Research Paper
AR3EX115

William Kosta
Student no. 5941369

Decomposing Architecture Atmospheres Using Foundation Models

Explore Lab 39
Nov 2024

Delft University of Technology
Faculty of Architecture and
the Built Environment

Table of Contents

Tutors:

Geert Coumans - Architecture
Seyran Khademi - Research
Rico Heykant - Building Technology
Casper van Engelenburg - Research Support

Abstract

The advancement of artificial intelligence provides humans with new tool sets that make various tasks more efficient and even unlock possibilities that were once unimaginable. There are a lot of different tools that AI materialises in and different ways those tools can be used. This paper aims to explore a small part of AI and apply it in architecture.

This paper focuses on architecture atmosphere, which can also be loosely interpreted as the prevailing tone or mood of a space. Architecture atmosphere is difficult to grasp and create accurately. Part of this is because there are no established definitive classification of architectural atmospheres which could be referred to and followed. Architecture atmosphere is a property that emerges from measurable elements. Foundation models (machine learning models) excel at processing measurable high dimensional data (images) at scale in a consistent way and then recognising implicit and abstract concepts out of it. Therefore it can be used to create vector representations of images, which then can be used to create a 2D plot that could be used to navigate architecture atmospheres in a meaningful and more quantitative way.

This paper examines the effectiveness of foundation models through experiments and evaluation of the resulting plots. Having done the experiments, it becomes apparent that using foundation models to explore architecture atmospheres has potential. The models are able to pick up relevant features, however it needs to be adjusted to prioritise these features more. Fully representing architecture atmosphere digitally is not easy. However, there are a lot of influential factors that affect atmospheres through visual means. In regards to these influential factors, with a larger data set and some re-training, foundation models are promising tools to use for addressing architectural atmospheres.

1

Problem Statement & Research Questions

Problem Statement

Introduction

The advancement of artificial intelligence (AI) has been growing rapidly. The realisation that technology is able to perform tasks previously considered possible only for humans has sparked the trend of applying AI in various fields of work. Architecture is no exception to this. One of my first interactions with AI was through Adobe's generative fill (Figure 1), where the software fills or replaces a highlighted area within a canvas with a desired image specified by the user in a prompt. The accuracy and quality of the resulting fill sparked my interest in exploring the extent of AI and its capabilities, especially in architecture.

In architecture, more recently, popular AI image-generation tools like DALL-E (Ramesh et al., 2021) has created the possibility of using machine learning models to seek inspiration. However, so far, aside from this, AI has been largely used to interact with the more pragmatic and quantifiable aspects of architecture, such as predicting building performance.

This application of AI in architecture is imbalanced because architecture lies at the intersection of science and art. On the practical side, it must be pragmatic and comply with structural and engineering requirements, but on the artistic side, architecture should be able to 'move' people. One quality that helps achieve this is atmosphere. This is a term that Peter Zumthor uses to describe this quality (Zumthor, 2006), which can also be loosely interpreted as the prevailing tone or mood of a space. Atmosphere is an emergent quality influenced by many factors. One of these factors is natural light. This is a factor I would like to focus on, especially later in my design, because it has dual nature: it is physically measurable, but in an atmospheric sense it is also evocative.

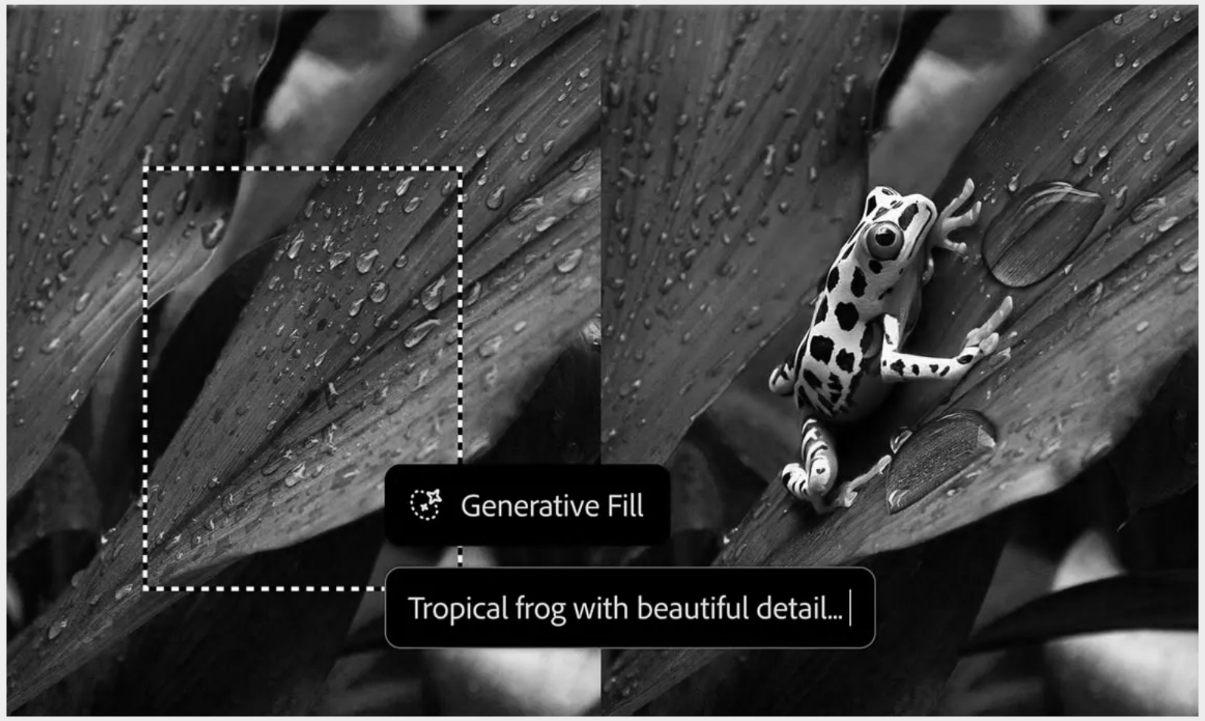


Figure 1. Adobe generative fill

Figure 2. St. Pierre, Firminy, Le Corbusier

Problem Statement

There are far more architects whose designs aim at publicity and attention, compared to those who focus on atmospheres. At present, atmosphere-driven design, like that of Peter Zumthor, or the later work of Le Corbusier (Figure 2 & 3), is becoming a niche within architecture. When atmosphere is neglected, we have less control over it, leaving the outcome to chance. Without a systematic way to approach atmospheres, we risk losing the ability to create spaces with prominent atmospheres—such as those that approach the sublime, like churches or monuments—which are essential to human culture and expression. One reason this decline in importance of atmosphere occurs is that designing with atmosphere in mind is not easily transferable.

My research aims to help solidify atmospheres as an important quality to be considered in design by making it more accessible. Because atmosphere is a fluid concept, I take inspiration from *Anchoring the Design Process* (Van Dooren, 2020), where the author's goal is to analyse, break down, and work on the design process in architecture (which is also a fluid and implicit concept). In the text, the 'design process' is abstracted, and a framework is created. With this framework, it becomes possible to address the fluid concept of the 'design process' in a more accurate and meaningful way. It is under this parallel condition of the 'design process' and atmospheres both being fluid and lacking vocabulary, framework and tools to be explored properly, that we can identify a shared problem of measurability and therefore a need to address them.

The nature of atmosphere being an abstract quality that humans feel, does not mean that it is born out of purely immeasurable elements. Atmosphere is conceived through a combination of measurable elements in the building that humans perceive. These perceived elements are then processed on an abstract level in the brain, which is then felt as atmosphere. This process of recognising implicit and abstract concepts out of measurable input is something that foundation models do well. Foundation models are AI models trained on broad data that can be adapted to a wide range of tasks (Bommasani et al., 2021). They effectively convert input data into representation that can be used for classification. This is done through an implicit training process rather than through hard coding or explicit instructions. In addition to that, foundation models gives access to investigating this topic in large scale data. This is important because there are patterns that can only become apparent when a large dataset is being used, such as similarities and trends. This makes foundation models an appropriate tool to explore for this research.

Given the limited of examples of AI being applied to the intangible aspects of architecture, this leads to the following questions:

Figure 3. Colline Notre Dame du Haut
Ronchamp, Le Corbusier

Research Question

To what extent are foundation models an effective tool to approach and address architectural atmospheres?

Sub Questions

1. How can we systematically approach the topic of atmospheres?

2. How can we effectively collect a large dataset of images that visually convey atmosphere?

3. Can atmospheres be clustered into different groups? What are the main groups?

4. What are the different ways natural light can be used to contribute to the creation of atmospheres?

2

Framework

Framework (Conceptual)

In this conceptual section, the key concept of atmospheres, which the research tries to capture will be explained, through literary sources which speak about the topic.

(The following four pages are based on *Atmospheres* (Zumthor, 2006))

Atmospheres

In his book *Atmospheres* (Zumthor, 2006), Peter Zumthor describes atmospheres as a quality in a building that manages to move people every single time. It is something people can sense within seconds of entering a building, much like a first impression. This sensation is closely linked to our primal survival instincts, through which we evolved to perceive and judge environments quickly—a contrast to our more logical and slower linear thinking. This suggests that atmosphere is something deeply embedded and natural to humans.

When designing for a specific atmosphere, Zumthor considers the 'Body of Architecture', where he considers architecture just like the human body, consisting all of the different parts and layers. He talks about how the body of architecture can 'touch' its users, through physical touch in elements such as handrails. Experiencing a building is rich and the way elements are put together, such as how beams and columns come together is full of character and contributes to the user experience.

He also talks about 'Material Compatibility'. He views materials as something that has endless possibilities. Such as stone, whether it is drilled into, split, polished or not, will all result in different outcomes. To him, materials react with one another and is radiant. It brings out different feelings and weight when interacting with one another. There is also the dynamic nature of how it ages. Therefore material composition always gives rise to something unique.

In 'Sounds of Space' Zumthor talks about a more sensory observations in a building. He talks about sound that results due to the shape of the room and its literal acoustic properties and the kind of feeling it evokes. But he also thinks about it in a tactile way, such as how the floor sound when it is just hollow timber compared to when it is stuck on a concrete slab. On an even more abstract level, he thinks about what sound the building emits when all foreign sound is removed.

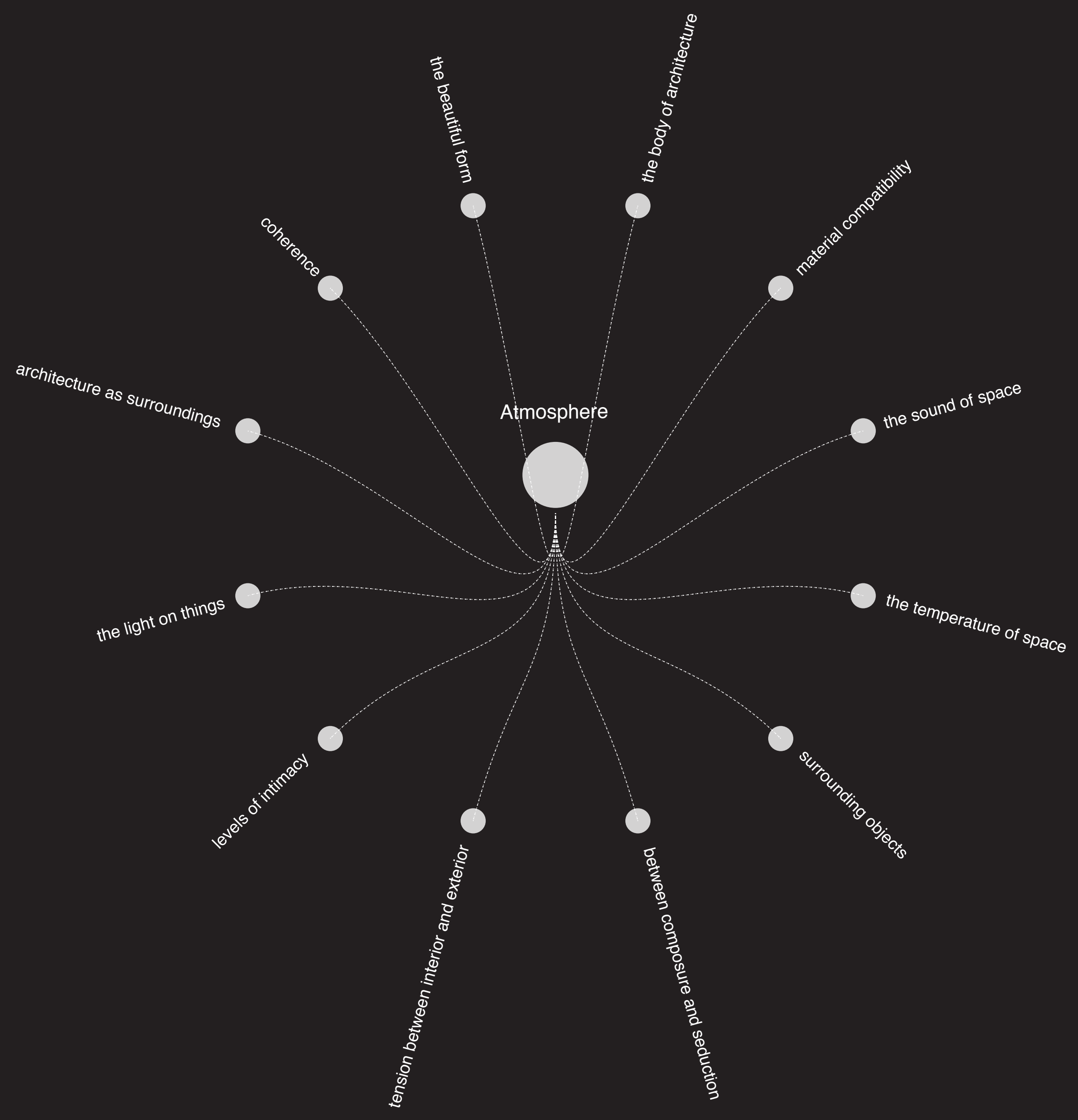


Figure 4. Factors affecting atmospheres

Framework (Conceptual)

'The Temperature of a Space' refers to the literal temperature of a building, and also how materials are perceived, which is also correlated to its physical property, for example steel is often considered as cold and industrial, whereas timber is considered as warm. But he also thinks of the verb 'to temper', similar to tempering pianos, where it is tuned to adjust for its imperfection, the same is done to the atmosphere of a space.

With 'Surrounding Objects', Zumthor refers to the different kinds of activities and objects that will be added to the building after the construction is finished. He writes that when entering a space, especially a personal space that has been used by an individual a lot, the space starts to be shaped and has the presence of the user, even without the user being there. This anticipation of the activities and the possible items that will be brought into the space in the future, is something to be taken into account.

In the section 'Between Composure and Seduction' he talks about movement in a building. In buildings with high functional needs, such as hospitals, have clear requirements of being explicit in regards to directing people to where they need to be. However, Zumthor brings the example of the baths he designed, where there was the intention of bringing the different spaces together by directing the users, but doing it in a subtle way, a contrast to the way directing would be done in a hospital. In the case of the baths he does this by designing points of interests in the space, such as corners, or the way light falls in certain parts, attracting the users to approach and explore without forcing it architecturally. With this, Zumthor is also then able to curate the sequence of spaces he wants the users to experience, similar to how scenes are arranged in a film.

In 'Tension between Interior Exterior', Zumthor talks about how differently he treats the building façade and the interior. The exterior expresses the wall and what the outside sees, expressing what the building wants to 'say', but not exposing everything. Only when a user goes through the door can the inside is the rest revealed. How the building reacts to the context and what the context sees from the outside is something that affects atmosphere.

With 'Levels of Intimacy' he talks about sizes of buildings and its elements relative to humans. He talks about how tall and short doors, thick and thin walls, or the relative sizes of ceilings to a person evokes different experiences. The feeling that we have when the size of a space is much bigger than a person and the feeling we have when the scale is more 'domestic' different. There is no good or bad sizes, but it has to be considered and used appropriately.

With 'The Light on Things' Zumthor talks mainly about natural lighting. How materials look, specifically when sunlight enters a space is something he considers when choosing them. The idea of seeing the plan as a void, then carving into the plan allowing light to seep in is also something he proposes in order to really put daylight into focus instead of an afterthought. He writes that sunlight is something that has an almost spiritual quality not belonging to this world, suggesting that there is something beyond that is greater. Depending on how it is used, this nature of light greatly affects the atmosphere of a space.

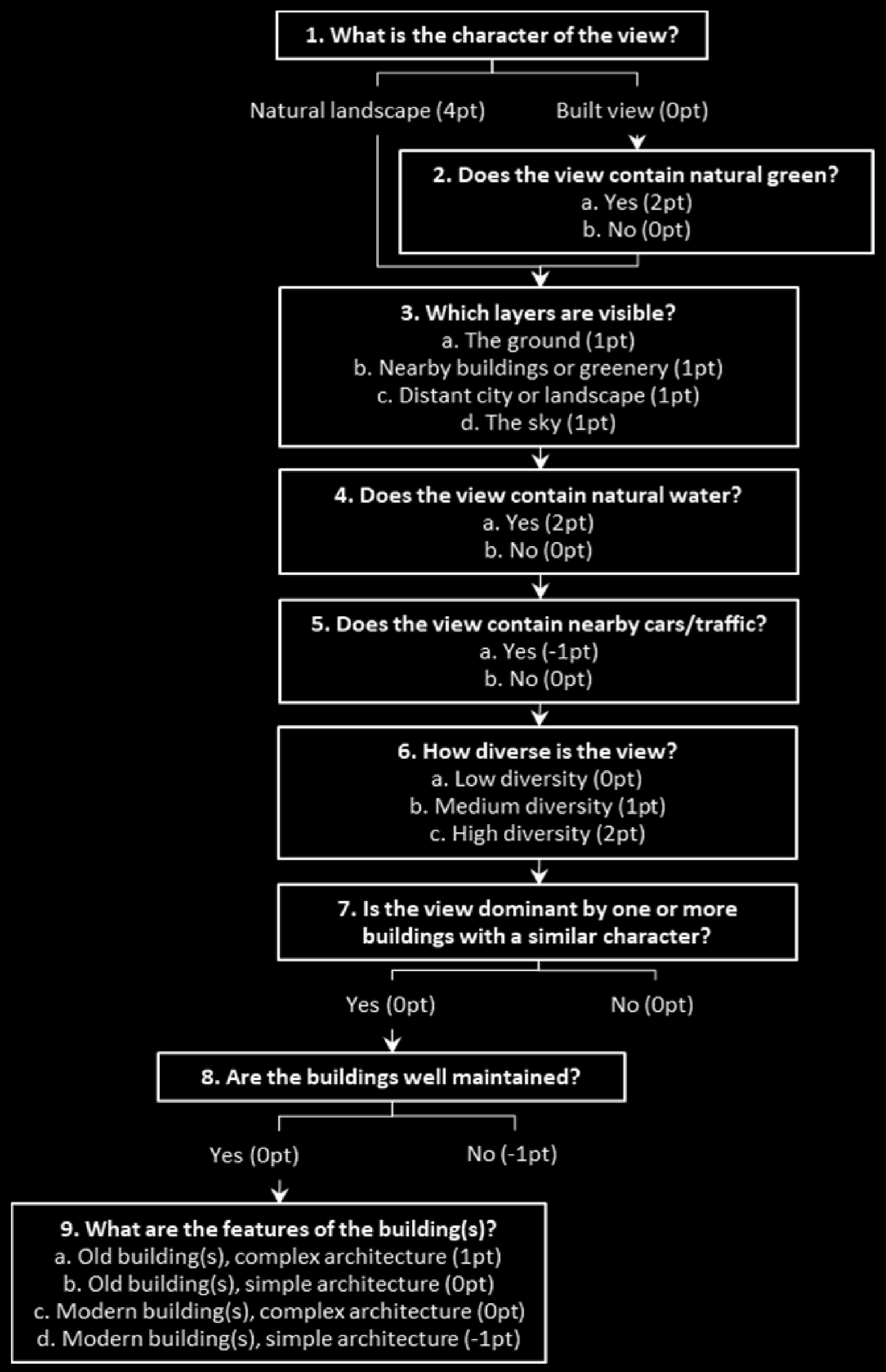
The points mentioned by Zumthor are linked to the way humans experience a building, a theme also discussed in *The Eyes of the Skin* (Pallasmaa, 2005). In this text, the author argues for an emphasis on human senses beyond sight, such as touch, sound, and smell. Pallasmaa suggests that more attention to these senses will result in a richer architecture. He also proposes that sight can be understood as an extension of touch; when we see, we are not merely looking—our brain incorporates experiences associated with the visual input, connecting us more intimately with materials and textures, rather than perceiving them only as visual elements.

