# Experiential Views: Towards Human Experience Evaluation of Designed Spaces using Vision-Language Models


Bon Adriel Aseniero
bon.aseniero@autodesk.com
Autodesk Research
Toronto, Ontario, Canada

Michael Lee
michael.lee@autodesk.com
Autodesk Research
Toronto, Ontario, Canada

Yi Wang
yi.wang@autodesk.com
Autodesk Research
San Francisco, California, USA

Qian Zhou
qian.zhou@autodesk.com
Autodesk Research
Toronto, Ontario, Canada

Nastaran Shahmansouri
nastaran.shahmansouri@autodesk.com
Autodesk Research
Toronto, Ontario, Canada

Rhys Goldstein
rhys.goldstein@autodesk.com
Autodesk Research
Toronto, Ontario, Canada

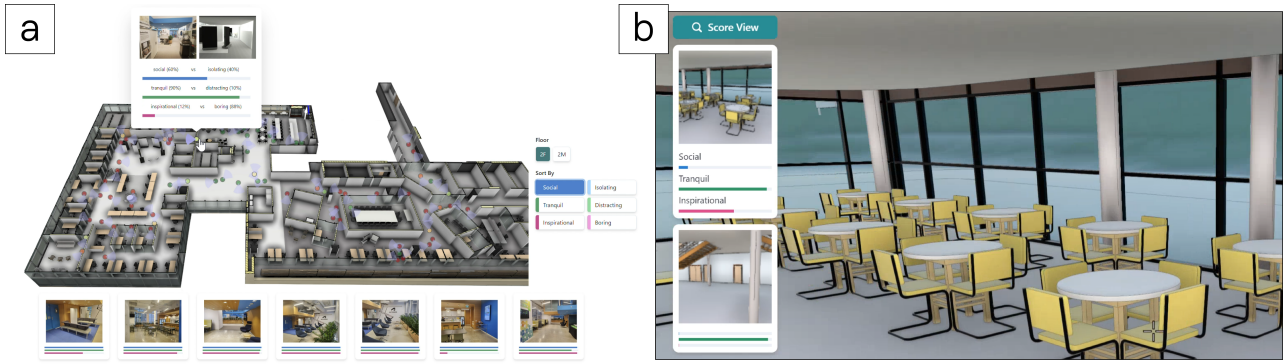




**Figure 1: The Experiential Views user interface with (a) a floor plan visualization and (b) its integration in a WebGL-based 3D-viewer with a dynamic view of the building.**


## ABSTRACT

Experiential Views is a proof-of-concept in which we explore a method of helping architects and designers predict how building occupants might experience their designed spaces using AI technology based on Vision-Language Models. Our prototype evaluates a space using a pre-trained model that we fine-tuned with photos and renders of a building. These images were evaluated and labeled based on a preliminary set of three human-centric dimensions that characterize the Social, Tranquil, and Inspirational qualities of a scene. We developed a floor plan visualization and a WebGL-based 3D-viewer that demonstrate how architectural design software could be enhanced to evaluate areas of a built environment based on psychological or emotional criteria. We see this as an early step towards helping designers anticipate emotional responses to their designs to create better experiences for occupants.



## CCS CONCEPTS

• **Human-centered computing** → **Interactive systems and tools**; *Visualization*; • **Computing methodologies** → *Artificial intelligence.*

## KEYWORDS

vision-language models, architectural design, human-centric building design

**ACM Reference Format:**
Bon Adriel Aseniero, Michael Lee, Yi Wang, Qian Zhou, Nastaran Shahmansouri, and Rhys Goldstein. 2024. Experiential Views: Towards Human Experience Evaluation of Designed Spaces using Vision-Language Models. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3613905.3650815




*CHI EA '24, May 11–16, 2024, Honolulu, HI, USA*



ACM ISBN 979-8-4007-0331-7/24/05
https://doi.org/10.1145/3613905.3650815


## 1 INTRODUCTION

Recent developments in artificial intelligence (AI), including large-language models (LLMs) and vision-language models (VLMs), have opened multiple opportunities for developing new paradigms in design computing. An industry we anticipate to benefit from this disruption is architecture and building design. Creative professionals whose work involves building design such as architects and

