diverse styles and textures further complicates recognition. In contrast, lower-body garments tend to have more uniform textures, standardized fits, and fewer layering variations, contributing to their higher recognition accuracy. These findings underscore the need for enhanced recognition techniques capable of distinguishing subtle differences among clothes.

We also examined the model's performance on 250 zoomed-in clips and found that, with expert-selected keyframes, the recognition success rate is 76.8% (192/250), surpassing the overall accuracy despite the increased difficulty. We attribute this to the high quality of expert-curated keyframes, which emphasize critical visual details. This also highlights the importance of high-quality input data in achieving accurate CI recognition.

To the best of our knowledge, our approach is the first to implement zero-shot LLMs for CV-based CI recognition. By leveraging the advantages of synthetic datasets, we introduce a new multimodal CI dataset that encompasses the most diverse range of clothing types (i.e., 111 unique outfits) compared to previous studies, facilitating future benchmarking efforts. Moreover, our framework is highly adaptable—the synthetic dataset generation process can be easily replicated using a custom UE-based platform, whilst alternative recognition models (e.g., Vision Transformers) are also allowed for testing to explore potential improvements in both accuracy and computational efficiency.

## 5. CONCLUSIONS

To pave the way for more advanced personal factor modeling and OCC systems, this study demonstrates the potential of zero-shot LLMs for CI recognition in TC assessments. By introducing a high-quality synthetic dataset generated using UE, we provide a scalable and privacy-preserving alternative to conventional data collection methods. The dataset, consisting of 2,250 monocular RGB videos with expert-annotated CI values, enables robust model evaluation and benchmarking. Our model inference results show that a SOTA LLM achieved a 75.6% success rate in clothing recognition and an MAE of 0.03 in CI predictions, validating both the dataset's quality and the feasibility of leveraging LLMs for occupant-specific personal factor estimation. Additionally, the findings highlight key challenges, such as the higher failure rate in upper-body garment recognition due to variability in clothing styles, textures, and layering complexities.

Beyond demonstrating the effectiveness of LLMs in CI recognition, this study underscores the broader implications for TC assessments. The proposed approach enables more precise, scalable, and non-intrusive methods for evaluating occupant-specific factors, contributing to the development of smarter, more adaptive indoor environments. Future work will explore refining recognition accuracy by increasing dataset diversity and exploring alternative models that are more lightweight and fine-tunable. Improving the precision of CI recognition will contribute to more accurate occupant TC evaluations, ultimately supporting the development of advanced OCC systems that deliver improved IEQ while optimizing energy use.

**ACKNOWLEDGMENTS**

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC).

**REFERENCES**

Arif, M., Katafygiotou, M., Mazroei, A., Kaushik, A., & Elsarrag, E. (2016). Impact of indoor environmental quality on occupant well-being and comfort: A review of the literature. *International Journal of Sustainable Built Environment*, *5*(1), 1-11.

Aryal, A., & Becerik-Gerber, B. (2020). Thermal comfort modeling when personalized comfort systems are in use: Comparison of sensing and learning methods. *Building and Environment*, *185*, 107316.

Black, M. J., Patel, P., Tesch, J., & Yang, J. (2023). Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 8726-8737).

Choi, E. J., Park, B. R., Kim, N. H., & Moon, J. W. (2022). Effects of thermal comfort-driven control based on real-time clothing insulation estimated using an image-processing model. *Building and Environment*, *223*, 109438.

Choi, H., Na, H., Kim, T., & Kim, T. (2021). Vision-based estimation of clothing insulation for building control: A case study of residential buildings. *Building and Environment*, *202*, 108036.

Cui, W., Cao, G., Park, J. H., Ouyang, Q., & Zhu, Y. (2013). Influence of indoor air temperature on human thermal comfort, motivation and performance. *Building and environment*, *68*, 114-122.

Cvetković, B., Szeklicki, R., Janko, V., Lutomski, P., & Luštrek, M. (2018). Real-time activity monitoring with a wristband and a smartphone. *Information Fusion*, *43*, 77-93.

De Carli, M., Olesen, B. W., Zarrella, A., & Zecchin, R. (2007). People's clothing behaviour according to external weather and indoor environment. *Building and Environment*, *42*(12), 3965-3973.

De Dear, R. J., & Brager, G. S. (2002). Thermal comfort in naturally ventilated buildings: revisions to ASHRAE Standard 55. *Energy and buildings*, *34*(6), 549-561.

De Giuli, V., Da Pos, O., & De Carli, M. (2012). Indoor environmental quality and pupil perception in Italian primary schools. *Building and Environment*, *56*, 335-345.

Fanger, P. O. (1970). Thermal Comfort: Analysis and Applications in Environmental Engineering.