Another source that elaborates on the theme of atmospheres is *The Poetics of Space* (Bachelard, 2014). In this book, Bachelard discusses space in a phenomenological way. He introduces the concept of the 'poetic image'—images linked to space that have the power to resonate deeply and emotionally with people. He argues that such images evoke universal feelings shared by many. For example, an image conveying the idea of 'home' is often associated with warmth, protection, and intimacy.

Framework (Conceptual)

Drawing from the sources mentioned before, it could be speculated that an image can provide a glimpse of a space's atmosphere. While it may not be as powerful as experiencing the building itself, the right images can convey significant (emotional) information about a space.

The concepts mentioned above are qualitative. However, there has been precedent in making qualitative concepts become something that is measurable. The 'Quantifying Window View Quality' paper (Abd-Alhamid et al., 2022) attempts to close the gap of the non-existence of established approaches or regulations when it comes to window view quality. In the paper, some methods such as making a points-based system on what is visible on a window view (Figure 5) is examined. My research paper will take this as precedent and use it as inspiration, adapting as needed for the theme of architectural atmospheres. Different experiments will be conducted and the qualitative theme of architecture atmosphere will be approached in an attempt to make it more measurable.



1. What is the character of the view?

Natural landscape (4pt)

Built view (0pt)

2. Does the view contain natural green?
a. Yes (2pt)
b. No (0pt)

3. Which layers are visible?
a. The ground (1pt)
b. Nearby buildings or greenery (1pt)
c. Distant city or landscape (1pt)
d. The sky (1pt)

4. Does the view contain natural water?
a. Yes (2pt)
b. No (0pt)

5. Does the view contain nearby cars/traffic?
a. Yes (-1pt)
b. No (0pt)

6. How diverse is the view?
a. Low diversity (0pt)
b. Medium diversity (1pt)
c. High diversity (2pt)

7. Is the view dominant by one or more buildings with a similar character?

Yes (0pt)

No (0pt)

8. Are the buildings well maintained?

Yes (0pt)

No (-1pt)

9. What are the features of the building(s)?
a. Old building(s), complex architecture (1pt)
b. Old building(s), simple architecture (0pt)
c. Modern building(s), complex architecture (0pt)
d. Modern building(s), simple architecture (-1pt)

Figure 5. Diagram showing one method of using a points based system to quantify the quality of a window view

Framework (Technical)

Having set out the concept in the previous section, this section will explain the technical approach of the research and why it is being used to address atmospheres in the research.

Machine Learning (ML)

ML is a subset of AI, which involves training machines to perform complex tasks such as face verification (Sengupta et al., 2016), object detection (Redmon et al., 2015), and prompt-guided text generation (Radford et al., 2018). ML also includes Deep Learning (DL), which uses multi-layered neural networks, called deep neural networks (As & Basu, 2021).

The training needed to create a model (As & Basu, 2021) can be seen as parallel to how people train to become architects, relying primarily on examples and experience rather than explicit instructions (Van Dooren, 2020). For example, to train a model to recognise whether an image shows bricks or timber, it must be provided with a set of training images. With each image, the model examines 'features' similar to how a person might visually assess whether an object is brick or timber. In this example, features could include the colour, texture, and shape of the object, among others.

Based on these features, the model assigns the image to a class. In this example, only two classes are needed: Class 1 for 'bricks' and Class 2 for 'timber'. If the input is an image of bricks, the correct output would be '1' for Class 1 and '0' for Class 2. However, if the output is incorrect, back-propagation allows the model to be adjusted for improved accuracy.

With DL (Figure 7), this learning process is automated (As & Basu, 2021): from raw data, to feature extraction, to classification, all steps are learned by the model. The model extracting features on its own is the reason why it is interesting to apply this to architecture atmospheres, as there is currently no fixed and definitive classification method of atmospheres.

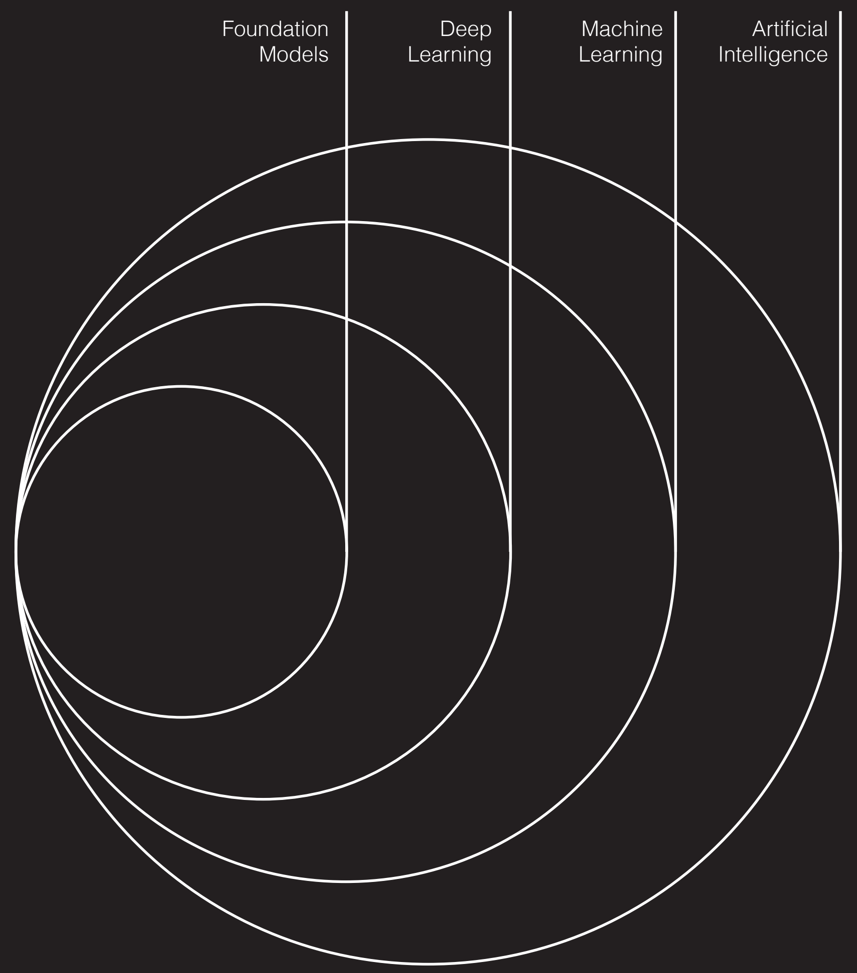


Figure 6. Diagram of Artificial intelligence hierarchy. At present, all foundation models are deep learning based

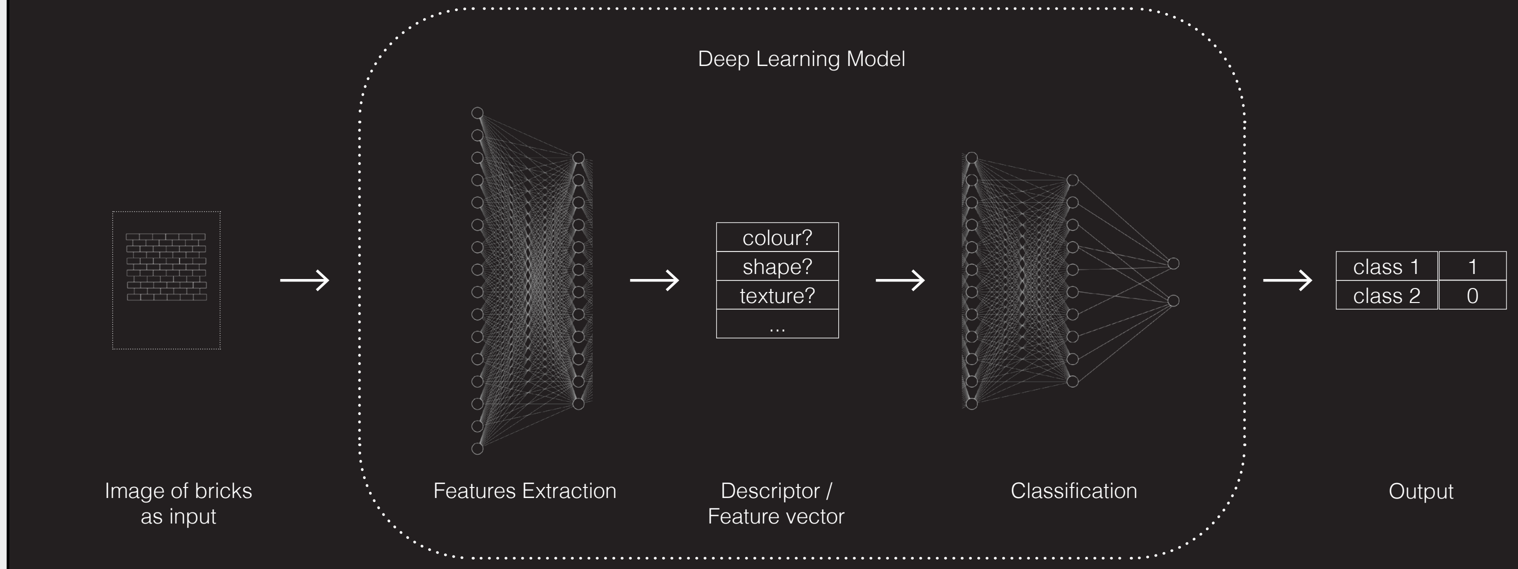


Figure 7. The fact that the models extract features and classifies on its own, based on a learning technique parallel to the way a humans train in architecture is why this method is relevant to the experiment

Framework (Technical)

Foundation Models

Collecting and annotating data for specific tasks is labour intensive. Today, the vast amount of available data, the ability to perform large-scale computations, and the existence of proper optimisation algorithms (such as DL) make it possible to develop models that can generalise across specific tasks and domains. These models are referred to as foundation models. Popular examples include GPT (Radford et al., 2018) for language understanding and generation, and DALL E (Ramesh et al., 2021) for image generation.

DINOv2 (Caron et al., 2021), EfficientNet (Tan & Le, 2019), and CLIP (Radford et al., 2021), are foundation models that will be used in this research, due to its affinity with images.

DINOv2 (Caron et al., 2021), is one of the first successful models that uses unsupervised learning. It achieves this using self-distillation (Zhang et al., 2019) and a Vision Transformer or ViT (Dosovitskiy et al., 2020). This allows it to be trained on a large and any collection of images. DINOv2 is trained on a curated dataset of 142 million images.

Self-distillation (Zhang et al., 2019) is a process of where the model creates two systems creating a 'teacher-student'-like learning system. These two systems are 2 versions of the same models. The 'teacher' system generates a prediction based on its understanding and the 'student' system tries to match the prediction while looking at the data in a different way (e.g. the image is cropped or orientated differently, etc.). Eventually the student gets better at matching the prediction and the teacher gradually updates its knowledge based on the student's progress. This allows a model to develop strong representations without the need of labelled data.

ViT (Dosovitskiy et al., 2020) applies the principles of transformers (originally developed for natural language processing) to images. ViT divides an image into 16x16 pixel patches and turns it into a vector embedding. The model then uses a process called self-attention where it analyses each patch and how they relate to one another. For instance, if the patches forms a face, the model learns that certain patches (e.g. of the nose and eyes) are related. This is powerful because ViT considers the picture as a whole, unlike other methods where small regions are analysed in isolation at a time.

EfficientNet (Tan & Le, 2019) is a family of Convolutional Neural Network (CNN) designed for efficiency at image recognition. EfficientNet consists of several models labelled from B0 to B7 with increasing size and performance. EfficientNet uses supervised learning based on the ImageNet dataset, consisting of 14 million labelled images.

CNN is a type of neural network designed to process structured grid like data, such as images. It uses a small filter to scan over an image to detect patterns like edges, textures or shapes. CNN learn in a hierarchical way. The early layers detect simple patterns such as edges, and the deeper layers combine these patterns and detects more complex structures like shape. CNN used to be the dominant method for image recognition before ViT. When compared to ViT, CNN uses labelled data and is therefore more expensive to train and does not scale as well as ViT.

CLIP (Radford et al., 2021), is a model designed to process image and text together, allowing connections between what the model 'sees' in an image and 'reads' in a text description. CLIP does this by utilising two neural networks: one for image and the other for text. CLIP is trained on 400 million text-image pairs.

The network that focuses on images creates vector embeddings (typically using ViT). While the other network focusing on text creates vector embeddings using a transformer model. After having both image and text embeddings, CLIP uses a method called contrasting learning. This is where vector representations of texts and images are adjusted in the embedding space (a high dimensional space where vector embeddings exist). The goal of contrasting learning is to bring the vector representation of the image and the corresponding representation of the text closer together (indicating that they are similar/point to the same thing).

Having merged both image and text representation in the same embedding space allows CLIP to process image and words as correlating concepts. After training, for example, in the embedding space, a vector of an image of a dog and the vector of the text 'dog running in the park' will be located close to each other, while a vector of an image of a car would be far away.

Framework (Technical)

Dimensionality Reduction

The output of most foundation models is high-dimensional, typically in the order of 1000 features. To visualise the dataset and analyse the relative distances of the dataset's instances in the outcome, a method to project the high-dimensional data into 2D or 3D is needed.

To reduce the dimensions of the output, algorithms such as UMAP (McInnes et al., 2018) or t-SNE (Van Der Maaten & Hinton, 2008) can be used. T-SNE 'compresses' the high dimensional data without losing too much meaningful information. It preserves the local distances found in high dimensional data when expressing it in 2 or 3 dimensions. Taking the example of a model recognising hand written digits from 0 to 9 (Figure 8), this is useful because if the high dimensional data output clusters a lot of pictures of the number '9' together (something that suggests that the model is working), we can still see it even when we compress the data in 2 dimensions.

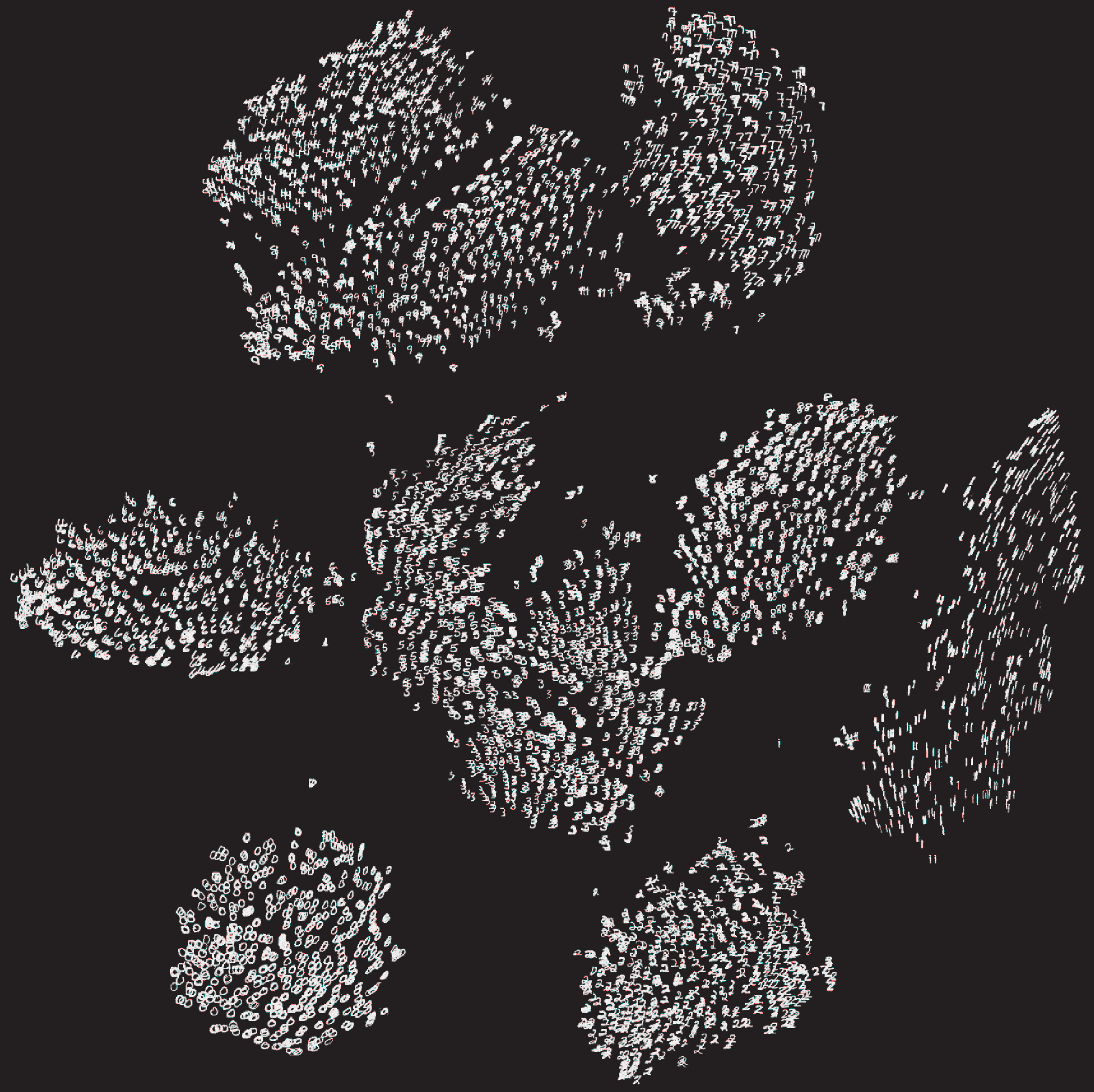


Figure 8. The graphs above are t-SNE visualisations of the MNIST dataset (Van Der Maaten & Hinton, 2008). With the large scale of the MNIST dataset, the visualisation by t-SNE reveals patterns, such as the frequency of how much people tend to write the digit '1' leaning to the right instead of left, or how similar the digit '3' and '5' is and how often it is indistinguishable. This is something that would be difficult to identify without scale and t-SNE's visualisation. The same insight is expected to be found with the topic of atmospheres in this research.

3

Research Intent, Dataset & Method

Research Intent

Sub question 1: How can we systematically approach the topic of atmospheres?

Architecture atmospheres is difficult to grasp and create. However, it is something designers interact with regularly. Currently, when trying to achieve a specific atmosphere, the preferred method is to find precedents of buildings with the desired atmosphere and studying the elements that contribute to the atmosphere. However, it is difficult to find a set of buildings that are grouped based on architecture atmosphere outside of buildings we already know of. Part of this is because there are no established definitive classification of architectural atmospheres. However, as mentioned before, architecture atmospheres is a property that emerges from measurable elements. Because foundation models are good at processing high dimensional data (images) at scale in a consistent way, its vector embeddings can be used to create a 2D plot that could be used to classify and navigate architecture atmospheres in a meaningful way.

This research attempts to examine the effectiveness of foundation models in architecture atmospheres through four experiments. The first experiment is to evaluate the effects of perplexity (a parameter in the dimensionality reduction algorithm), the second experiment is to evaluate the effects of using different foundation models, the third experiment is to tackle architecture atmosphere using words and captions, lastly, the fourth experiment is to judge the outcome of precedent finding, using the method itself.

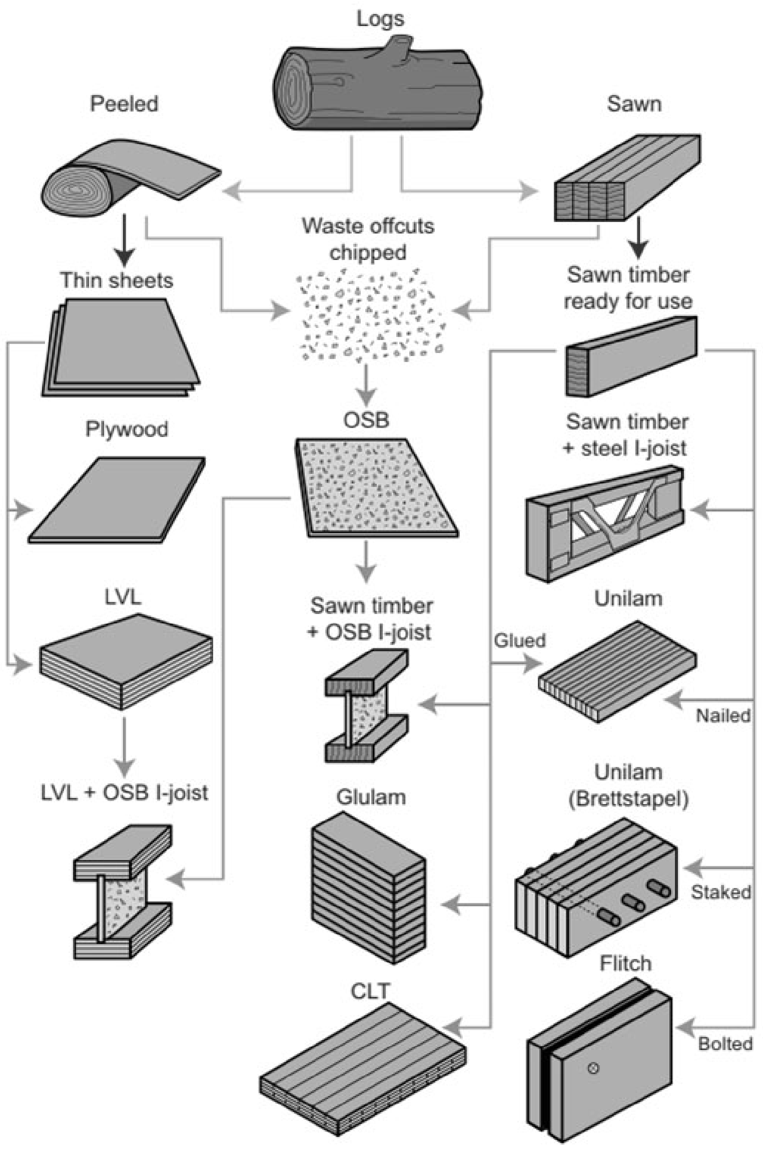


Figure 9. Classification of the different forms timber can take. This kind of classification and relations is something that does not exist for architecture atmosphere. The different kinds of atmospheres cannot be easily listed. Therefore, how closely related one atmosphere to the other does not exist in a taxonomical representation.