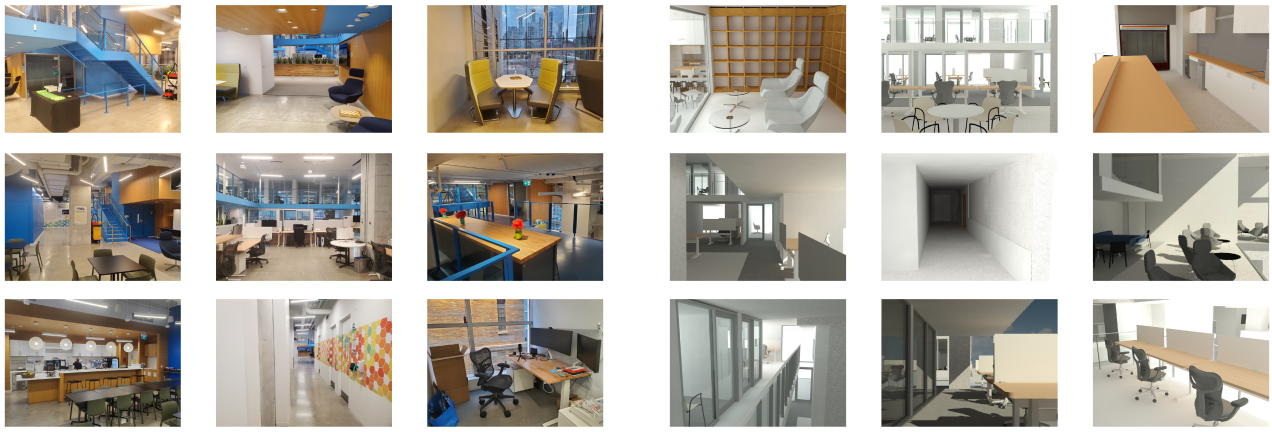


**Figure 2: Sample scenes from our dataset composed of photos of the space and 3D renderings.**

architectural designers[1] use Computer-Aided Design (CAD) software like AutoCAD [1] or Building Information Modeling (BIM) tools like Revit [2] to render designed spaces and get a sense of their look-and-feel. Extracting helpful insights about these spaces is challenging, especially insights about how a space is or will be experienced by its occupants. While some tools provide simulations of the physical environment (e.g., lighting and airflow) that designers can use to gain insights into how occupants might experience a space, there exists a gap regarding possible psychological or emotional responses to a space [13, 20].

Human-computer interaction has a long history of modeling human emotional responses [3] as part of computing systems such as affective computing research [7]. This has been expanding to include emergent research regarding emotional responses [9] combined with LLMs and other AI techniques. For instance, Andres et al.'s "System of a Sound" looked into the relationships between human activity, the built environment, and the surrounding natural environment using an LLM-based engine that interprets emotions [5]. Other work has modeled users' scene perception using graphs [19] and statistics [18]. Thus, we believe that there is potential for AI and machine-learning techniques to help incorporate the psychological or emotional responses of people in the design of built environments.

As a preliminary exploration, we developed a proof-of-concept application—*Experiential Views*. Similar to previous work using VLMs for visual sentiment analysis [8, 11, 16], Experiential Views leverages a pre-trained VLM that we fine-tuned with our own labeled data. Our aim is to help designers predict how people might experience the spaces that they design based on a preliminary set of human-centric criteria (*dimensions*). Photos or renders of a space (*scenes*) are evaluated and given a score by the application. We developed two main methods to interact with Experiential Views: (1) floor plan visualization – This interface shows the building floor plan where designers can see select evaluated scenes visualized (Figure 1a). (2) 3D viewer - This interface integrates Experiential Views into an existing WebGL-based 3D model viewer in which designers can dynamically evaluate their building design (Figure 1b).

To our knowledge, there is currently a lack of tools or methods that leverage VLMs to help designers assess possible emotional responses to their designed spaces. Thus, we contribute:

(1) A method that uses a VLM to predict how people might experience a designed space, and
(2) The Experiential Views application that demonstrates our approach with a floor plan visualization interface and an integration into an existing 3D interface.

## 2 METHODOLOGY

To explore the feasibility of our concept, we decided to evaluate the design of an office building that we have access to and are personally familiar with. The building's office space has two floors (2F and a mezzanine, 2M) composed of desk areas, working spaces, meeting rooms, kitchens, hallways, etc. We used real photos and 3D renders of the space in two datasets, each further split into training and testing sets. We refer to these images of the space as *"scenes"* (Figure 2).

### 2.1 Preliminary Human-Centric Dimensions

Our main goal in developing Experiential Views is to help designers consider human experiences and emotional responses more readily while designing a built environment. Hence, we sought to fine-tune a pre-trained VLM with images we labeled according to the following set of three human-centric dimensions:

- Social dimension - whether a scene looks *"social"* or *"isolating."*
- Tranquil dimension - whether a scene looks *"tranquil"* or *"distracting."*
- Inspirational dimension - whether a scene looks *"inspirational"* or *"boring."*

We chose these dimensions because they seemed (1) sufficiently distinct from one another, (2) applicable to a wide range of building types including offices as well as university campuses, community

[1]For simplicity, we will refer to architects and architectural designers as *designers* in this paper.