Gunay, B., Nagy, Z., Miller, C., Ouf, M., Dong, B. (2021). Using Occupant-Centric control for commercial HVAC systems. *ASHRAE Journal*, *63*(5), 30-40.

Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., ... & Kivlichan, I. (2024). Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

International Organization for Standardization. (2005). *Ergonomics of the thermal environment: Analytical determination and interpretation of thermal comfort using calculation of the PMV and PPD indices and local thermal comfort criteria*. International Organization for Standardization.

Jenkins, P. L., Phillips, T. J., Mulberg, E. J., & Hui, S. P. (1992). Activity patterns of Californians: use of and proximity to indoor pollutant sources. *Atmospheric Environment. Part A. General Topics*, *26*(12), 2141-2148.

Jung, S., Jeoung, J., Kong, M., & Hong, T. (2025). Occupant activities and clothes detection based on semi-supervised learning for occupant-centric thermal control. *Building and Environment*, *267*, 112178.

Kanazawa, A., Black, M. J., Jacobs, D. W., & Malik, J. (2018). End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7122-7131).

Klepeis, N. E., Nelson, W. C., Ott, W. R., Robinson, J. P., Tsang, A. M., Switzer, P., ... & Engelmann, W. H. (2001). The National Human Activity Pattern Survey (NHAPS): a resource for assessing exposure to environmental pollutants. *Journal of exposure science & environmental epidemiology*, *11*(3), 231-252.

Lee, J. H., Kim, Y. K., Kim, K. S., & Kim, S. (2016). Estimating clothing thermal insulation using an infrared camera. *Sensors*, *16*(3), 341.

Lee, K., Choi, H., Choi, J. H., & Kim, T. (2019). Development of a data-driven predictive model of clothing thermal insulation estimation by using advanced computational approaches. *Sustainability*, *11*(20), 5702.

Lee, K., Choi, H., Kim, H., Kim, D. D., & Kim, T. (2020). Assessment of a real-time prediction method for high clothing thermal insulation using a thermoregulation model and an infrared camera. *Atmosphere*, *11*(1), 106.

Liang, J., & Du, R. (2005, August). Thermal comfort control based on neural network for HVAC application. In *Proceedings of 2005 IEEE Conference on Control Applications, 2005. CCA 2005.* (pp. 819-824). IEEE.

Li, R., & Zou, Z. (2025). How far back shall we peer? Optimal air handling unit control leveraging extensive past observations. *Building and Environment*, *269*, 112347.

Liu, J., Foged, I. W., & Moeslund, T. B. (2022). Automatic estimation of clothing insulation rate and metabolic rate for dynamic thermal comfort assessment. *Pattern Analysis and Applications*, *25*(3), 619-634.

Li, Z., Liu, J., Zhang, Z., Xu, S., & Yan, Y. (2022, October). Cliff: Carrying location information in full frames into human pose and shape estimation. In *European Conference on Computer Vision* (pp. 590-606). Cham: Springer Nature Switzerland.

Naboni, E., Lee, D. S. H., & Fabbri, K. (2017). Thermal comfort-CFD maps for architectural interior design. *Procedia engineering*, *180*, 110-117.

Na, H., Choi, J. H., Kim, H., & Kim, T. (2019). Development of a human metabolic rate prediction model based on the use of Kinect-camera generated visual data-driven approaches. *Building and Environment*, *160*, 106216.

OpenAI, R. (2023). GPT-4 technical report. *ArXiv*, *2303*, 08774.

Schiavon, S., & Lee, K. H. (2013). Dynamic predictive clothing insulation models based on outdoor air and indoor operative temperatures. *Building and Environment*, *59*, 250-260.

Sim, S. Y., Koh, M. J., Joo, K. M., Noh, S., Park, S., Kim, Y. H., & Park, K. S. (2016). Estimation of thermal sensation based on wrist skin temperatures. *Sensors*, *16*(4), 420.

Soleimanijavid, A., Konstantzos, I., & Liu, X. (2024). Challenges and opportunities of occupant-centric building controls in real-world implementation: A critical review. *Energy and Buildings*, 113958.

Team, G., Georgiev, P., Lei, V. I., Burnell, R., Bai, L., Gulati, A., ... & Batsaikhan, B. O. (2024). Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Tian, X., Shi, L., Wang, Z., & Liu, W. (2023). A thermal comfort evaluation model based on facial skin temperature. Building and Environment, 235, 110244.

Vaswani, A. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.

Zhu, H., Wang, H., Liu, Z., Li, D., Kou, G., & Li, C. (2018). Experimental study on the human thermal comfort based on the heart rate variability (HRV) analysis under different environments. *Science of the Total Environment*, *616*, 1124-1133.

Zou, Z., Yu, X., & Ergan, S. (2020). Towards optimal control of air handling units using deep reinforcement learning and recurrent neural network. *Building and Environment*, *168*, 106535.