Dataset

Sub question 2: How can we effectively collect a large dataset of images that visually convey atmosphere?

In order to get a relevant dataset for the experiment, the images are collected from two architecture websites: Archdaily and Divisare. The two are chosen because they have categories that are of interest to the research. In Archdaily, the advantage is that the website is categorised by building use such as ‘religious buildings’, ‘coffee shops’, and ‘libraries’ (Figure 10). Archdaily is also relatively more popular as it is free to use and the content is often submitted by the architecture practice instead of the website collecting the projects.

Divisare was chosen as a secondary source because despite being a much more curated website, it has a lower amount of projects. In addition to elemental categories, Divisare has categories that are much more specific to architects, such as ‘curves’, ‘building in historical context’ and most importantly ‘daylighting’ (Figure 11). The disadvantage with Divisare is that it requires logging in to access the website, and is therefore more difficult to automate image collection for.

Pinterest is another website which is common to use in architecture. However, Pinterest’s image dataset is broader as it does not only specialise in architectural images, making it difficult to set up a focused and automated image collection method for.

In order to automatically collect data from the two websites, two different methods of collection were made. They are both done in Python using the Playwright library, to automate web browser activities, and the Beautifulsoup library to parse through the html files.

Divisare displays approximately 20 images (projects) per page. For the ‘daylighting’ category, there are 18 pages. Using Playwright, it is possible to go the page link and iterate through 1 to 18 to make sure to visit every page. For every page opened, it can be configured to wait 2 seconds so that all the elements load properly, and then download the HTML file. Once the HTML file is downloaded, the Beautifulsoup library can be used to process the HTML file and search for the link to the thumbnail of the project, which is usually found under a consistent class determined by the website (visible using the ‘inspect elements’ tool in a typical browser). For Divisare, the thumbnail link proved to be enough, as it actually linked to an image large enough for the use of this research. Once the link is found, it can be downloaded and saved to a folder.

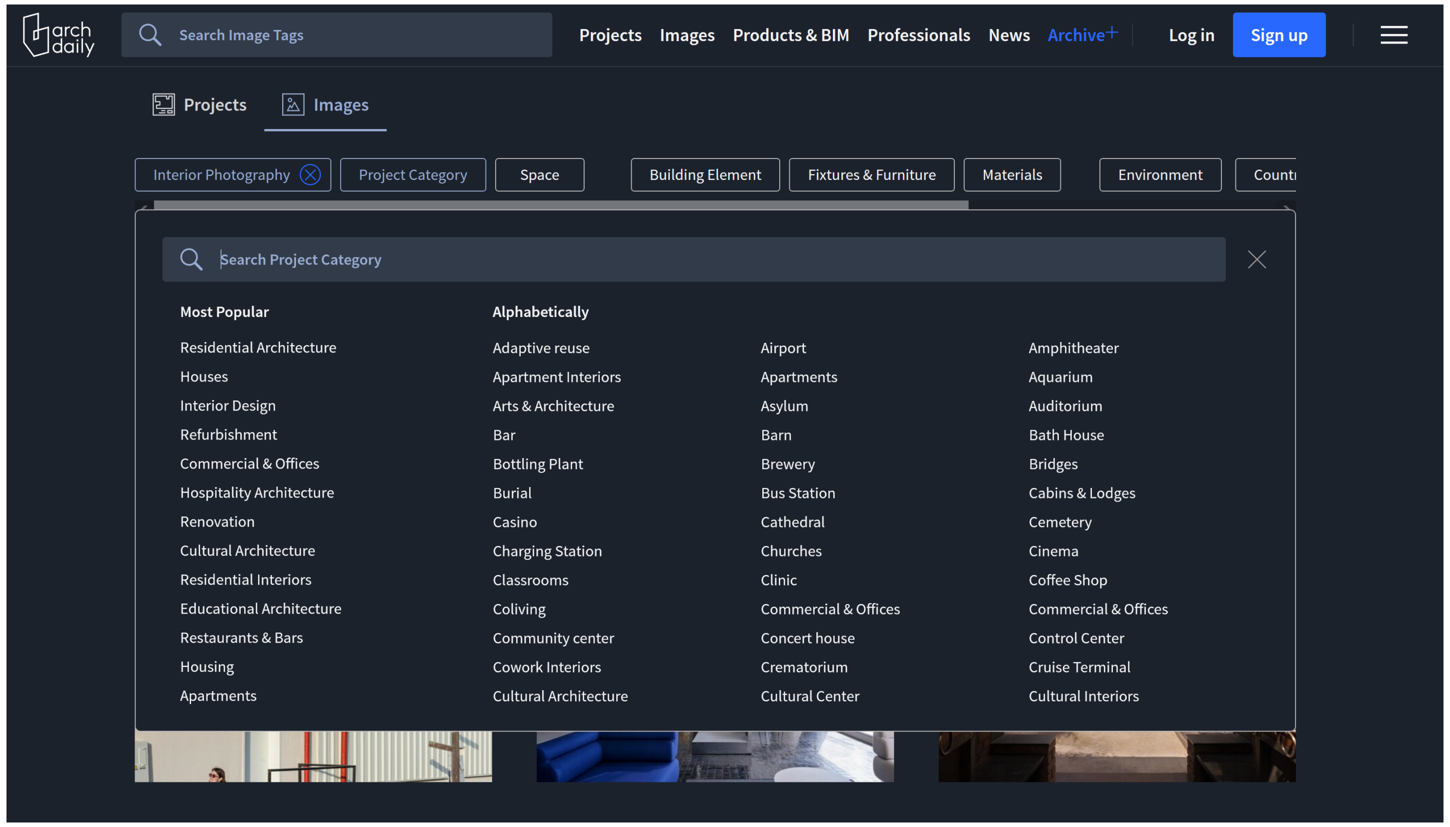


Figure 10. Archdaily available categories

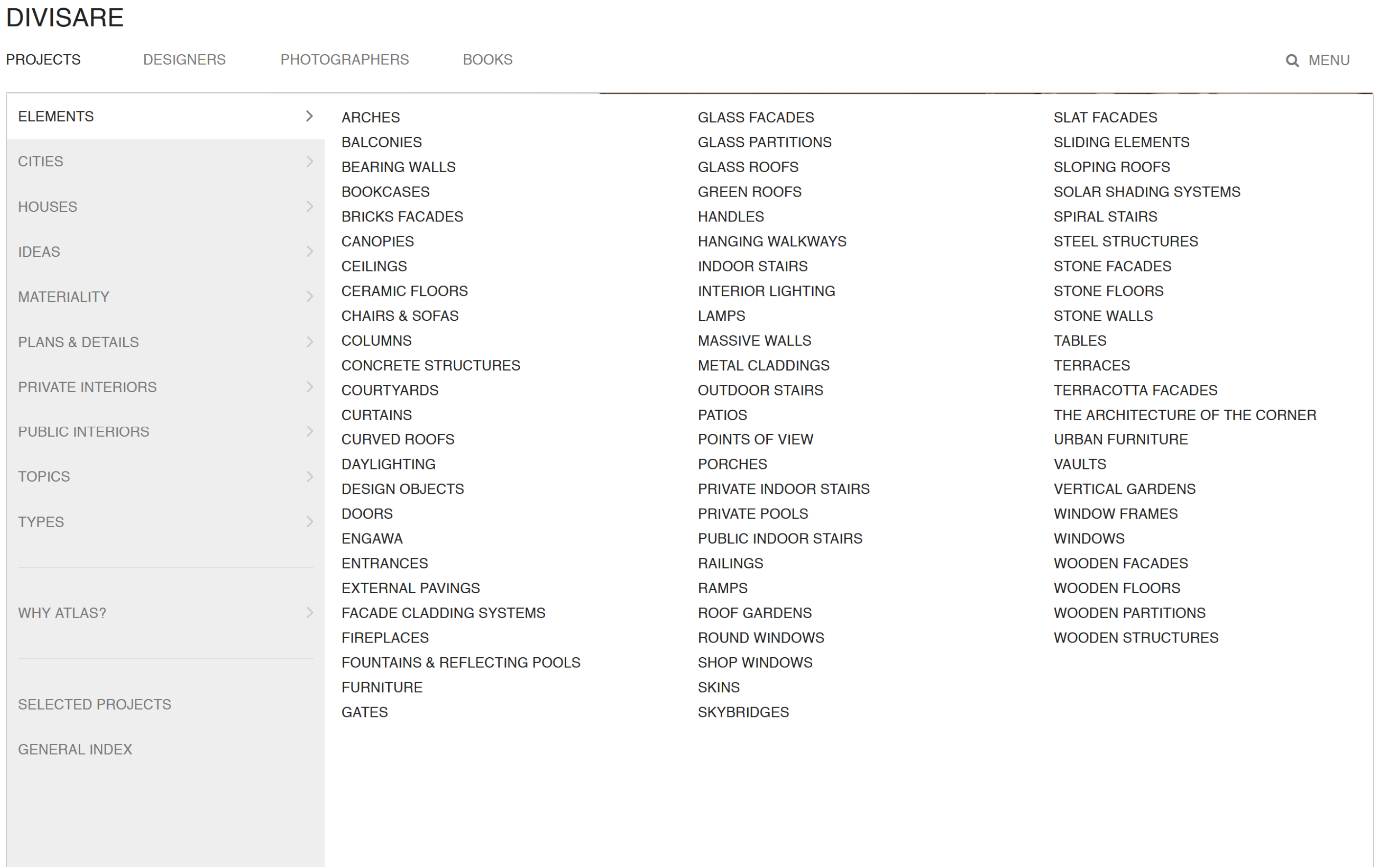


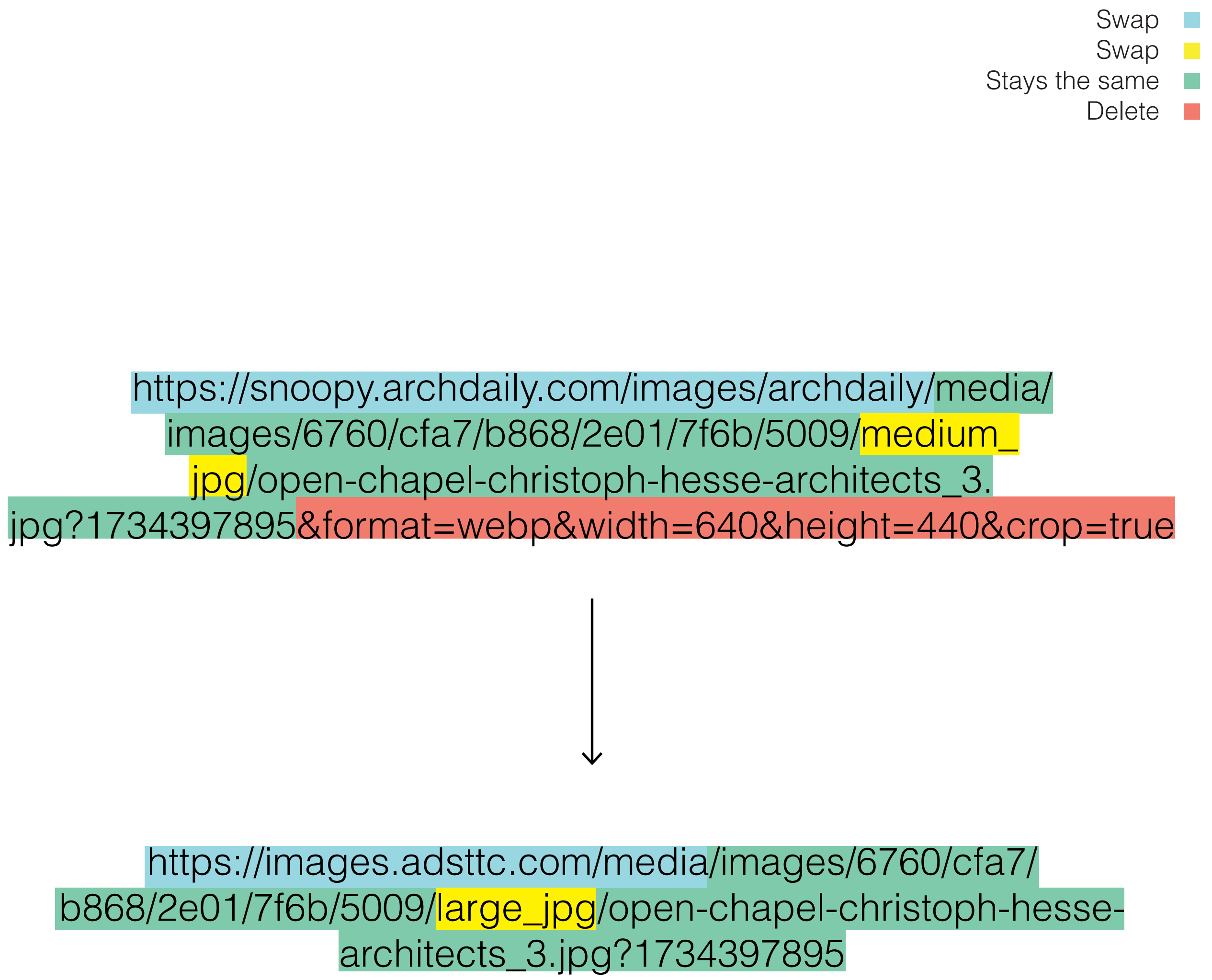
Figure 11. Divisare available categories

Dataset

For Archdaily, because the website uses infinite scrolling on the website, there are no pages in the website link (users simply scroll down to load more content). In order to load as much of the desired content, Playwright was used to automatically press the 'End' key on the keyboard, bringing the page scroll all the way to the bottom, triggering the website to load more content. It is then configured to wait 2 seconds for the content to load before pressing the key again. This can be configured as many times until the images (projects) loaded reach a satisfactory amount. Then the HTML file can be downloaded and passed to the Beautifulsoup library. Again, the link for the thumbnail can be found under a specific class determined by the website.

However, with Archdaily, the link for the thumbnail actually points to a small image. After comparing the thumbnail link and the full sized link in Archdaily's interface, the two are actually similar. Archdaily hosts the full size image elsewhere, but uses the same file structure. Therefore, what needed to be done was to replace the website with their server website, and also replace 'medium_jpg' with 'large_jpg' to create the link for the full size image (Figure 12). After having all the links, the images can then be downloaded and saved in a folder.

The images are mainly collected from Archdaily from the category of 'Burial', 'Coffee', 'Gallery', 'Library', 'Living room', and 'Worship', reaching a total of approximately 2200 images. Another 400 or so images are collected from Diviare's 'Daylighting' category. This brings the combined total of roughly 2500 images after culling duplicate images.



Swap 🟦
Swap 🟨
Stays the same 🟩
Delete 🟥

https://snoopy.archdaily.com/images/archdaily/media/images/6760/cfa7/b868/2e01/7f6b/5009/medium_jpg/open-chapel-christoph-hesse-architects_3.jpg?1734397895&format=webp&width=640&height=440&crop=true

↓

https://images.adsttc.com/media/images/6760/cfa7/b868/2e01/7f6b/5009/large_jpg/open-chapel-christoph-hesse-architects_3.jpg?1734397895

Figure 12. Diagram showing how the Archdaily thumbnail link is transformed to show the full size image link for higher resolution downloading

Method

In order to explore the topic of atmosphere, using foundation models as a tool, I will be attempting to create a 2D representation whose objective is to map different buildings based on atmosphere (a taxonomy of architecture atmosphere). This will be done through foundation models because of its ability of performing this task of classification at a large scale in a consistent way. The steps are as follows:

Data curation

First, I will prepare a dataset of ~2500 architecture photos that display a variety of atmospheres. The method of collection will be both finding images by myself manually, and also setting up an automatic web scraper that collects photographs from websites such as Archdaily and Divisare.

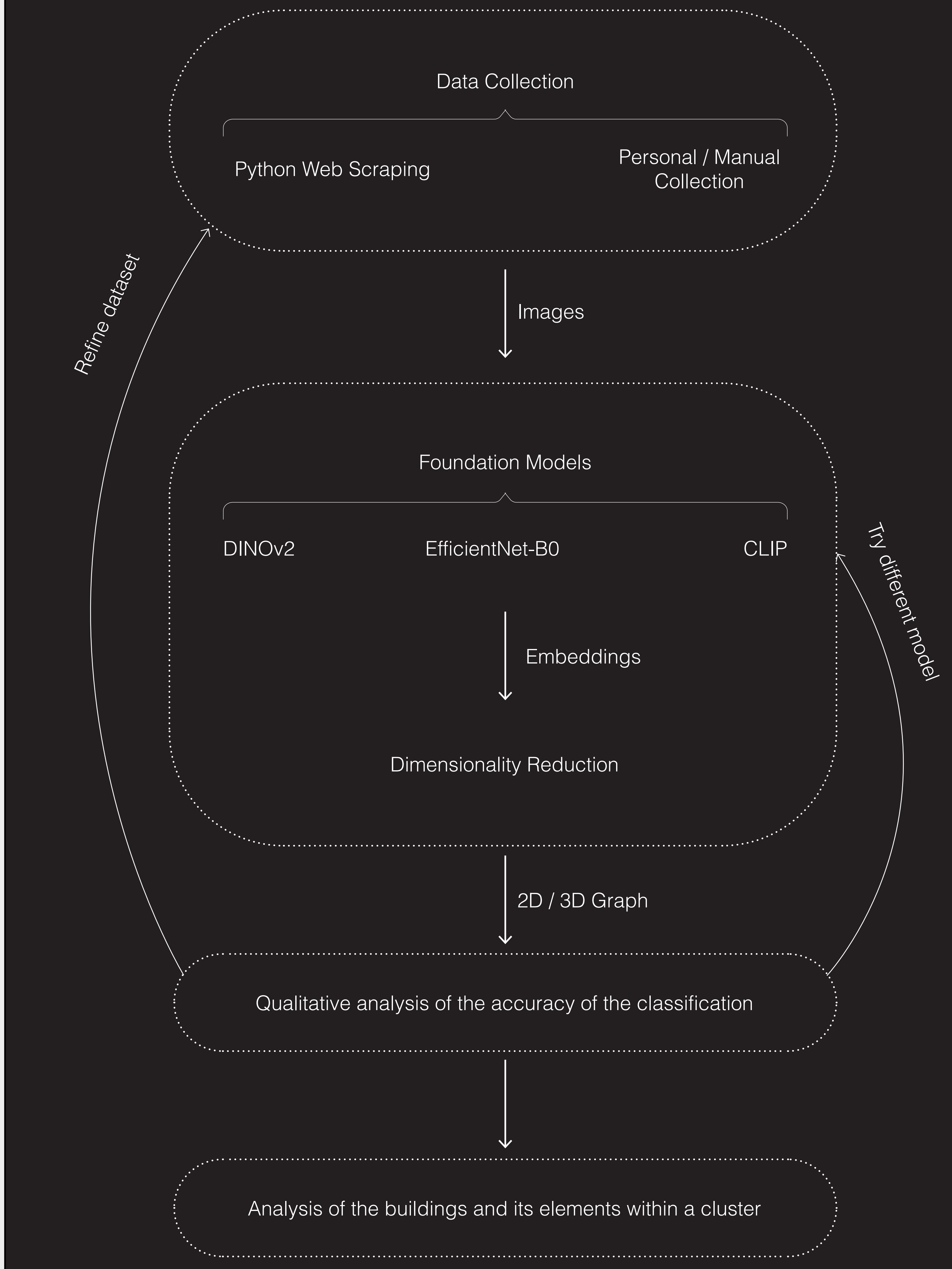
Image processing and visualisation

Second, I will investigate whether the foundation models available today are able to recognise atmospheres from an image and therefore cluster them into different groups. I will use different foundation models such as DINOv2 (Caron et al., 2021), EfficientNet (Tan & Le, 2019) and CLIP (Radford et al., 2021). For each trial, the foundation model would calculate the vector embedding of each data point. These then serve as input to t-SNE, which reduces the dimension of the vector to 2. The reduced vector are then visualised as points in a 2D scatter plot.