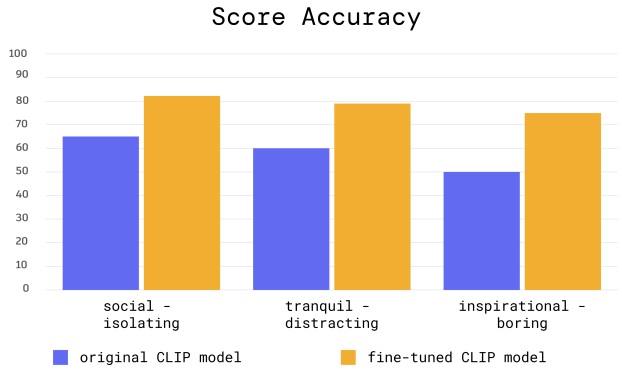



**Figure 3: A comparison of the accuracy of the original CLIP model vs our fine-tuned CLIP model (on photos).**

hubs, eldercare homes, and housing developments, and (3) consistent with recurring themes in both academic literature and online media concerning interior, architectural, and urban design for psychological comfort and human well-being. The Social dimension is supported by studies associating positive social interactions, high-quality gathering places, and a sense of community with positive outcomes for physical and mental health [12, 17]. The Tranquil and Inspirational dimensions are loosely based on ideas from the biophilic design community, which encourages the creation of spaces that are "*calming*," "*relaxing*" and tend to "*reduce stress*," as well as spaces that feel "*fresh, interesting, stimulating and energizing*" [10].

## 2.2 Model Fine-Tuning

To fine-tune a pre-trained VLM for evaluating spaces based on our human-centric dimensions, our method involved representing scenes as pairs of image and text; we conveyed the view of the space through an image (photo and 3D render), while we expressed the experience of the space as text (using our dimensions as keywords). We then let the model evaluate the scenes by computing the similarity scores between the images and text descriptions.

We chose to use OpenAI's CLIP model [21] because it can take an image and a set of text prompts as input and return a normalized score between 0.0 and 1.0 for each of the text prompts, reflecting the alignment between the content of the image and the corresponding text prompt (the closer it is to 1.0, the more aligned the image and the text are). Large pre-trained VLMs such as CLIP have shown great potential in learning representations that are transferable across a wide range of downstream tasks [24]. Furthermore, [11] showed evidence that CLIP can "learn to perform visual sentiment analysis with minimal training effort."

We then fine-tuned this model with our training datasets of photos and renders, respectively, to get specialized classifiers for our dimensions. One of the authors captured over 150 photos within the office space. Two of the authors selected 100 of these photos for evaluation and 50 photos for training while removing images that were sufficiently similar to be considered duplicates. These two authors then manually evaluated and classified the 150 selected images. At the inference phase, we provided the model with an image that was not included in the training dataset. For each of our three human-centric dimensions, we characterized the two extremes of the dimension with two words, for instance; we used the opposing words "social" vs. "isolating" for the Social dimension. We query the fine-tuned model with the image and the text prompts "*this feels like an* x *space,*" with x being either of the two words, and obtain scores for both text prompts. We then normalize the two scores to get the final score for the dimension.

The original CLIP model was trained on a dataset with diverse images that are mostly different from the views of an office building. This original model could return scores for the type of images and text in our dataset; however, the accuracy of the scores relative to the manual evaluations was ~58% on our testing sets, which is too low for our application. After fine-tuning with 50 photos from our training dataset[2], the accuracy increased to ~78% (Figure 3). The accuracy scores are computed using the evaluation set of 100 photos disjoint from the training set of 50 photos. We utilized the fine-tuned CLIP model as a central component of Experiential Views to evaluate new scenes, resulting in a higher level of accuracy.

If a VLM model of this nature were to be made available to practicing designers, we anticipate that it would be fine-tuned using a larger and more diverse set of images representing a greater variety of buildings. A designer would then have the option to further refine the model by supplying their own classified images, but they would not be required to do so.

# 3 THE SYSTEM: EXPERIENTIAL VIEWS

We developed two methods of interacting with Experiential Views that evoke common CAD/BIM environments:

(1) A *floor plan visualization*, which is a dedicated interface that collects and visualizes all the evaluated scenes in a single view (Figure 4) and
(2) An integrated UI to a WebGL-based *3D viewer*, which demonstrates how our concept can merge with preexisting BIM tools (Figure 5).

## 3.1 Floor Plan Visualization

The floor plan visualization displays the location and view cones of all of the evaluated scenes on a floor plan/map of the entire space. Figure 4 visualizes scenes in the office building that we evaluated to demonstrate our concept. In this interface, each scene is represented by a circle that is placed in the location where the scene was captured. The direction of the view for each scene is represented by a cone attached to its circle. This enables designers to see, at a glance, how much of the entire space has been evaluated in a way that is similar to the score map visualization by Li et al. [19]. More details about the scene are shown through a tooltip that contains the photo and 3D render of the scene and the scores produced by our fine-tuned model for each dimension (represented as bars).

Figure 4d shows the list of evaluated scenes as icons containing the scene photo and score bars. Clicking or hovering over these icons shows their location on the floor plan by activating their corresponding circle and tooltip. Designers can sort this list by selecting an option from the sort menu (Figure 4e). Selecting a dimension will sort the scenes based on their scores for the selected

[2]Note that each image can occur in multiple image-text pairs.

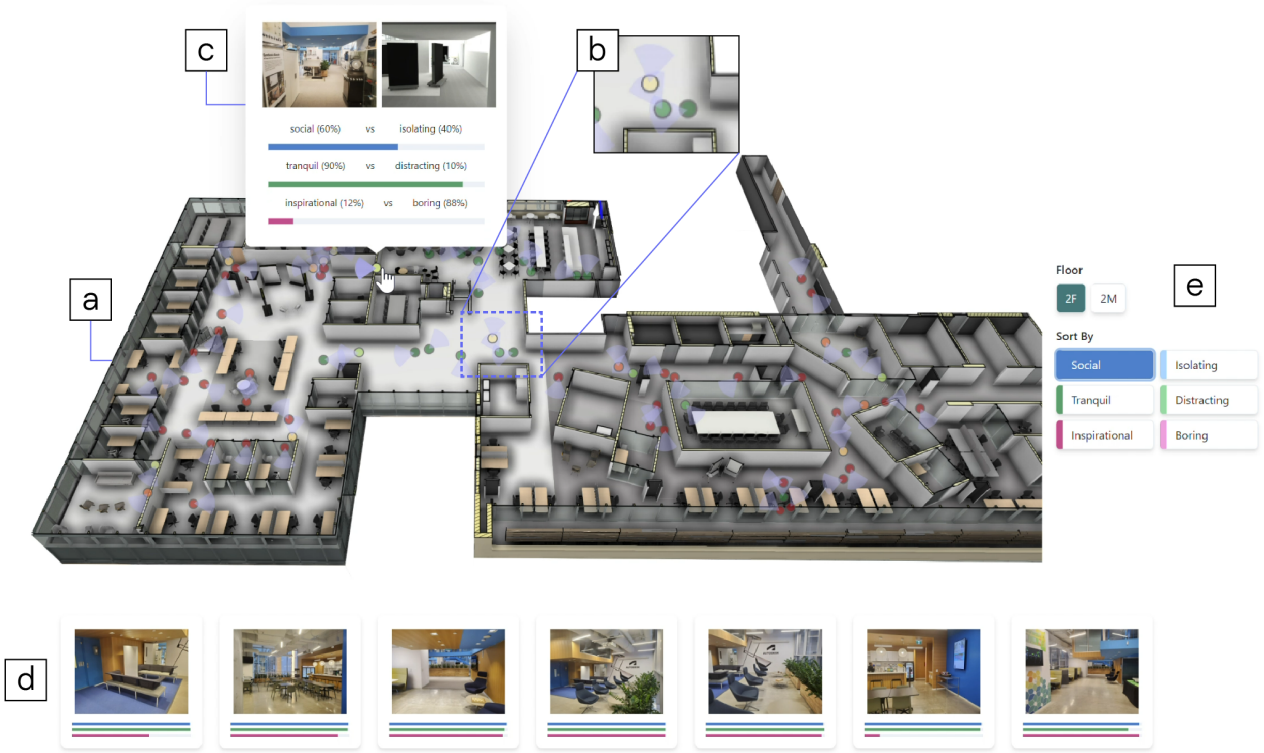



**Figure 4: The Floor Plan Visualization is composed of (a) a floor plan containing (b) evaluated scenes represented as circles with a cone corresponding to the direction of the scene's view. Hovering over a circle brings up the (c) tooltip containing the photo and render of the scene and its scores. (d) A list of the evaluated scenes. (e) Sort menu and floor picker.**

dimension in descending order. This is also reflected on the circles on the map—for the selected dimension, high-scoring scenes will appear more green, while low-scoring ones will appear more red.