Analysis

Third, I will analyse the resulting plot. In this plot, instances that are placed closer together are the images that the foundation model deems are similar. The analysis will be done in two stages. The first stage is to judge the output to determine whether the model is clustering the instances based on atmosphere or not. If not, the experiment will be repeated, changing variables such as the foundation model itself.

Figure 13.



Method

The second stage is when an acceptable result is achieved. The clusters itself will then be investigated, looking for patterns, overlap and trends. This can be done by selecting the cluster of interest, and then creating another plot, with a setting more suited for smaller datasets, providing a more ‘fine grain’ result of the particular cluster. The corresponding buildings of the images that are clustered together will also be analysed to see if there are any architecture elements in common, which contributes to the creation of a specific atmosphere.

Variables

The three main variables of this experiment are: The dataset, the different foundation models, and the hyperparameter in the dimensionality reduction algorithm.

The dataset is a major factor in determining the outcome of the experiment. In an ideal scenario, the images are all taken all from the same angle. The only thing that should be changing would be factors affecting the atmosphere of the space. That way, the outcome of the experiment can clearly indicate the models’ ability to judge architectural atmosphere. However, this is not the case with the current dataset collected through the limited method.

The foundation models are also a factor in the outcome of the plot because they are trained using different methods and training datasets. Therefore, the models will have different outputs when passed the same image dataset in this experiment.

Lastly the perplexity hyperparameter in the dimensionality reduction algorithm is also important to consider. A higher perplexity (such as a value of 50) will be more suitable on a larger dataset as more neighbours are considered and is therefore better at capturing broader pattern in the data. A lower perplexity (such as 10) will result in the algorithm providing a fine-grained local relationships, with less regard of the global structure.

Hypothesis

Due to the nature of foundation models being designed to be a general purpose model, it is expected that the models will take into account a high number of features. Added with the fact that the image dataset is diverse in terms of visual features, the hypothesis is that there will be a more uniform distribution throughout the plot will be observed, due to the models ability to create accurate representations and the uniqueness of the dataset.

4

Results

Experiment 1: Perplexity

The first experiment was done on a smaller dataset. The chosen set consists of 380 images, downloaded from the 'library' category in Archdaily. In this trial, DINOv2 was used, and the perplexity value chosen for the dimensionality reduction algorithm (t-SNE) was set to 10 (Figure 14) on one trial and 50 on another (Figure 15). From these two trials, we can see that t-SNE works quite well with a lower perplexity when the dataset is smaller. With a lower perplexity, clusters are more defined, making it easier to investigate which of the images are closely related and analyse it semantically.

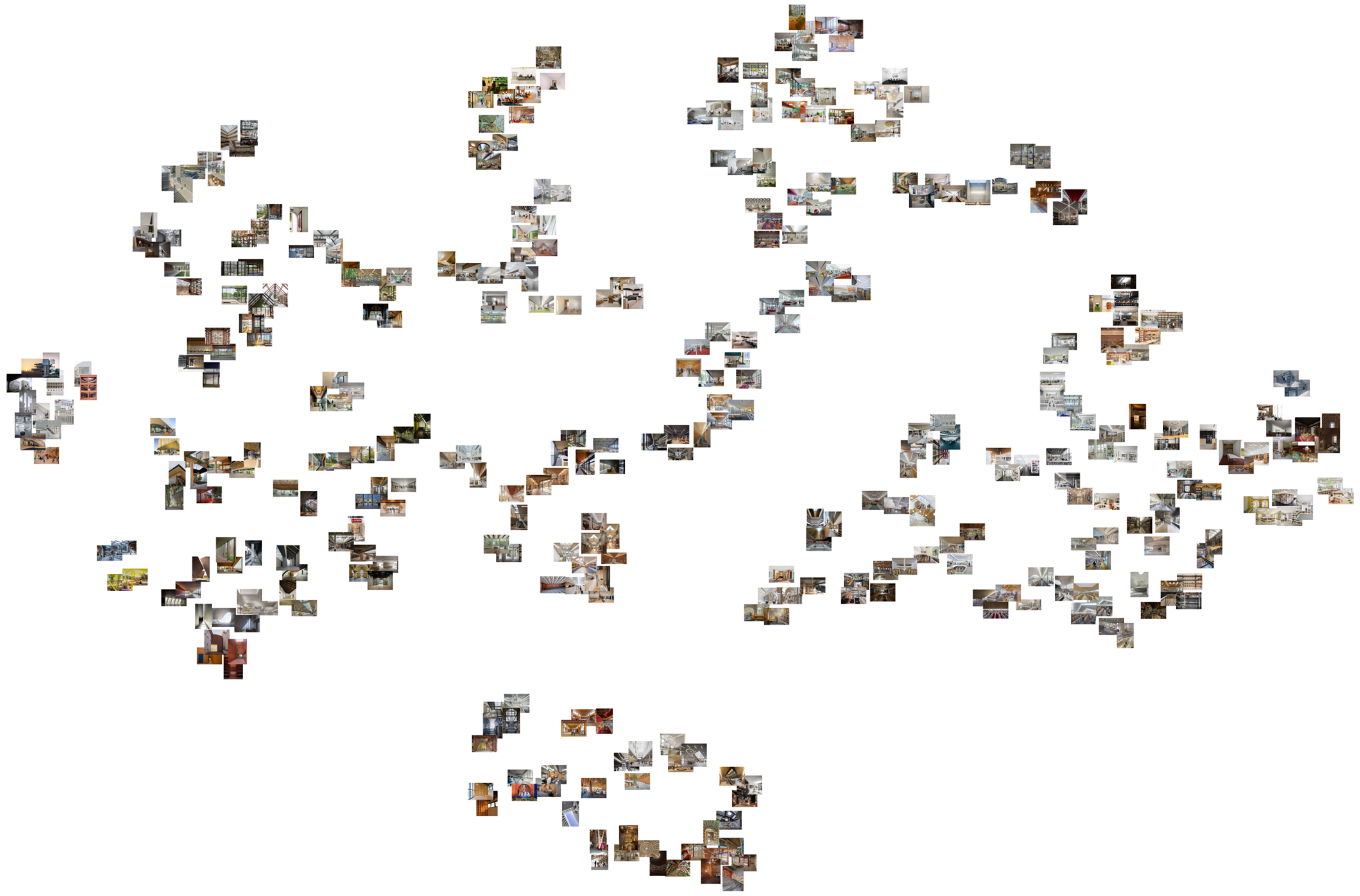


Figure 14. DINOv2 2D scatter plot at perplexity = 10 (low), library dataset (380 images)

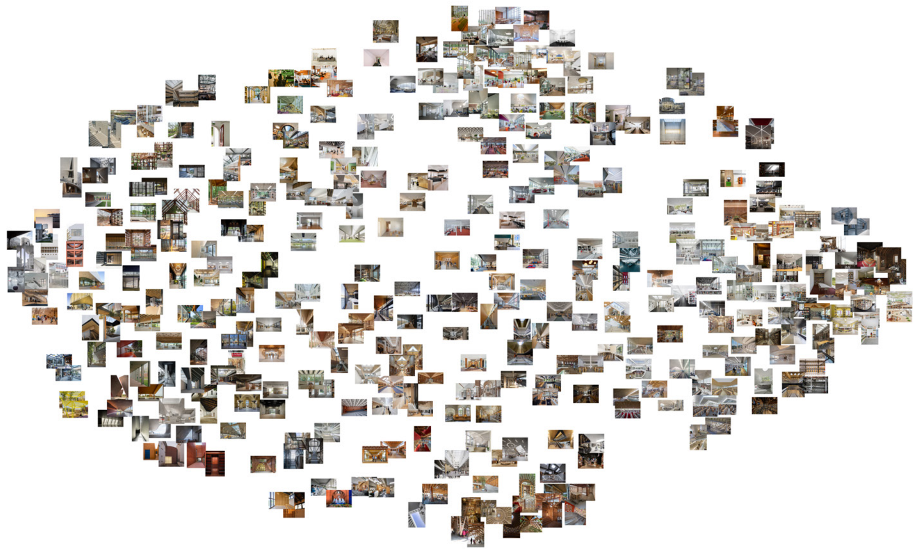


Figure 15. DINOv2 2D scatter plot at perplexity = 50 (high), library dataset (380 images)

Experiment 2: Different Foundation Models

This experiment is done with EfficientNet-B0 (Figure 16), DINOv2 (Figure 17) and CLIP (Figure 19). It includes the whole ~2500 image dataset. The perplexity for this trial is set to 50. We can see that with this high perplexity and a large dataset, local relationships between instances are still observable (although less explicit), but this plot reveals a global pattern in the dataset.

From investigating the plots in this experiment, it is apparent that all the models display at least some semantic logic in its clustering. CLIP proved to create the most distinct groups, with DINOv2 coming in close second, being slightly less distinct, but still meaningful groups. On the other hand, EfficientNet-B0 creates clusters that are smaller, which might be a result of having a much more specific and restrictive requirements when fitting images into a group. This could be a result of the size of the vector EfficientNet-B0 produces compared to the other models. DINOv2 vectors has 384 dimensions, CLIP has 512 dimensions, while EfficientNet-B0 vectors have 1000 dimensions. Dimensionality is merely the size of the resulting vector, but in the case of creating this plot, too large of dimensions might be the cause that it is more difficult to find semantically useful clusters.

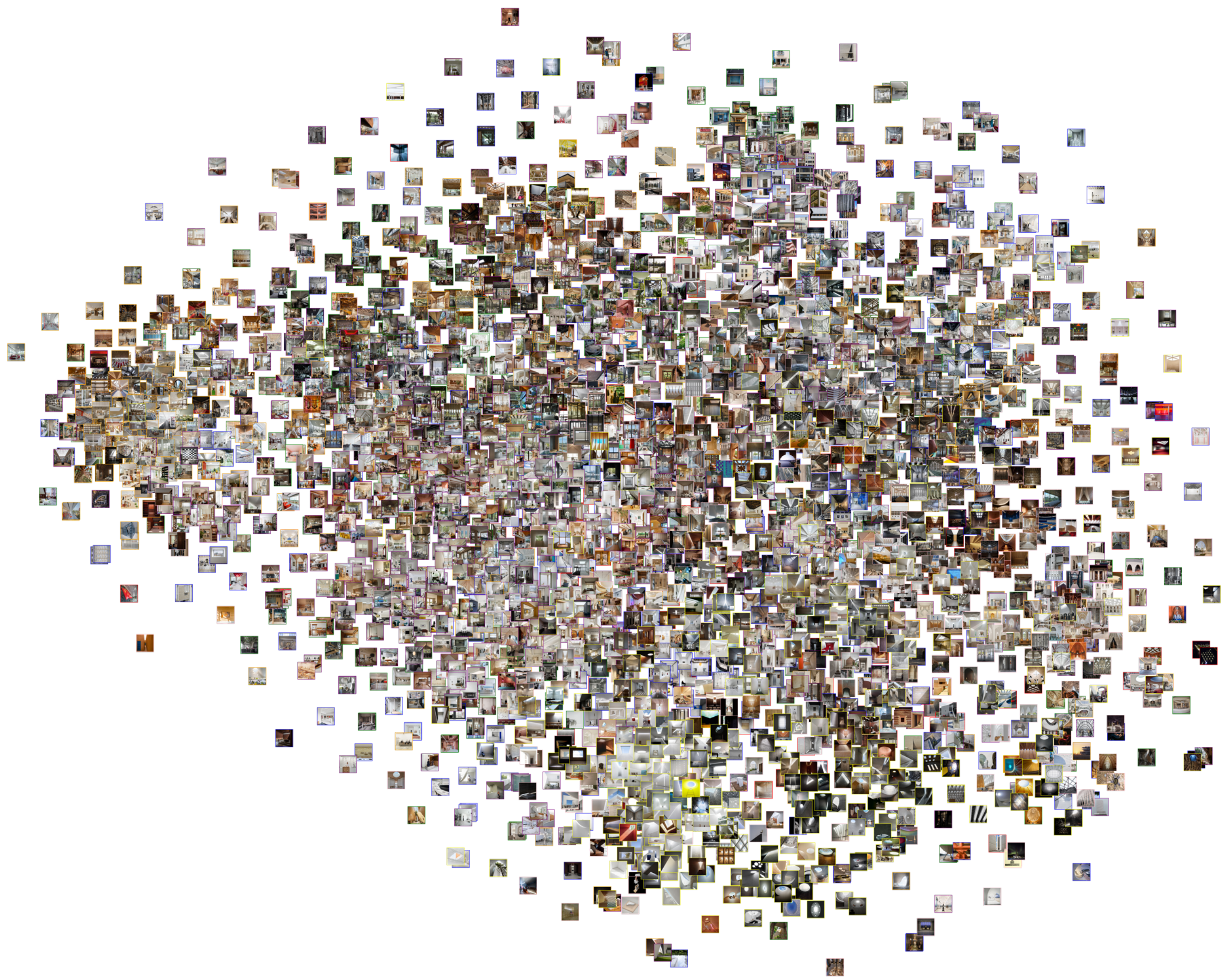


Figure 16. EfficientNet-B0 2D scatter plot at perplexity = 50, full dataset (2500 images)

Experiment 2: Different Foundation Models

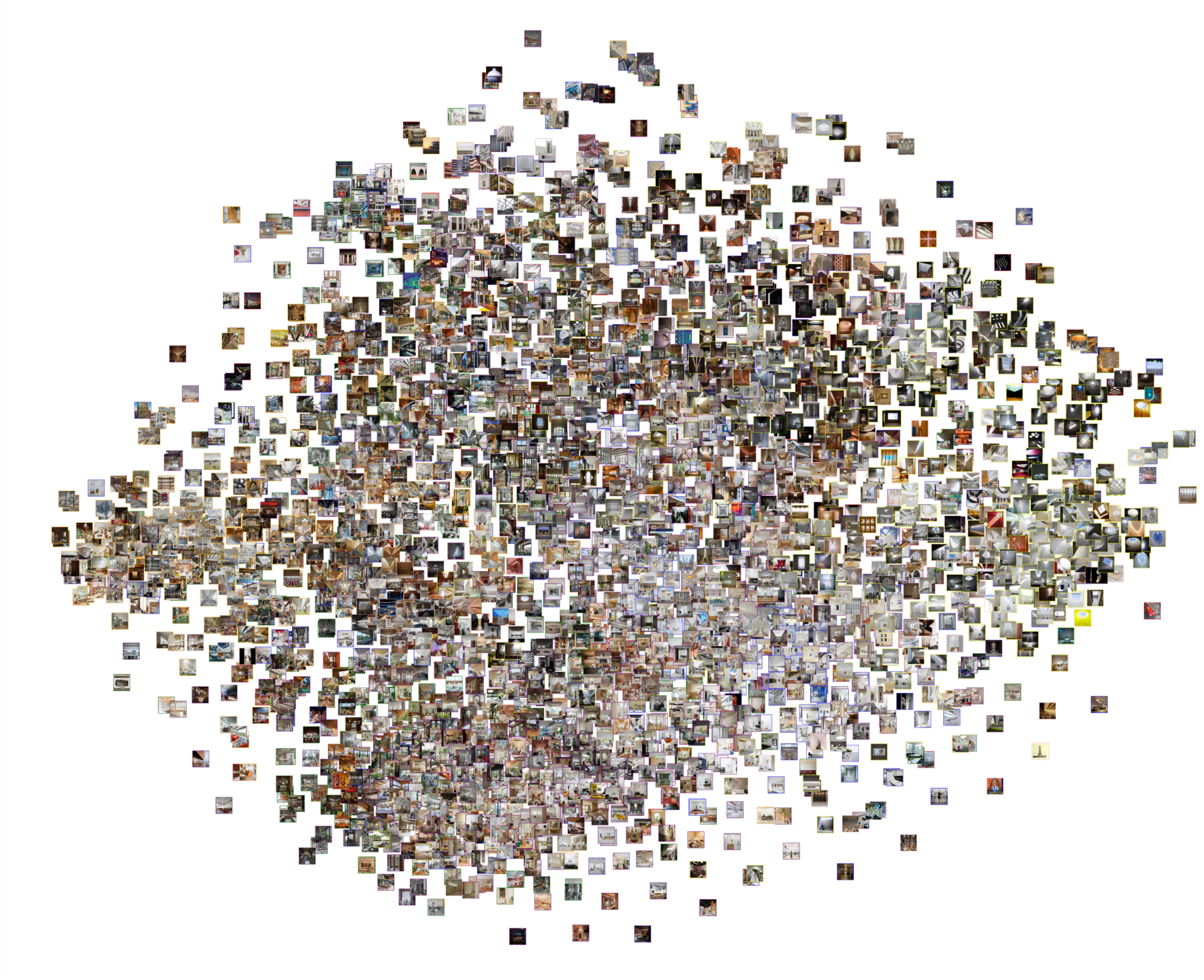


Figure 17. DINOv2 2D scatter plot at perplexity = 50, full dataset (2500 images)

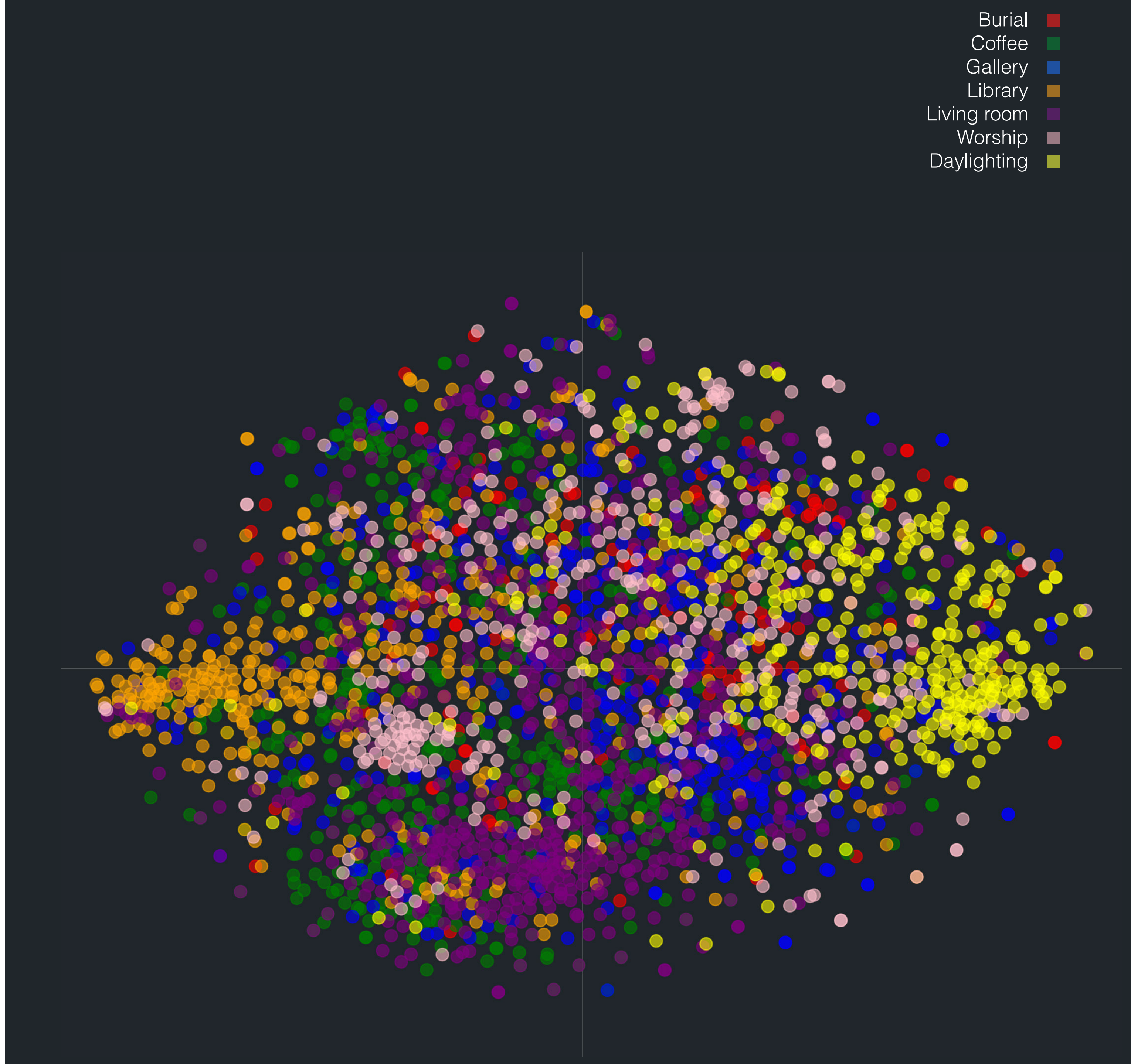


Figure 18. DINOv2 2D scatter plot at perplexity = 50, full dataset (2500 images), data points shown as dots instead of thumbnails, colour coded based on the category the image is taken from