### 3.2 3D Viewer Integration

We integrated Experiential Views with a widely used WebGL 3D model viewer to enable designers to dynamically navigate through their designed spaces and evaluate scenes (see Figure 5). This interface can be useful for evaluating the design of a space that has not yet been built. Using this interface, the designer can navigate through a model of a building and repeatedly evaluate the current view. When the user clicks the "Score View" button, the current view gets sent to a web service hosting our fine-tuned model. The model then computes the scores and sends them back to the viewer where they are displayed on the left (Figure 5b). Previous scenes and their scores are displayed underneath to facilitate comparisons between different views (Figure 5c). In future iterations, when the response time of the model becomes faster, the scoring can be performed in real time whenever the view changes.

## 4 DISCUSSIONS AND FUTURE WORK

The development of Experiential Views raises a number of issues regarding the practicality of the concept and the selection of human-centric dimensions. We discuss these topics and outline how future work might overcome the current limitations of the prototype.

### 4.1 Examining the Practicality of our Concept

While we were able to show a proof-of-concept that a VLM could be leveraged to evaluate spaces based on human-centric criteria, there are still some areas of improvement to ensure the practicality of our approach. For instance, we need to evaluate this concept with architects and designers to investigate how they might incorporate such a tool into their current workflows. We must also expand the scope of this work by including other types of buildings beyond office spaces and fine-tuning the model with considerably more data (photos and/or renderings).

Since our current implementation evaluates an already-built office space, another interesting direction is to explore how designers might use this concept to evaluate buildings before they are built. Our tool can theoretically handle this use case by evaluating 3D renders; in practice, however, we found that our 3D renders score less accurately than real photos of the space (see Figure 6). One possible reason could be that during the early stages of architectural design, it is common for 3D models to lack certain elements such as furniture, realistic wall colors, and decor. Consequently, the renderings obtained from these 3D models might fail to accurately portray realistic lighting conditions, color schemes, and furniture arrangements. The result is that the rendered spaces do not feel "lived in", which can greatly impact how people experience a space as well as undermine a predictive model. To investigate this issue,

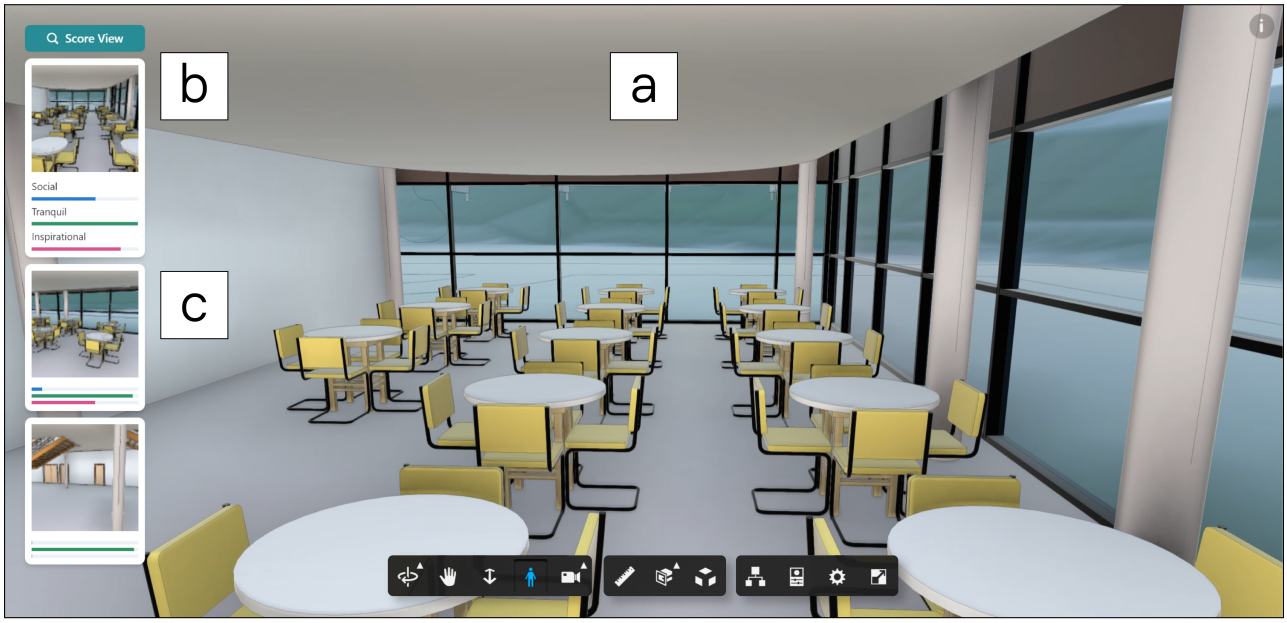



**Figure 5: Experiential Views' 3D viewer integration interface showing (a) the 3D model of the space, (b) scene capture button and scores, and (c) a minimalist view of previously evaluated scenes and their scores.**

we experimented with Stable Diffusion, a deep-learning text-to-image diffusion model [6, 22], as a means of adding stylization and detail to the renderings. The stylized renderings were not entirely faithful to the actual geometry of the building, but they tended to bring out the character of the actual real-world spaces. We noted that, at least in some cases, this technique resulted in improved prediction accuracy. Figure 6 provides a comparison of scores between a scene depicted as a basic 3D render and one that was enhanced using Stable Diffusion. Following the enhancement there was a noticeable increase in the "*inspirational*" score of the scene, better aligning with how we would evaluate the actual space.

It is worth noting that both architectural and interior design elements have been found to affect human experience in buildings [15], suggesting that it is important to predict the eventual qualities of a space before architectural decisions are finalized and only furnishing and decorating decisions remain. Future studies can look into how different stylistic renderings could affect how people perceive a space compared to scores given by a system like Experiential Views.

## 4.2 Choosing Human-Centric Dimensions

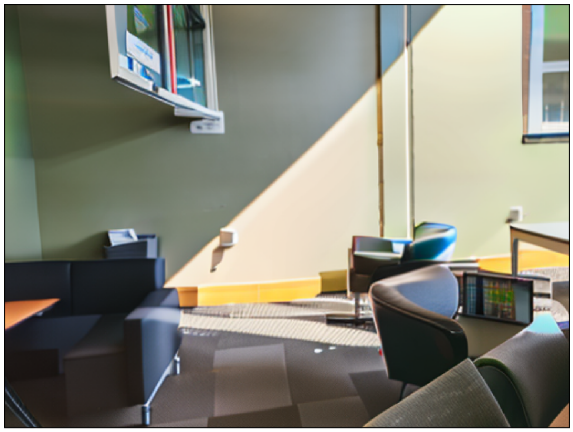
One question that arises with our proposed system is how to determine the dimensions and criteria for evaluating designed spaces. The selection of Social, Tranquil, and Inspirational dimensions for our initial exploration was influenced and supported by the existing literature [10, 12, 17], but also influenced by our personal preferences and potentially affected by cultural norms. It could be argued that any attempt to characterize psychological or emotional responses will be somewhat arbitrary and susceptible to bias, though lessons can be learned from recent efforts to crowdsource experiential data in a systematic manner.

In a study by Coburn et al. [14], roughly 800 online participants were recruited to collectively evaluate 200 photos of architectural interiors according to 16 subjective questions. Rather than pre-grouping similar questions, the authors used principal component analysis to determine a set of three dimensions that collectively explained most of the variance in the respondents' ratings of the scenes. The authors named one of the dimensions "Hominess", which appears similar to our Tranquil Dimension. The other two dimensions were named "Coherence", and "Fascination", which seem to overlap with our Inspirational dimension. The concept of a space feeling "*social*" was largely absent from the 16 questions that generated the Hominess-Coherence-Fascination framework. It was later found that different groups of participants—specifically design students, participants with autism spectrum disorder, and a neurotypical control group—expressed different preferences on average for homey, coherent, and fascinating spaces [23].

The Hominess-Coherence-Fascination framework is interesting in that it was obtained through scientific observation and statistical analysis, though it is possible that a set of dimensions based on values could prove easier for designers to interpret and apply. Altaf et al. [4] use crowdsourced data to evaluate photos of spaces at a university campus according to their three chosen "constructs of well-being": belonging, self-efficacy, and environmental efficacy. This study is notable for its use of indirect questions, such as "*I would pick up litter that is not my own in this space.*" We observe that their concept of belonging appears more closely related to our Social dimension than the dimensions in the Hominess-Coherence-Fascination framework.

Regardless of what framework or set of dimensions is chosen for human experience evaluation of designed spaces using VLMs, crowdsourcing data from a diversity of participants around the

```
Render Score: (inaccurate)
{
    "boring": "0.99253464",
    "distracting": "0.019161317",
    "inspirational": "0.007465348",
    "isolating": "0.9746235",
    "social": "0.025376452",
    "tranquil": "0.98083866"
}
```



```
With Stable Diffusion Score: (accurate)
{
    "boring": "0.10956594",
    "distracting": "0.020431139",
    "inspirational": "0.890434",
    "isolating": "0.0952176",
    "social": "0.9047824",
    "tranquil": "0.9795688"
}
```

**Figure 6: Comparison scores of a scene represented as a plain 3D render and one that was stylized through Stable Diffusion. After the stylization, the score of the scene for "*inspirational*" noticeably improved.**

world will likely improve the robustness of the prediction system and reduce bias. It may also be beneficial to display some form of diversity or uncertainty output, informing designers when people with different cultural backgrounds, neurodiversity or life experiences will likely experience radically different responses to a given space. Ideally, the system would support a diversity of languages, and allow the evaluated criteria to be customized based on a design team's objectives or even personalized according to an occupant's individual preferences.