Experiment 2: Different Foundation Models

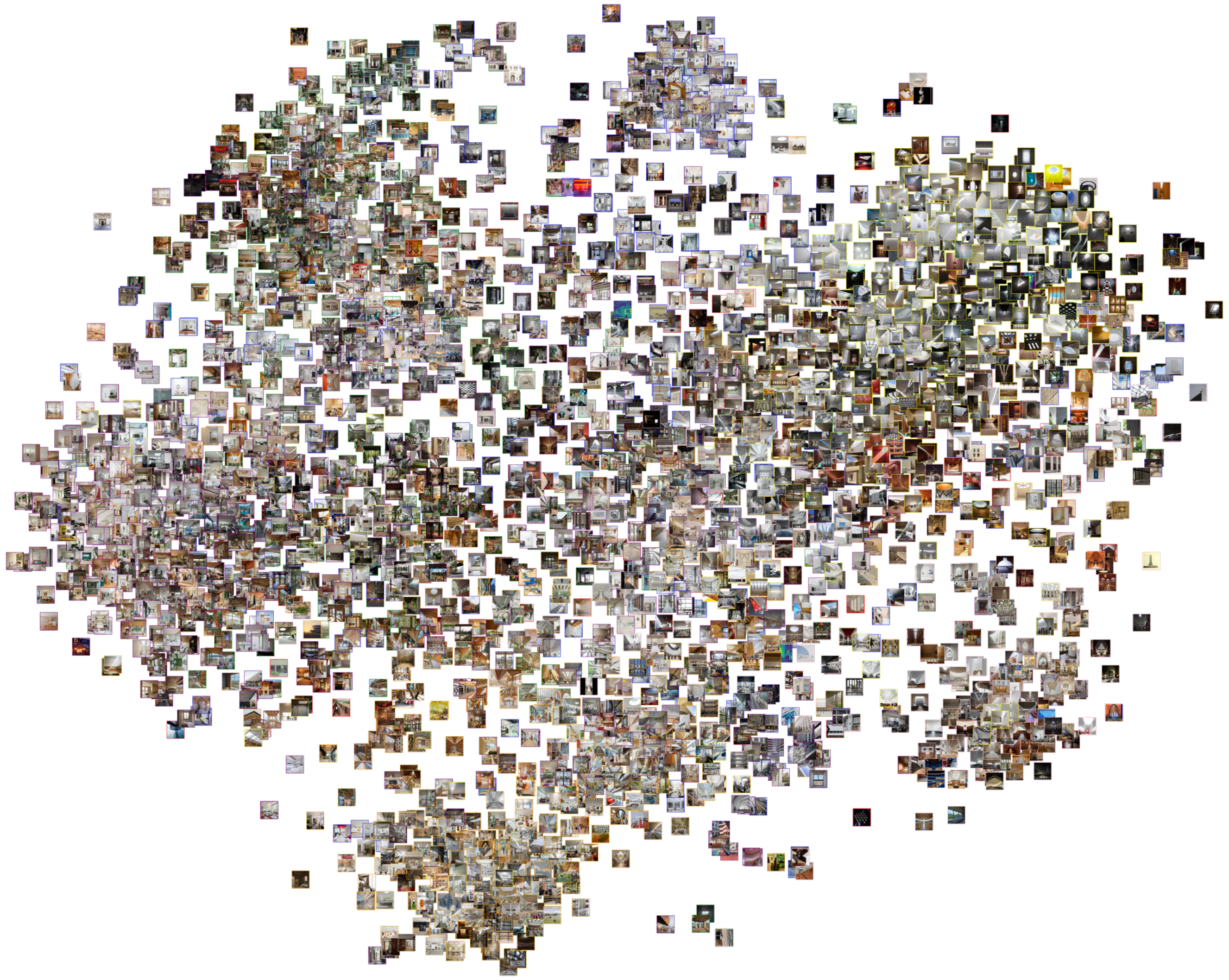


Figure 19. CLIP 2D scatter plot at perplexity = 50, full dataset (2500 images)

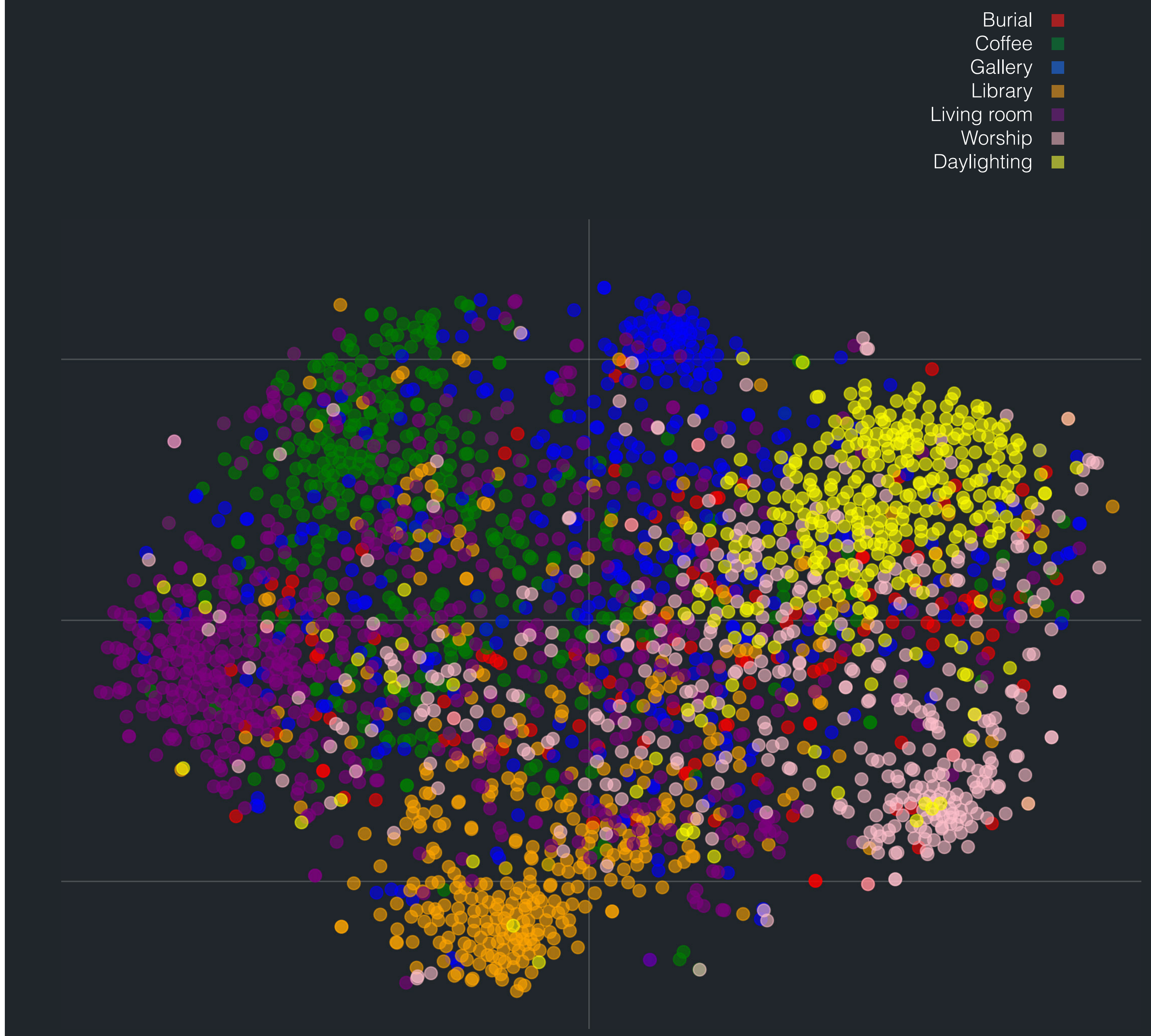


Figure 20. CLIP 2D scatter plot at perplexity = 50, full dataset (2500 images), data points shown as dots instead of thumbnails, colour coded based on the category the image is taken from

Experiment 2: Different Foundation Models

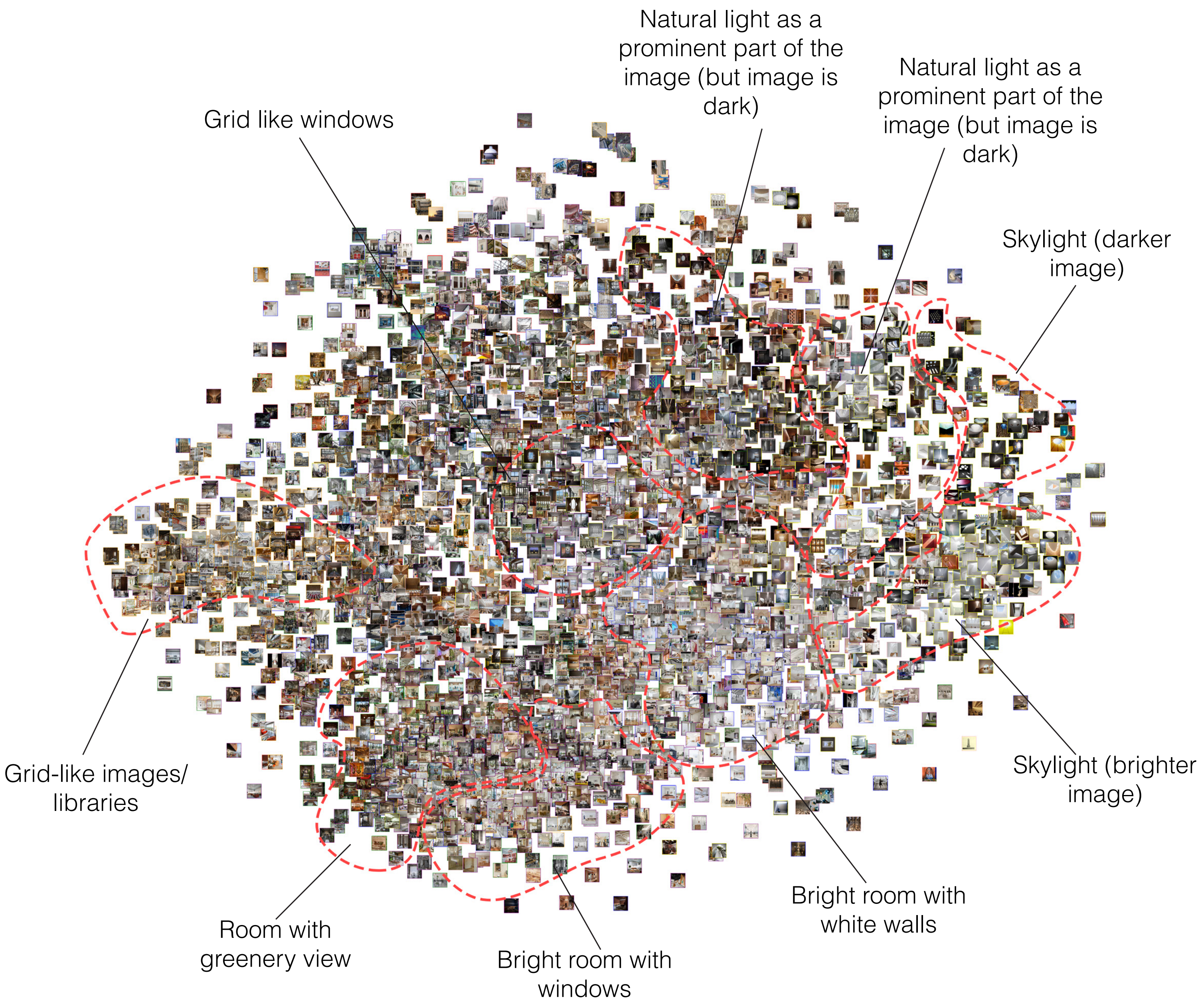


Figure 21. DINOv2 2D scatter plot at perplexity = 50, full dataset (2500 images), annotated semantically

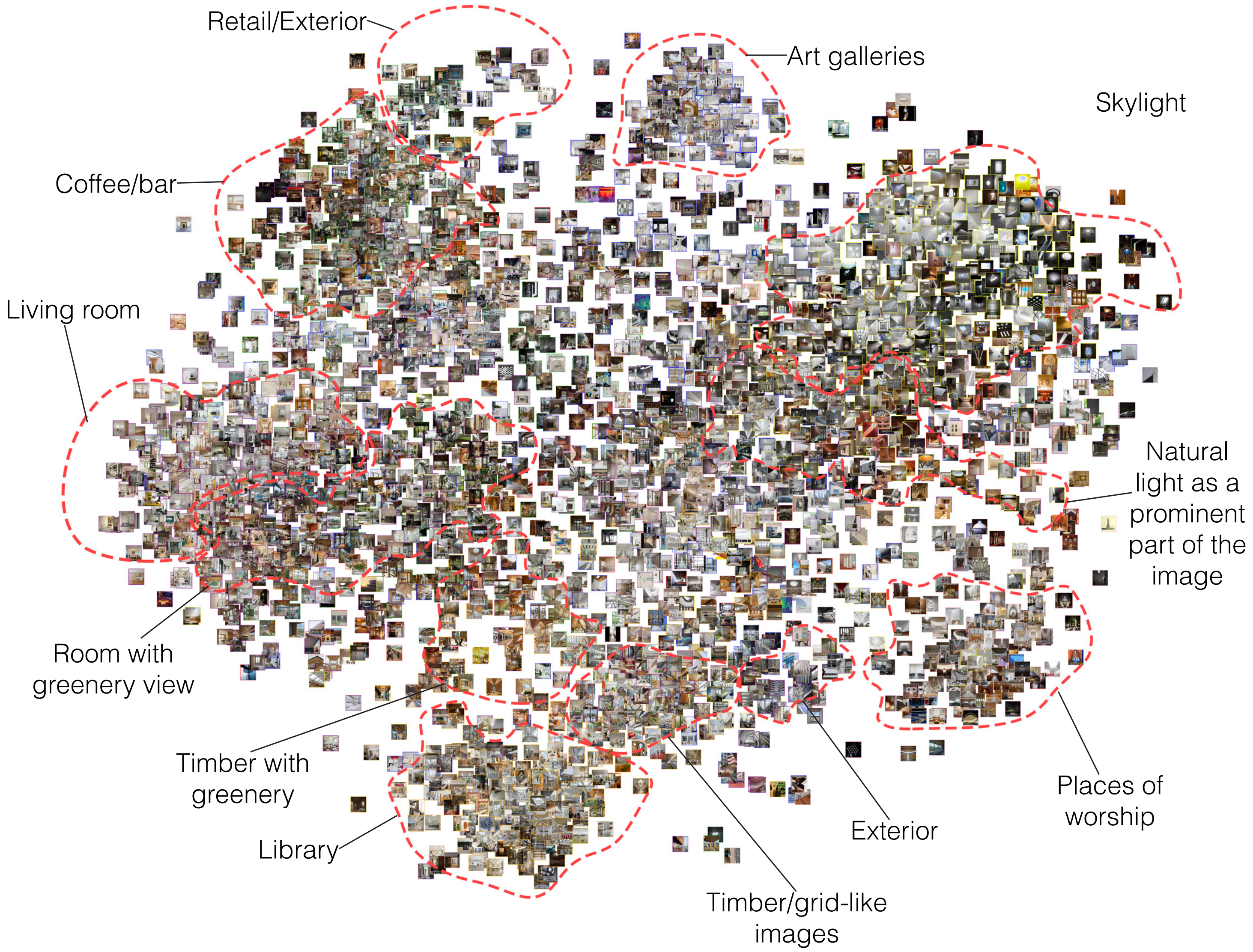


Figure 22. CLIP 2D scatter plot at perplexity = 50, full dataset (2500 images), annotated semantically

CLIP Cluster Analysis

Sub question 3: Can atmospheres be clustered into different groups? What are the main groups?

Semantically, we can also analyse the plots and speculate why certain images are placed closely to each other (Figure 21 & 22), while also relating them to the concepts of Peter Zumthor. In the CLIP plot, on the top right region, we can see a lot of yellow bordered image (Figure 23). The right side of this region is mostly images of circular skylight, whereas the left side is more of linear and angular shaped skylight with a streak of sunlight coming in. In this CLIP plot, on the lower left side of the skylight cluster, the images are no longer of only skylights, rather a mix of skylight and windows, but they all seem to feature sunlight as a prominent subject in the picture. This can be considered as an extreme case of 'Light on Things' where the ethereal quality of light is dominant.

Moving clockwise in the plot (Figure 24), there is a group of images of churches (mostly from the 'daylighting' and 'worship' category). Churches are something that CLIP successfully identified semantically. This maybe because of the 'standard' that loosely exists in church layout design and the pictures of it that are taken. Churches often have rows of seating and are symmetrical, leading to photographers naturally having at least one picture of the church from the centre, facing the main altar. This may lead to the abundance of image that are this 'type' making it easier for CLIP to recognise as a class.

These images could also have been looked at from the perspective of 'Levels of Intimacy' and 'Temperature of a Space'. However, it is apparent that these atmosphere related factors are not prioritised by the model. Instead it prioritises the church typology as the class of the image.

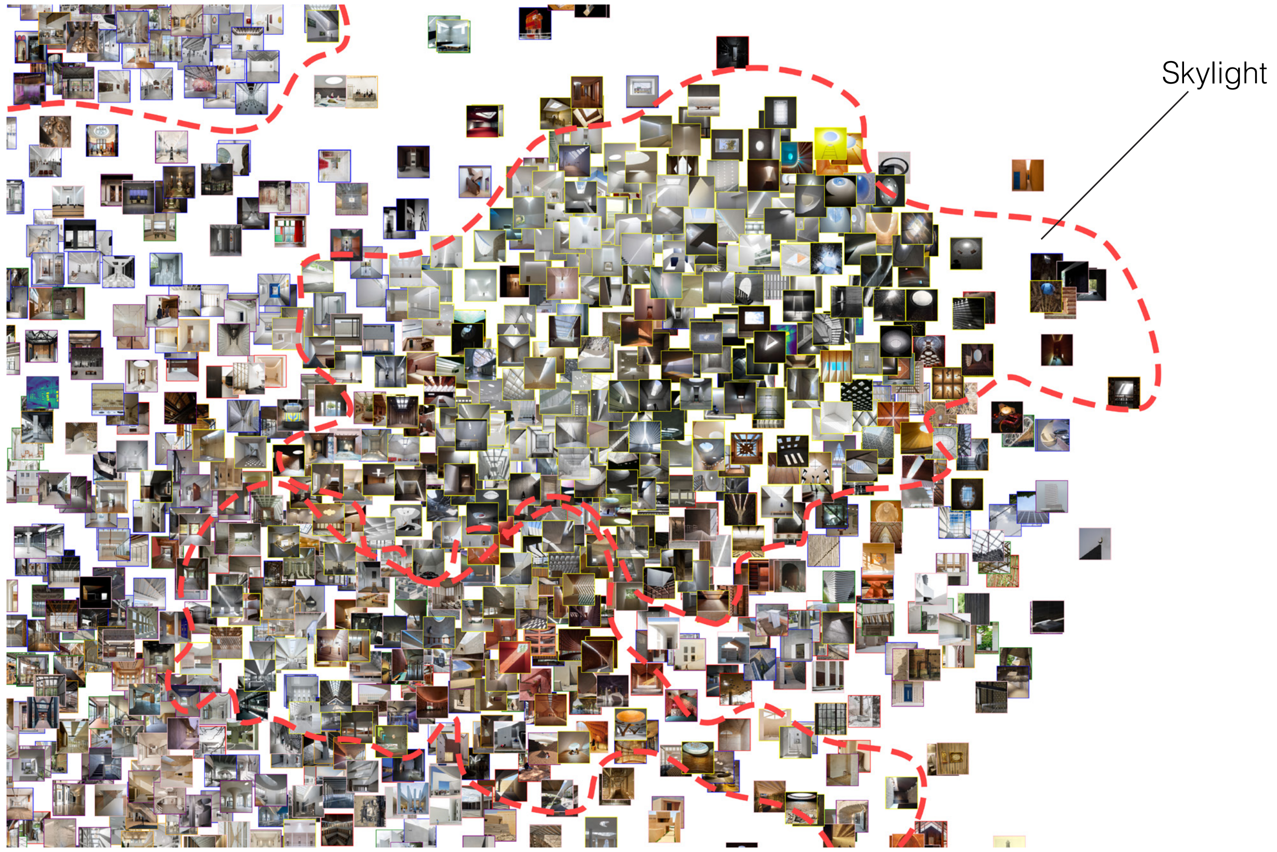


Figure 23. CLIP 2D scatter plot at perplexity = 50, full dataset (2500 images), annotated semantically, top right corner zoom