## 5 CONCLUSION

We presented Experiential Views, a proof-of-concept application that envisions how designers might evaluate designed spaces based on the possible experiences of the people who will inhabit them. Experiential Views leverages a pre-trained VLM (OpenAI CLIP) to predict these experiences according to a set of pre-selected human-centric dimensions. We demonstrated the approach by fine-tuning this VLM with scenes of an office space that we labeled as *social* or *isolating*, *tranquil* or *distracting*, and *inspirational* or *boring*. We also developed two interfaces—a floor plan visualization and a 3D Viewer Integration—that explore how building designers might apply this human experience prediction technology while navigating CAD/BIM models and viewing photos and renders.

Future work to make the system practical must include user studies, a dedicated open-source effort to collect data from a diversity of building types, and further investigation of rendering enhancement techniques. Strategies for choosing alternative sets of human-centric dimensions, and measuring psychological and emotional responses accordingly, also deserve serious consideration. It is worth noting that our prediction model proved surprisingly effective despite the fact we only used 50 images to fine-tune the VLM. The accuracy of the predictions improved from roughly 58% to 78% with the fine-tuning on the dataset of photos. This shows the potential of large pre-trained VLMs to be easily customized with only a small set of user data, encouraging further investigation of the approach as a whole.

We see Experiential Views as an early step towards novel tools and applications that take advantage of recent AI technology to enable human experience evaluation of designed spaces. The presented prototype represents a new opportunity to help designers improve the experiences of people in built environments.

## REFERENCES


[1] 2024. Autodesk AutoCAD. https://www.autodesk.com/products/autocad/overview. [Online accessed Jan-2024].

[2] 2024. Autodesk Revit. https://www.autodesk.com/products/revit/overview. [Online accessed Jan-2024].

[3] Sharmeen M.Saleem Abdullah, Siddeeq Y. Ameen, Mohammed A. M. Sadeeq, and Subhi Zeebaree. 2021. Multimodal Emotion Recognition using Deep Learning. *Journal of Applied Science and Technology Trends* 2, 02 (Apr. 2021), 52–58. https://doi.org/10.38094/jastt20291

[4] Basma Altaf, Eva Bianchi, Isabella P. Douglas, Kyle Douglas, Brandon Byers, Pablo E. Paredes, Nicole M. Ardoin, Hazel R. Markus, Elizabeth L. Murnane, Lucy Z. Bencharit, James A. Landay, and Sarah L. Billington. 2022. Use of Crowdsourced Online Surveys to Study the Impact of Architectural and Design Choices on Wellbeing. *Frontiers in Sustainable Cities* 4 (2022). https://doi.org/10.3389/frsc.2022.780376

[5] Josh Andres, Rodolfo Ocampo, Oliver Bown, Charlton Hill, Caroline Pegram, Adrian Schmidt, Justin Shave, and Brendan Wright. 2023. The Human-Built Environment-Natural Environment Relation - An Immersive Multisensory Exploration with 'System of a Sound'. In *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) *(IUI '23 Companion)*. Association for Computing Machinery, New York, NY, USA, 8–11. https://doi.org/10.1145/3581754.3584119

[6] AUTOMATIC1111. 2022. Stable Diffusion WebUI. https://github.com/AUTOMATIC1111/stable-diffusion-webui.

[7] Kirsten Boehner, Rogério DePaula, Paul Dourish, and Phoebe Sengers. 2007. How Emotion is Made and Measured. *International Journal of Human-Computer Studies* 65, 4 (2007), 275–291. https://doi.org/10.1016/j.ijhcs.2006.11.016

[8] Alessandro Bondielli and Lucia C. Passaro. 2021. Leveraging CLIP for Image Emotion Recognition. In *Proceedings of the Fifth Workshop on Natural Language for Artificial Intelligence (NL4AI 2021) co-located with 20th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2021), Online event, November 29, 2021 (CEUR Workshop Proceedings, Vol. 3015)*, Elena Cabrio, Danilo Croce, Lucia C. Passaro, and Rachele Sprugnoli (Eds.). CEUR-WS.org. https://ceur-ws.org/Vol-3015/paper172.pdf

[9] Jeffrey A. Brooks, Vineet Tiruvadi, Alice Baird, Panagiotis Tzirakis, Haoqi Li, Chris Gagne, Moses Oh, and Alan Cowen. 2023. Emotion Expression Estimates to Measure and Improve Multimodal Social-Affective Interactions. In *Companion Publication of the 25th International Conference on Multimodal Interaction* (Paris, France) *(ICMI '23 Companion)*. Association for Computing Machinery, New York, NY, USA, 353–358. https://doi.org/10.1145/3610661.3616129

[10] William Browning, Catherine Ryan, and Joseph Clancy. 2014. *14 Patterns of Biophilic Design: Improving Health & Well-Being in the Built Environment*. Technical Report. Terrapin Bright Green.

[11] Cristina Bustos, Carles Civit, Brian Du, Albert Solé-Ribalta, and Àgata Lapedriza. 2023. On the use of Vision-Language models for Visual Sentiment Analysis: a study on CLIP. In *11th International Conference on Affective Computing and Intelligent Interaction, ACII 2023, Cambridge, MA, USA, September 10-13, 2023*. IEEE, 1–8. https://doi.org/10.1109/ACII59096.2023.10388075

[12] Michael Campo and Habib Chaudhury. 2012. Informal Social Interaction Among Residents with Dementia in Special Care Units: Exploring the Role of the Physical and Social Environments. *Dementia* 11, 3 (2012), 401–423. https://doi.org/10.1177/1471301211421189

[13] Sunwoo Chang and Hanjong Jun. 2019. Hybrid Deep-Learning Model to Recognise Emotional Responses of Users towards Architectural Design Alternatives. *Journal of Asian Architecture and Building Engineering* 18, 5 (2019), 381–391. https://doi.org/10.1080/13467581.2019.1660663

[14] Alexander Coburn, Oshin Vartanian, Yoed N. Kenett, Marcos Nadal, Franziska Hartung, Gregor Hayn-Leichsenring, Gorka Navarrete, José L. González-Mora, and Anjan Chatterjee. 2020. Psychological and Neural Responses to Architectural Interiors. *Cortex* 126 (2020), 217–241. https://doi.org/10.1016/j.cortex.2020.01.009

[15] Susanne Colenberg, Tuuli Jylhä, and Monique Arkesteijn. 2021. The relationship between interior office space and employee health and well-being – a literature review. *Building Research & Information* 49, 3 (2021), 352–366. https://doi.org/10.1080/09613218.2019.1710098

[16] Sinuo Deng, Lifang Wu, Ge Shi, Lehao Xing, Wenjin Hu, Heng Zhang, and Ye Xiang. 2023. Simple But Powerful, a Language-Supervised Method for Image Emotion Classification. *IEEE Transactions on Affective Computing* 14, 4 (2023), 3317–3331. https://doi.org/10.1109/TAFFC.2022.3225049

[17] Jacinta Francis, Billie Giles-Corti, Lisa Wood, and Matthew Knuiman. 2012. Creating Sense of Community: The Role of Public Space. *Journal of Environmental Psychology* 32, 4 (2012), 401–409. https://doi.org/10.1016/j.jenvp.2012.07.002

[18] Wilson S Geisler. 2008. Visual Perception and the Statistical Properties of Natural Scenes. *Annual Review of Psychology* 59 (2008), 167–192. https://doi.org/10.1146/annurev.psych.58.110405.085632

[19] Changyang Li, Haikun Huang, Jyh-Ming Lien, and Lap-Fai Yu. 2021. Synthesizing scene-aware virtual reality teleport graphs. *ACM Transaction on Graphics* 40, 6, Article 229 (dec 2021), 15 pages. https://doi.org/10.1145/3478513.3480478

[20] Neda Norouzi, Antonio Martinez, and Zayra Rico. 2023. Architectural Design Qualities of an Adolescent Psychiatric Hospital to Benefit Patients and Staff. *HERD: Health Environments Research & Design Journal* 16, 4 (2023), 103–117. https://doi.org/10.1177/19375867231180907

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models from Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 8748–8763. https://proceedings.mlr.press/v139/radford21a.html

[22] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10684–10695. https://doi.org/10.1109/CVPR52688.2022.01042

[23] Oshin Vartanian, Gorka Navarrete, Letizia Palumbo, and Anjan Chatterjee. 2021. Individual Differences in Preference for Architectural Interiors. *Journal of Environmental Psychology* 77 (2021), Article 101668. https://doi.org/10.1016/j.jenvp.2021.101668

[24] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Learning to Prompt for Vision-Language Models. *International Journal of Computer Vision* 130, 9 (2022), 2337–2348. https://doi.org/10.1007/s11263-022-01653-1