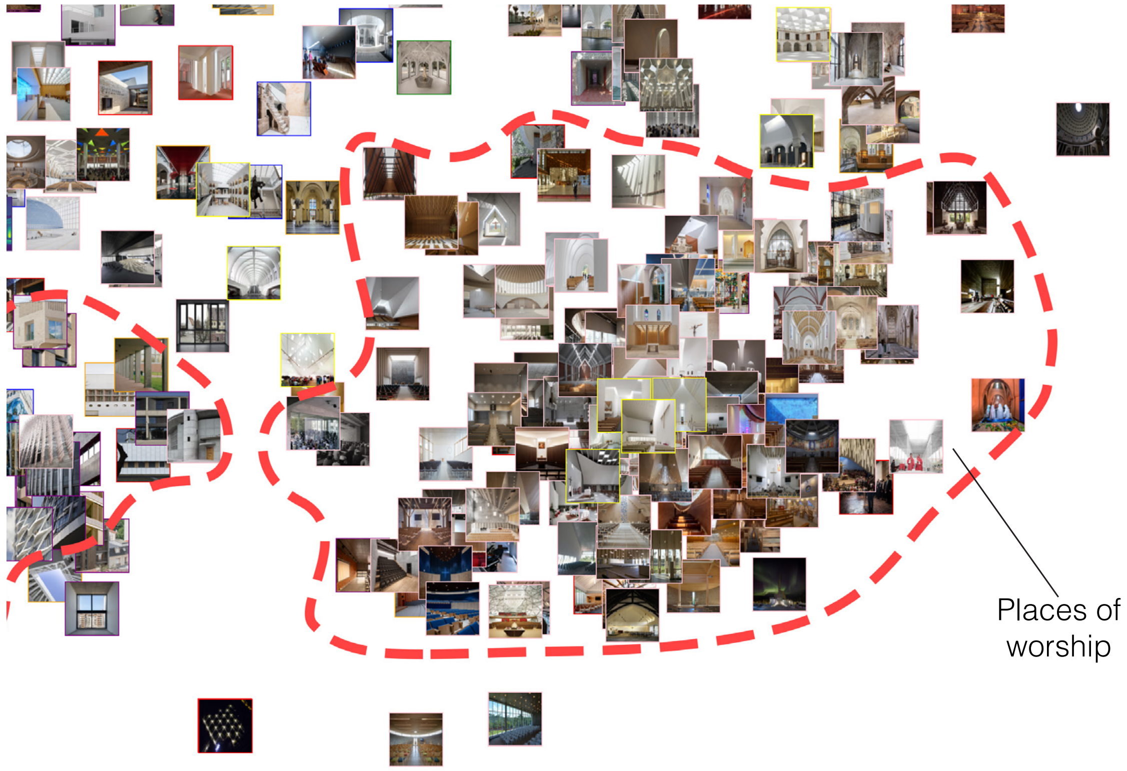


Figure 24. CLIP 2D scatter plot at perplexity = 50, full dataset (2500 images), annotated semantically, bottom right corner zoom

CLIP Cluster Analysis

Next (Figure 25), there is a group of clusters that gradually transform and blend from one to the other. Lower on the plot is the group of images of libraries. This cluster is again something that the model recognises based on its typology. However, near this cluster is a group of images which presumably does not fit into the library typology, but has similar 'primitives', such as having a lot of grid elements (beam and column structures), a lot of timber, and has a medium scale of 2-4 storeys.

This is apparent in the group of exterior images with diagonal grids, floor plates, and façade fins. Next to it there is a cluster with a logic similar to this, but this time it is interior images.

The upper side of the library group is images of libraries with timber elements. Moving further up, it transitions into images with timber elements and vegetation. This then it makes a transition into rooms that feature have vegetation, placed next to living rooms in general.

The proximities of these clusters are interesting as this also happens close to the 'skylight' cluster, where if the model does not recognise the image as an image of a skylight, but has a similar features, it places these image just outside of the cluster.

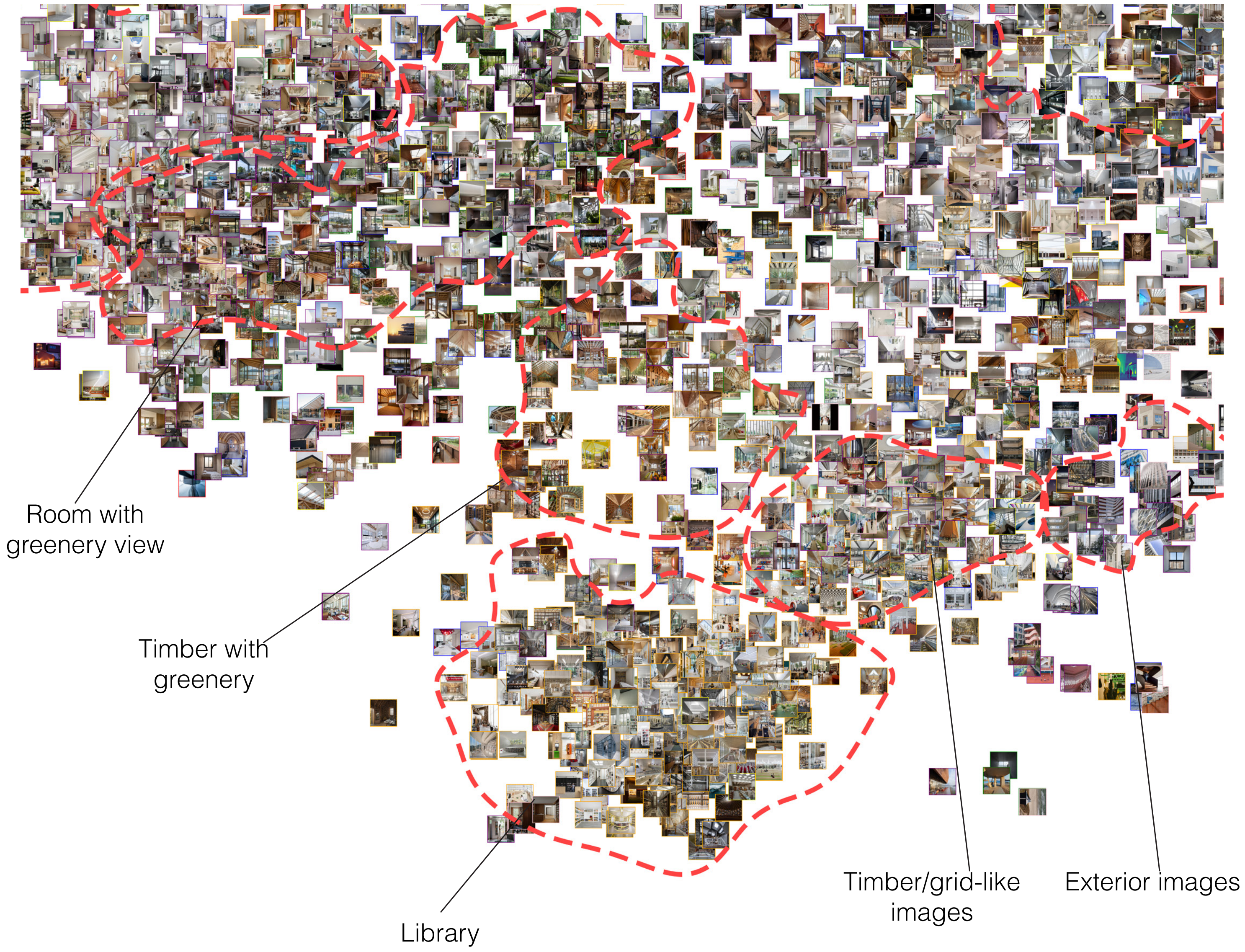


Figure 25. CLIP 2D scatter plot at perplexity = 50, full dataset (2500 images), annotated semantically, bottom part zoom

CLIP Cluster Analysis

Moving further in a clockwise manner brings us to the group of café images (Figure 26), with the lower part of it (closer to the living room cluster) showing cafes that are more cosy and domestic-looking, whereas the upper part of the cluster is much more commercial. Naturally the cluster close to this are images of shop-fronts, retail spaces and the exterior of galleries. Finally the last distinct group CLIP plotted is the art galleries group.

The results suggest that the model is able to pick up features relevant to atmospheres, but is more biased when an image can be mapped to known classes such as 'skylight', 'libraries', or 'churches'. To get the model to give the result desired in the 2D plot, a more elaborate modification of the process is needed.

However, this current classification is already promising as it recognised the 'skylight' cluster, which is not a typical class and is directly related to what Zumthor describes as 'Light on Things'. Furthermore, when not classifying based on typical classes, the model classifies based on texture and colour, which links to materiality and light. This is related to Zumthor's 'Material Compatibility', 'Temperature of a Space' and 'Light on Things'. The model also recognises patterns quite well, such as images with a lot of grids (beam and column structures) from images that are more smooth (monolithic load-bearing structures), closely linked to the 'Body of Architecture'. Lastly, the model recognises 'Surrounding Objects' quite well, something often not given a lot of attention to in architecture, but is obvious to the model.

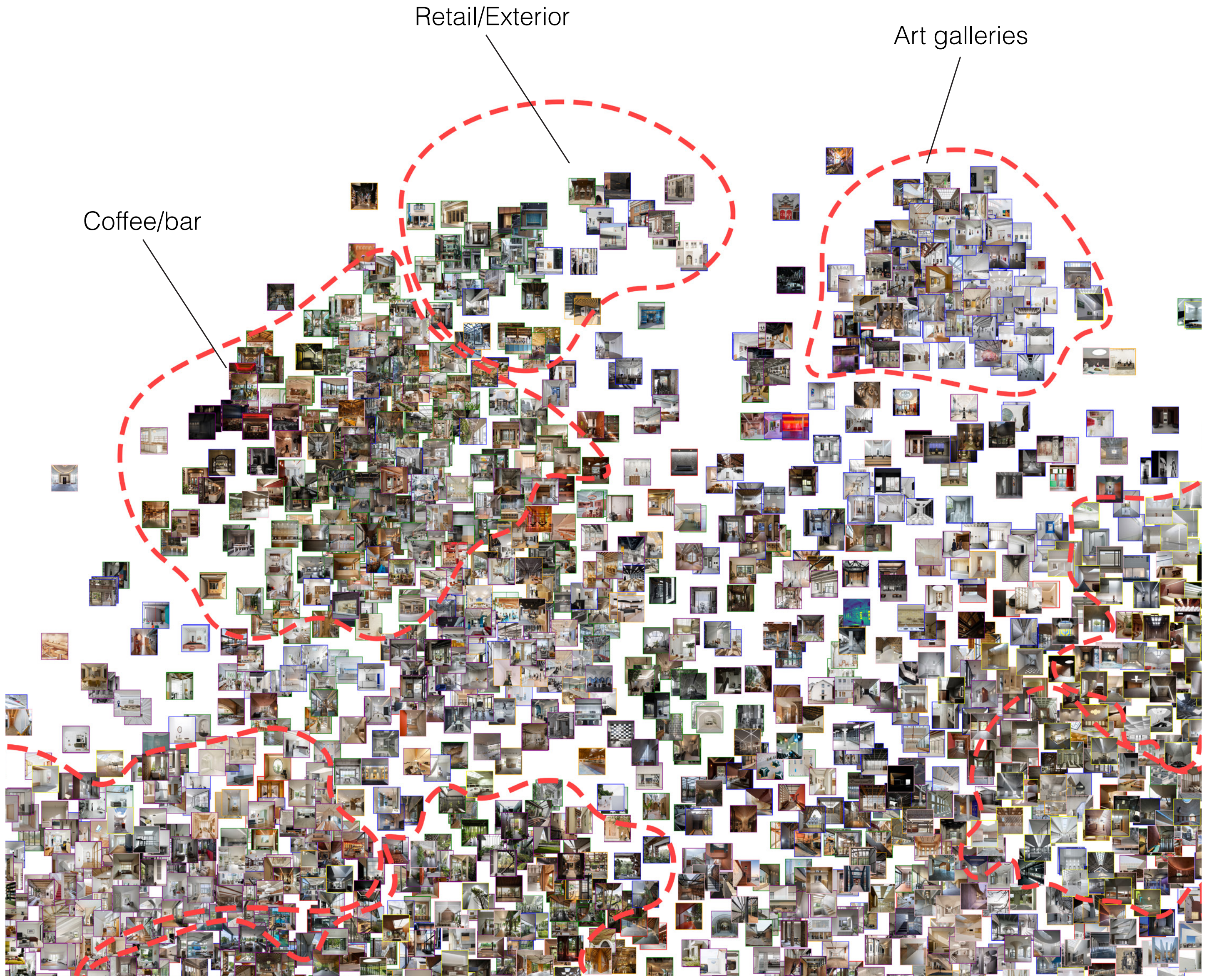
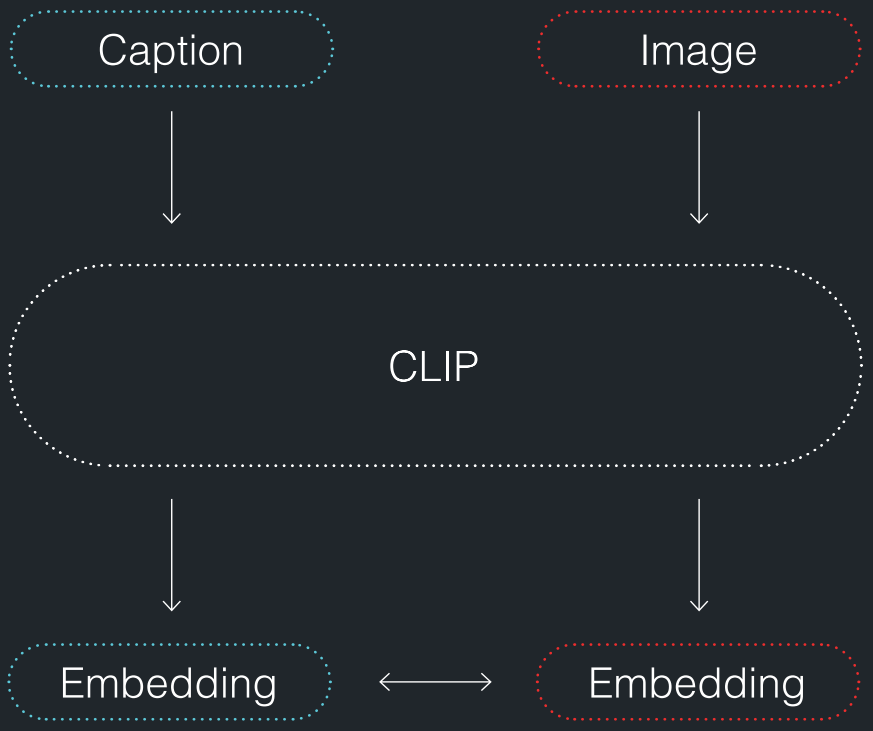


Figure 26. CLIP 2D scatter plot at perplexity = 50, full dataset (2500 images), annotated semantically, top left zoom

Experiment 3: CLIP with Captions



Caption

Image

CLIP

Embedding

Embedding

Figure 27. CLIP caption and image diagram

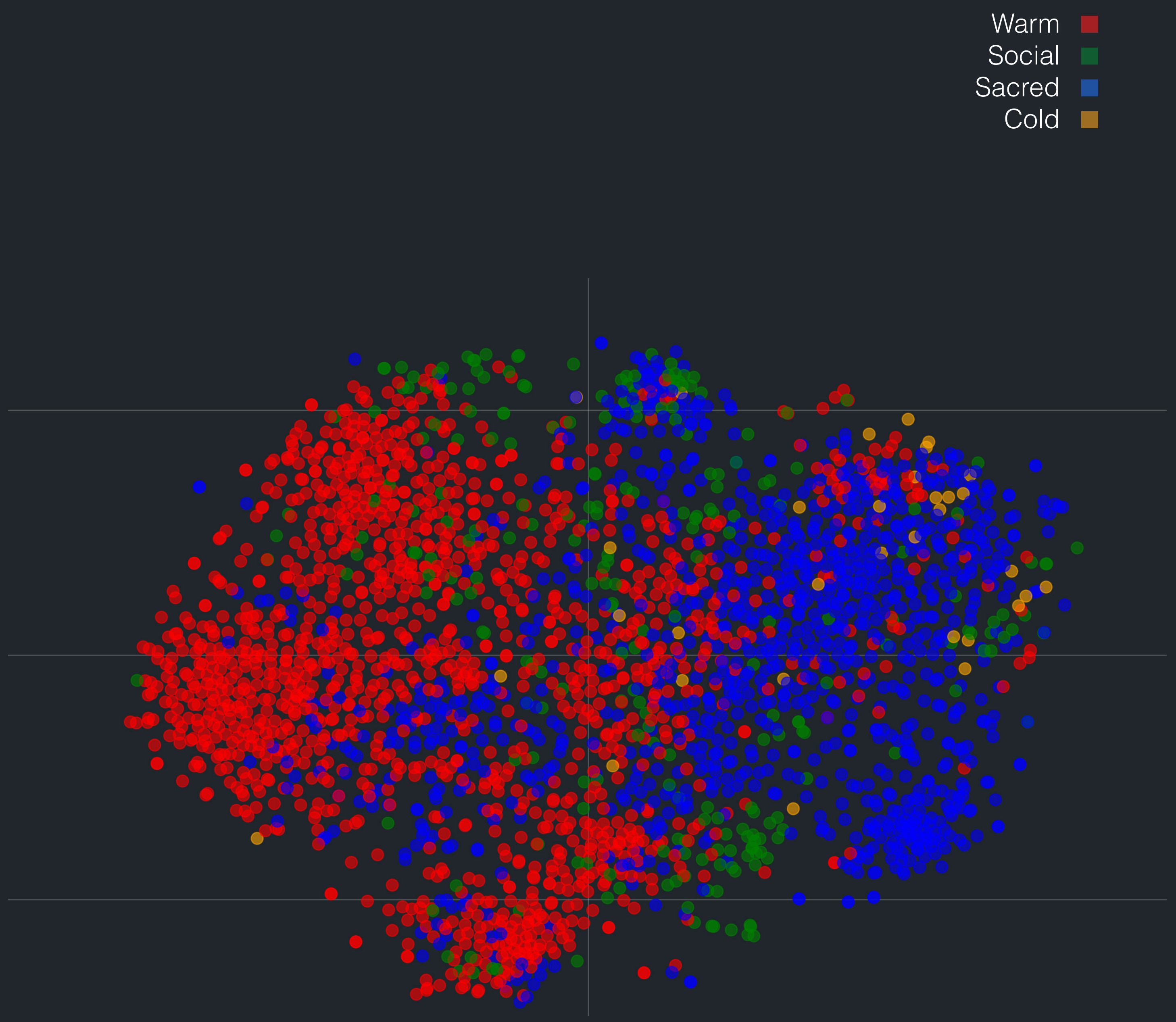


Figure 28. CLIP 2D scatter plot at perplexity = 50, full dataset (2500 images), data points shown as dots instead of thumbnails, colour coded based on the caption CLIP assigns for each image

The analysis before shows that the models produce a relatively meaningful plot for atmospheres. However, specifically with CLIP, there is the possibility of pushing the models further. It is possible to write a few captions and have the model chose one and pair it with an image that would match the text based on their embedding (Figure 27). With this method, in the first trial, four captions: 'sacred atmosphere', 'social atmosphere', 'warm atmosphere', and 'cold atmosphere' were used. The results were promising, however there are some images that clearly mis-labelled.

Experiment 3: CLIP with Captions

In order to address this, the next attempt added more captions: ‘retail atmosphere’, ‘sacred atmosphere’, ‘sterile atmosphere’, ‘contemplative atmosphere’, ‘cosy atmosphere’, ‘calm atmosphere’, ‘relaxed atmosphere’, ‘warm atmosphere’, ‘serene atmosphere’, and lastly ‘spiritual atmosphere’ (Figure 29). With these captions, the results were better and it was interesting to see the model separate the different images within the same clusters and tagging them with different captions. Comparing Figure 29 to figure 20, It excelled in labelling the images that were in Figure 20’s coffee cluster, tagging the images appropriately with either ‘retail’, ‘cosy’, or ‘calm’ atmosphere. However it struggled with the group of images around Figure 20’s ‘skylights’ cluster. In this group, it mainly tags the images mostly appropriately with the caption ‘contemplative atmospheres’, however, there are some outliers where ‘sterile’ and ‘retail atmosphere’ is being assigned instead.

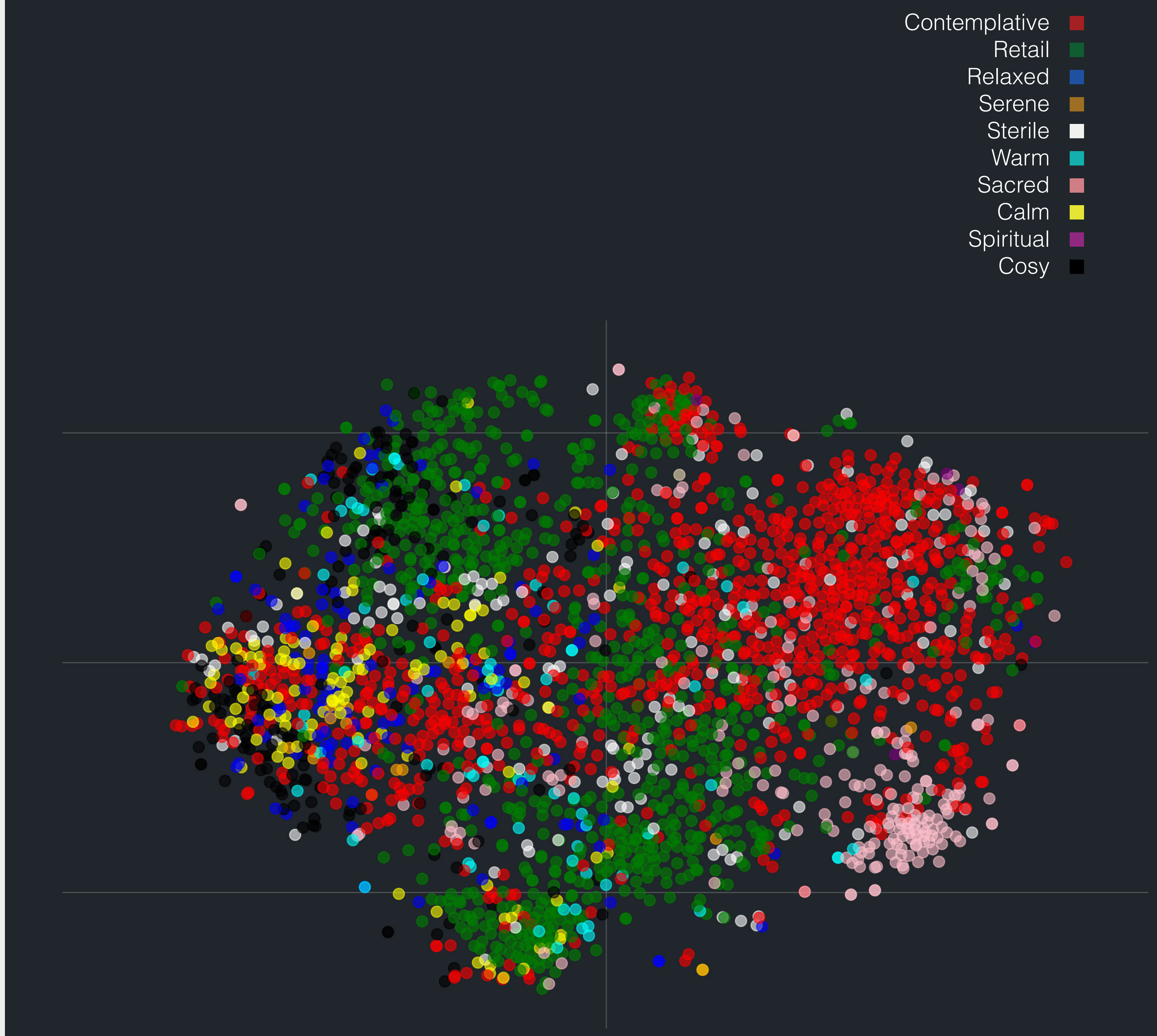


Figure 29. CLIP 2D scatter plot at perplexity = 50, full dataset (2500 images), data points shown as dots instead of thumbnails, colour coded based on the caption CLIP assigns for each image

Experiment 3: CLIP with Captions

The third trial was run with captions that are more elaborate (Figure 30). Since CLIP is designed to generate captions that are more like sentences, the captions chosen were ‘an image of a place that has a transcendental atmosphere’, ‘an image of a place that has a retail atmosphere’, ‘an image of a place that has a cosy atmosphere’, ‘an image of a place that has a sacred atmosphere’, ‘an image of a place that has a serene atmosphere’, ‘an image of a place that has a spiritual atmosphere’, and ‘an image of a place that has a relaxed atmosphere’.

Despite modifying the captions, it can be seen in Figure 30 that it tagged some images inaccurately such as Peter Zumthor’s baths in Vals, which was tagged as ‘an image of a place that has a retail atmosphere’.

Comparing with the original ‘classes’ from Figure 20 with Figure 30, the outcome was still not accurate within the Figure 20’s ‘skylight’ cluster. With these images, CLIP struggles to find dominant features to properly assign an accurate caption for the image. These were also images that were hard for myself to provide captions for and are therefore images of interest. Several attempts have been made in terms of providing different caption so that CLIP would assign it to these specific group of images, however the captions provided never seem to be able to capture the whole set of images accurately. The fact that there are no precise words to capture it, but CLIP being able to place them in a similar location is a promising result in this experiment. It shows that there is a gap in language that can be used to identify it, but visually, there is a consistency that these images share.

This ability of integrating captions to the images provides the opportunity to be more accurate when processing the images. When appropriate captions are found and assigned for the images, it would be possible to train a small model to create a model that includes the captions as a feature when creating an embedding.

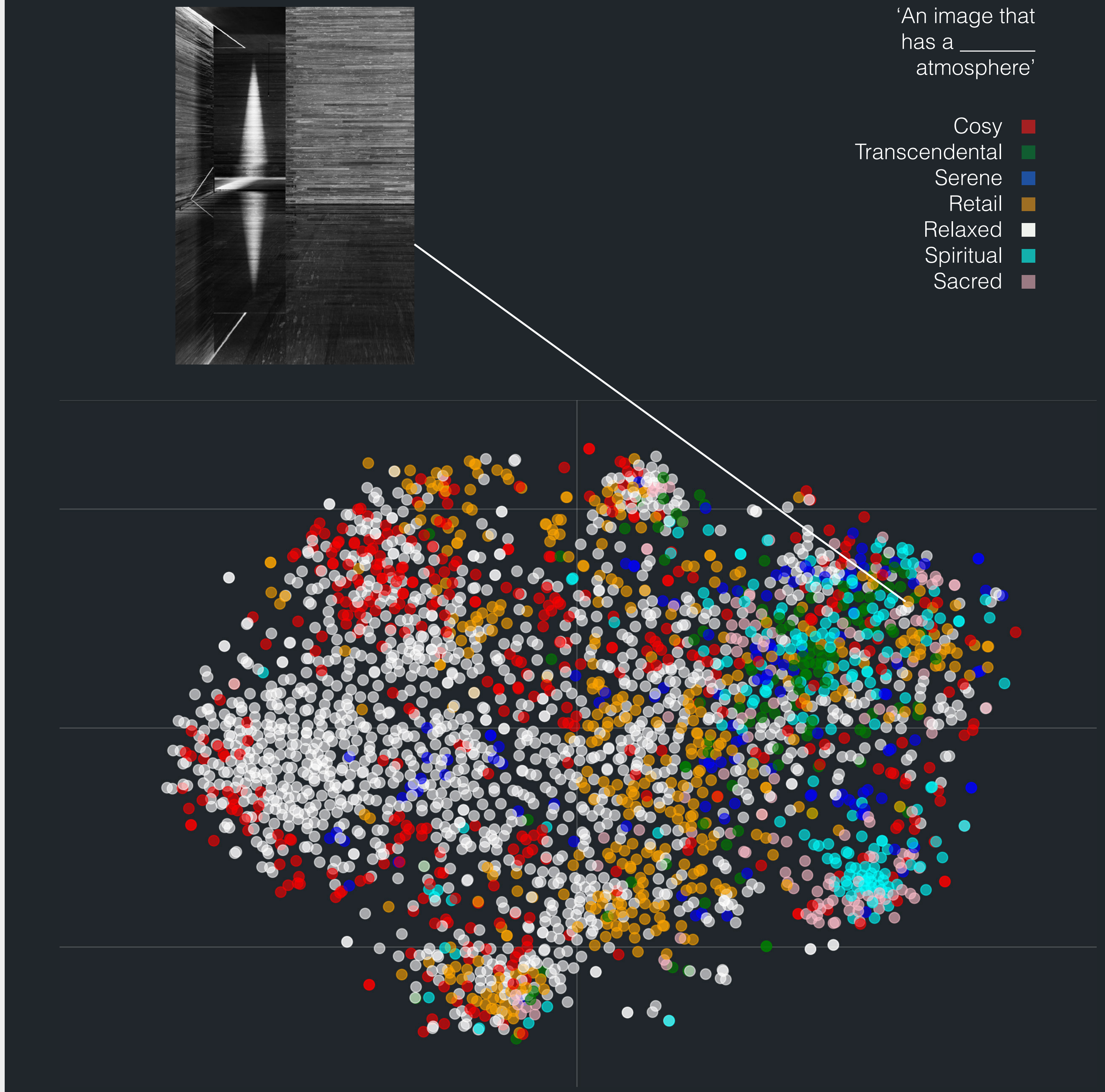


Figure 30. CLIP 2D scatter plot at perplexity = 50, full dataset (2500 images), data points shown as dots instead of thumbnails, colour coded based on the caption CLIP assigns for each image. The Therme Vals is wrongly labelled as retail in the graph showing the inaccuracy of CLIP’s label.

Experiment 4: Potential Use-case

In this experiment the goal is to test if the foundation models can be used to find precedents that have a relevant atmosphere. To do that a single reference image is given as input to CLIP, together with the full dataset of 2500 images to be processed as a 2D plot. In this case, it is Peter Zumthor's Chapel (Figure 31). Already from Figure 32-33, we can see that there are similar images located around the image reference in the 2D plot.

However, in order to have a more fine-grained result that focuses more on local distances, the images in the cluster of the reference image is selected (Figure 34) and then used as an input again, this time for DINOv2 (Figure 35-36). DINOv2 is used because after several tries, it shows that it is able to distribute the plot more gradually and is more relevant for this purpose.

With this method, some interesting results arise. From the fine-grained plot, six closest images to the reference image were examined (Figure 37-39). Two of them (Figure 37) are images of building I recognise. However, the other four are images that are new to me. They are images that I would not have found without this method. Three out of the four images (Figure 38 and the first image in Figure 39) are also visually different images, but with the same essence of the reference image. This being deemed as similar by the model suggests that the classification method used is effective for this use-case.

Another use case of the foundation models and the 2D plot is to evaluate the design of a space. When designing a proposal, an architectural render could be made and then used as input for the foundation model, together with the 2500 images. Then, when the 2D plot is made, it would be possible to see if the design proposal falls close to the desired cluster. Furthermore it could be used to create a description of a design proposal.

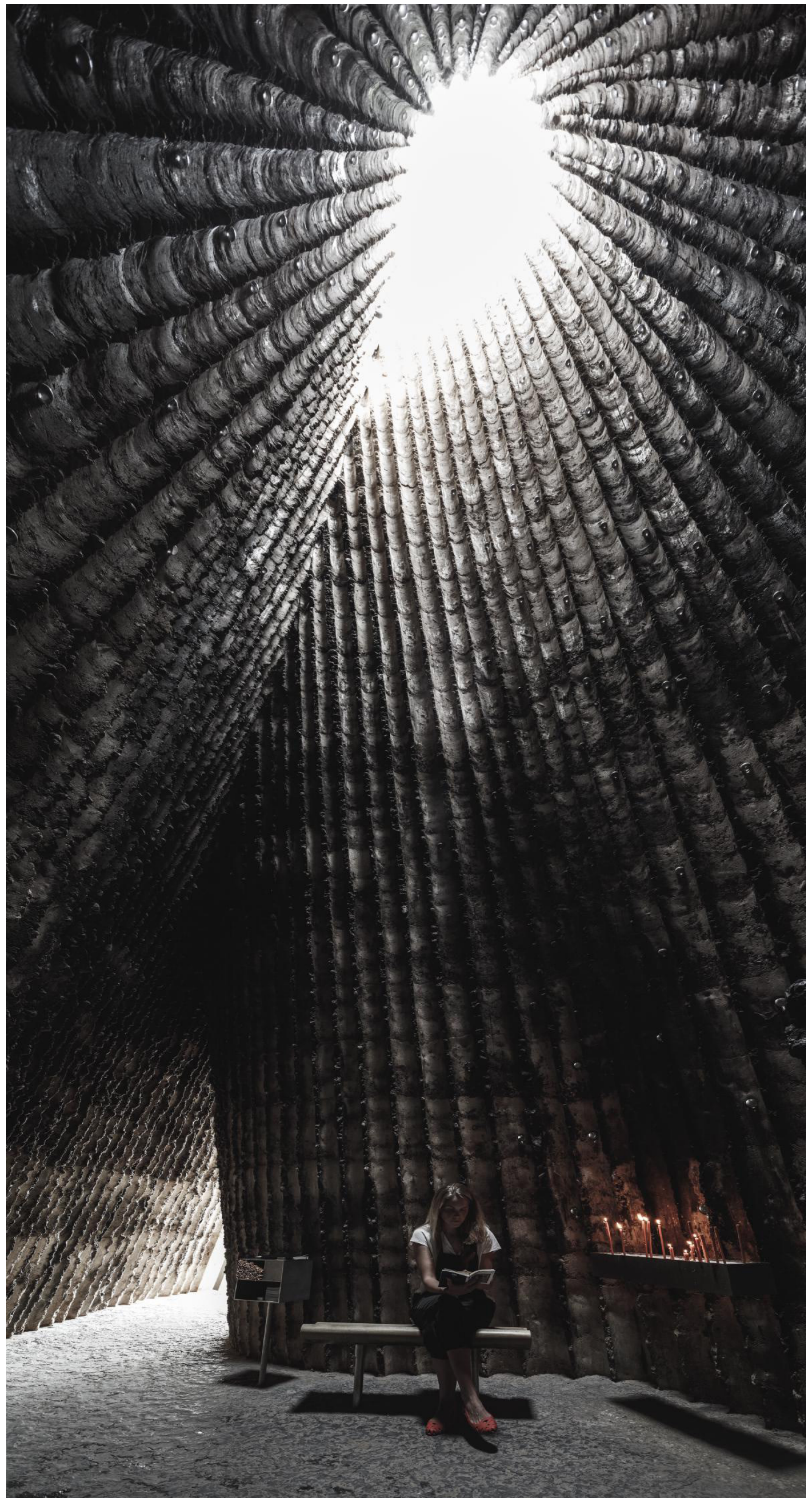


Figure 31. Bruder Klaus Feldkapelle, by Peter Zumthor

Experiment 4: Potential Use-case

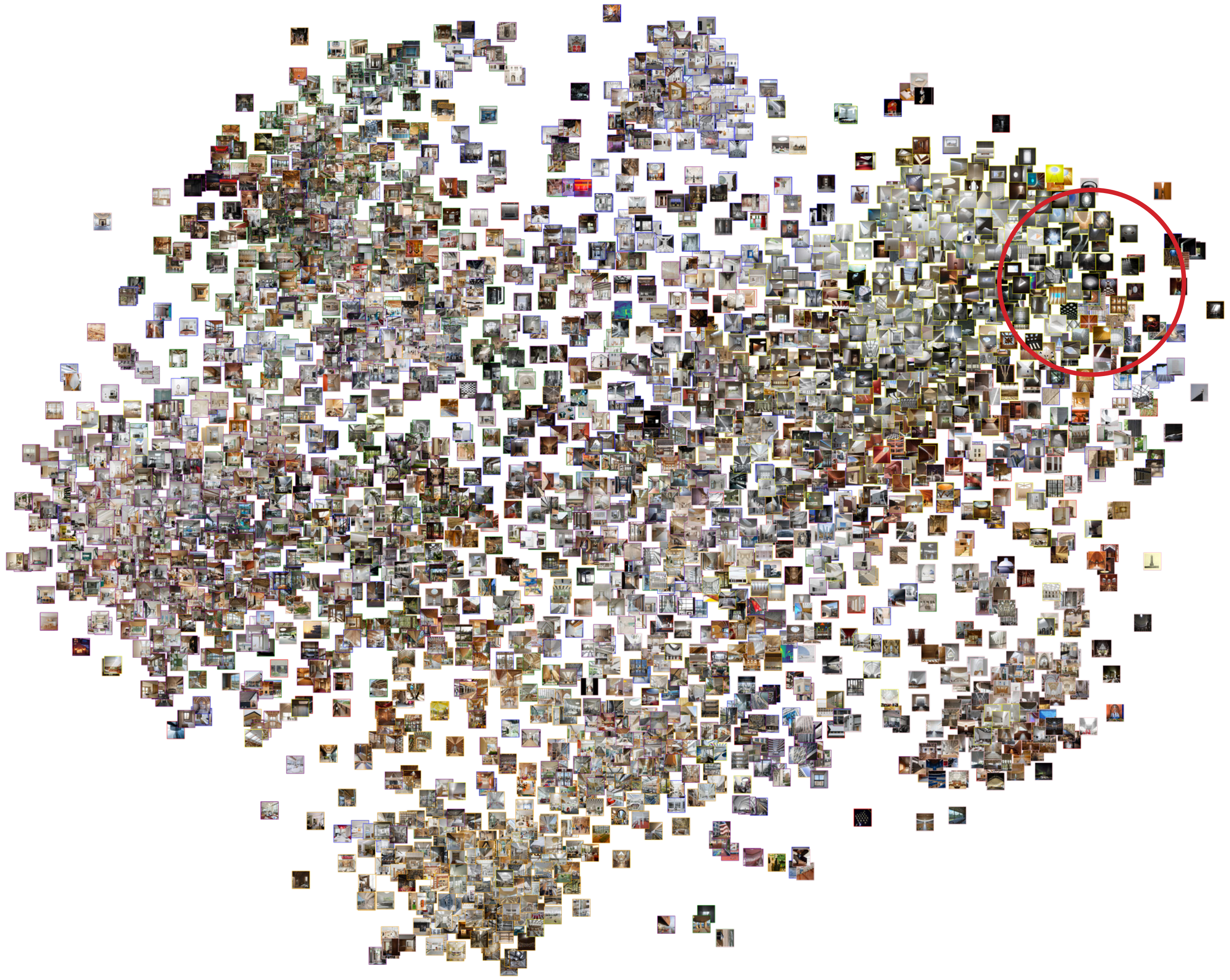


Figure 32. CLIP 2D scatter plot at perplexity = 50, full dataset (2500 images)

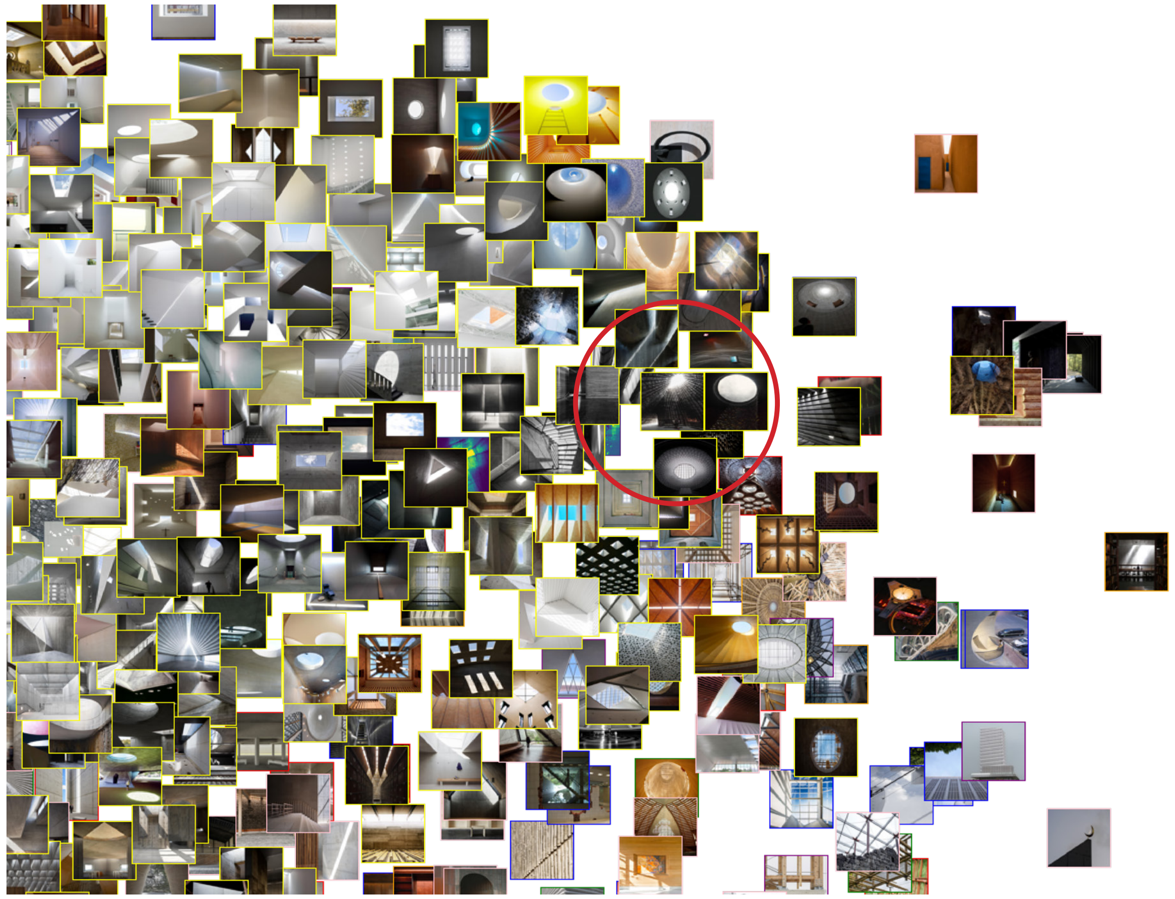


Figure 33. CLIP 2D scatter plot at perplexity = 50, full dataset (2500 images), zoomed on the top right corner

Experiment 4: Potential Use-case

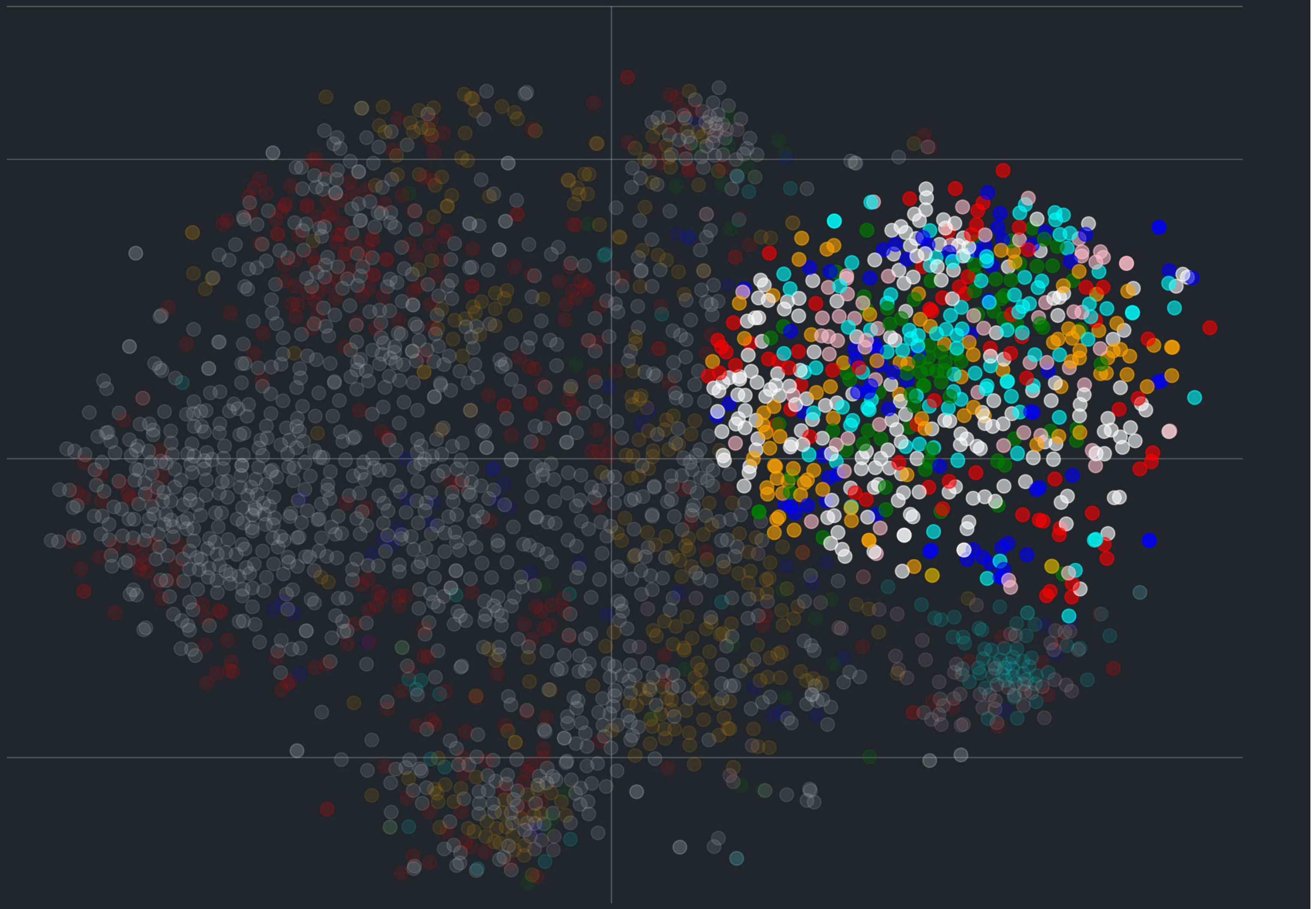


Figure 34. CLIP 2D scatter plot at perplexity = 50, full dataset (2500 images), data points shown as dots instead of thumbnails. Highlighted data-points are the images selected as input for the smaller plot

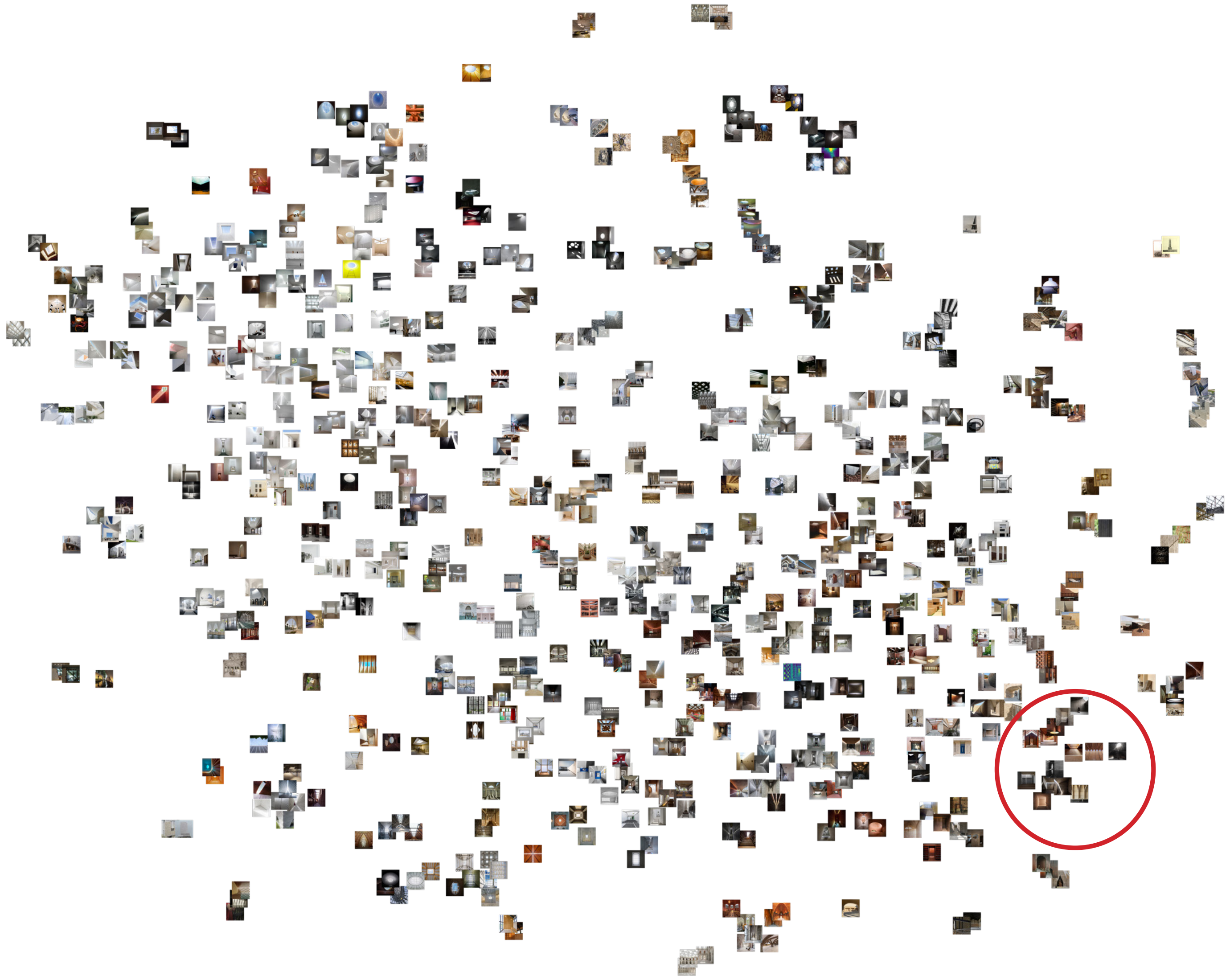


Figure 35. DINOv2 2D scatter plot at perplexity = 10, selected dataset (~400 images)

Experiment 4: Potential Use-case

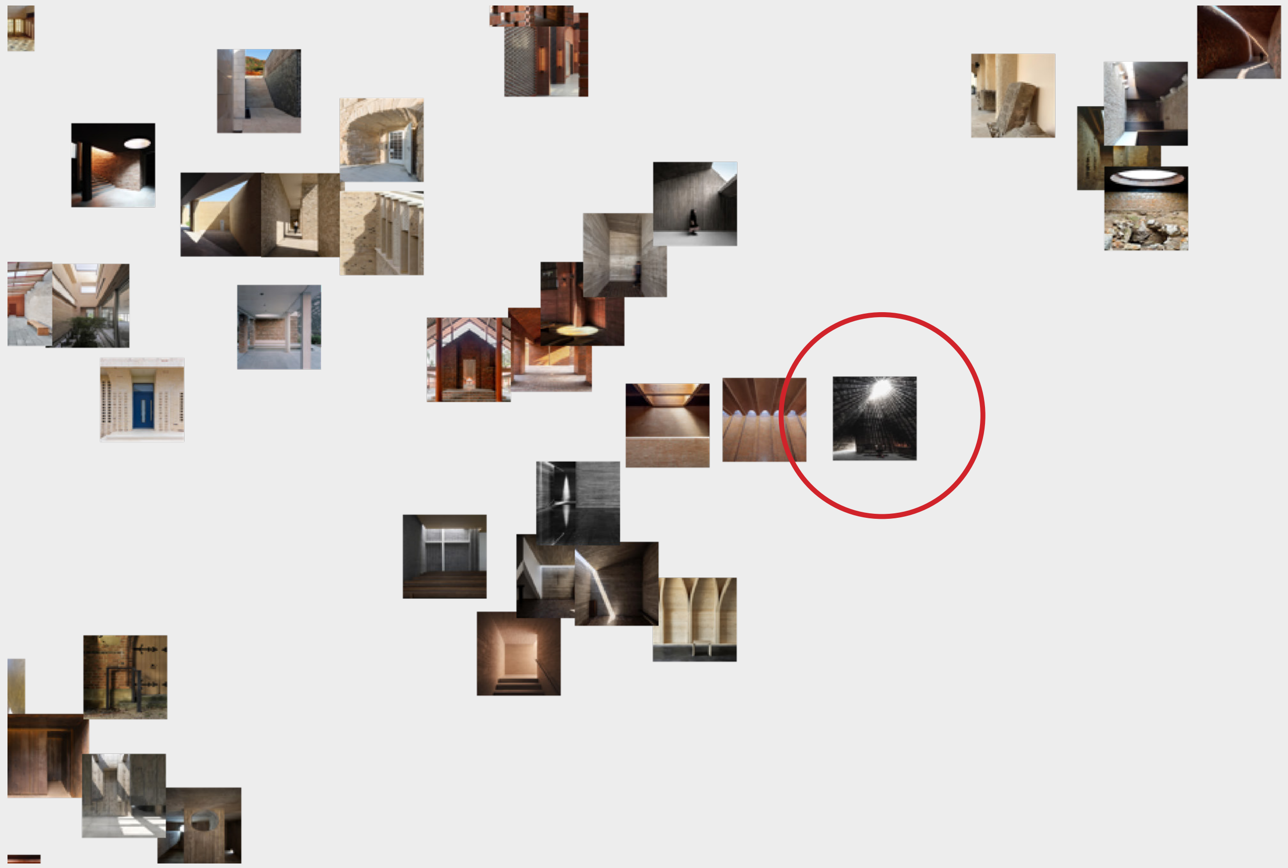


Figure 36. DINOv2 2D scatter plot at perplexity = 10, selected dataset (~400 images) zoomed on bottom right

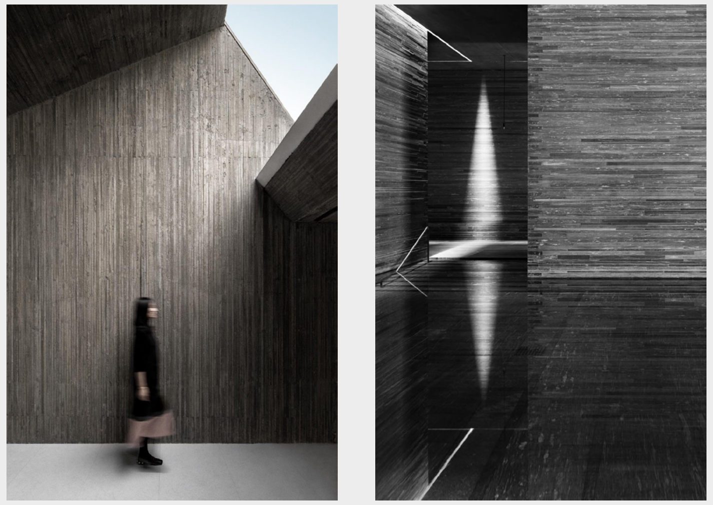


Figure 37. Two precedents found from the experiment. Waterside Buddhist Shrine by Archstudio, China (left). Therme Vals by Peter Zumthor, Switzerland (right).

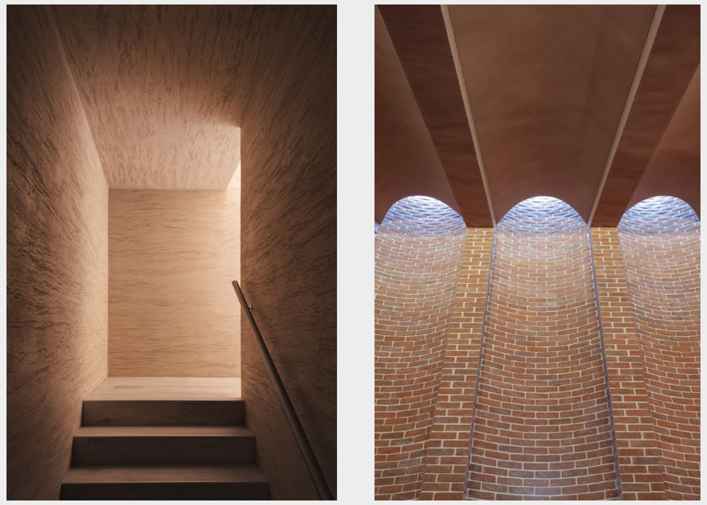


Figure 38. Two precedents found from the experiment. Cafube Funeral home, Switzerland (left). The International Rugby Experience, Cultural Institution, UK (right)

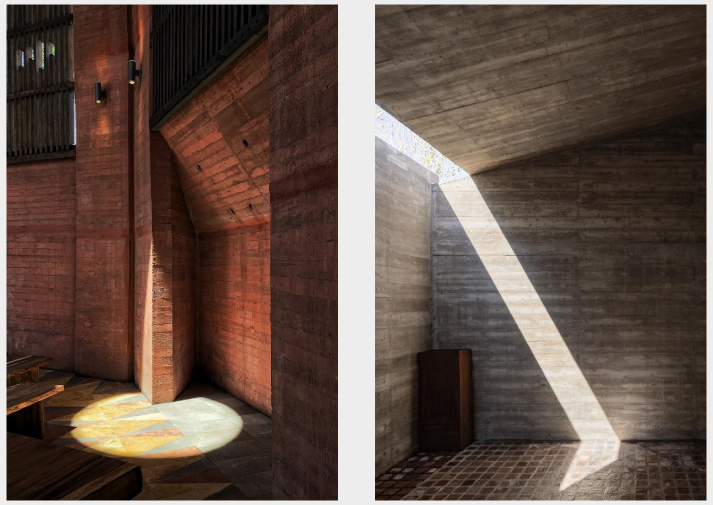


Figure 39. Two precedents found from the experiment. Capilla de la Santa Cruz, Mexico (left). San Peregrino Oratory, Praying Room, Argentina (right)

Conclusion

Looking at 2D plot output from CLIP and DINOv2, it is clear that because of the goal of these models being more general purpose orientated, the models seem to put more emphasis on more typical classes, such as clustering more based on typology as seen in the case of CLIP. Even with these typical classes, with the case of the dataset used in the experiment, it can already start to give an answer to the fourth sub question: 'What are the different ways natural light can be used to contribute to the creation of atmospheres?'

Natural light is present in every building, but at the same time is also versatile and can take very different forms. It can take centre stage and be the most defining factor of atmosphere, radiate an otherworldly presence (Figure 40), or on the opposite end, it could merely be used just for its physical property to illuminate space (Figure 41). In the two cases, the model is able to separate and determine that the two belong in different extremes. There are definitely other images that can be arranged as a gradient between these two extremes. However, at the moment the foundation model is not tuned for this.

The investigation shows that the foundation model has the potential to be adjusted to do this. When an image does not fall strictly within these typical classes, it has been observed that they are placed outside of defined clusters not at random, but rather the image are placed there because they share some similar features to the images inside a defined class. When investigating these outlier images, it becomes clear that some of the features the model picks up are relevant to architectural atmospheres.

Furthermore, it is also possible extract a more fine grained classification by isolating the specific images and running the whole process again. As demonstrated in Experiment 4, it yielded results that are similar in terms of architecture atmospheres.

Lastly, CLIP's ability of integrating captions adds another method of increasing the accuracy of the foundation model for the desired purpose, as shown in experiment 3, opening the possibility of retraining a smaller model for the specific purpose of dissecting architecture atmospheres.

Main research question: 'To what extent are foundation models an effective tool to approach and address architectural atmospheres?'

Having looked at the various results and analysis from the experiments, it is apparent that there are a number of promising results with regards to using foundation models to explore architecture atmospheres. The models have the ability and picks up relevant features, however it needs to be adjusted to prioritise these features more. Fully representing architecture atmosphere digitally is not easy. However, there are a lot of influential factors that affect atmospheres through visual means. In regards to these influential factors, with a larger data set and some re-training, foundation models are promising tools to use for addressing architectural atmospheres.

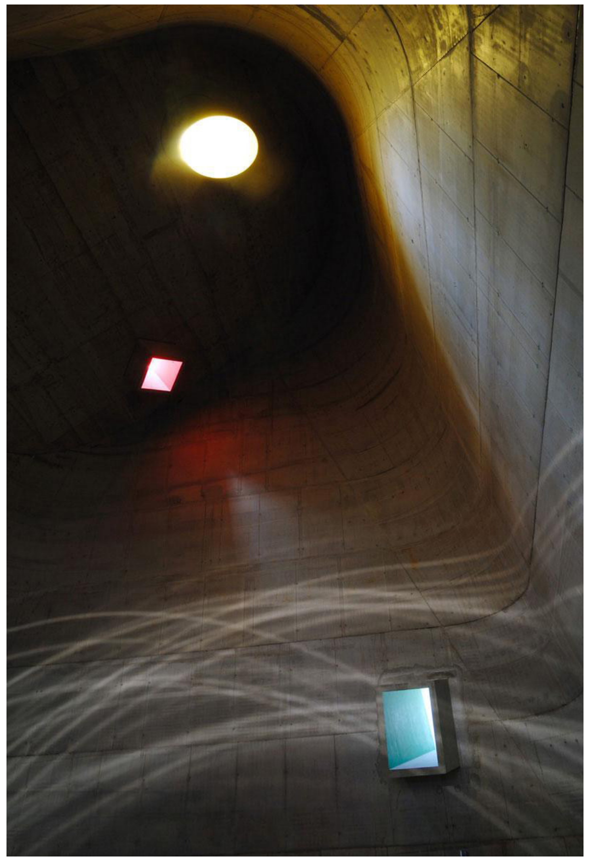


Figure 40. St. Pierre, by Le Corbusier

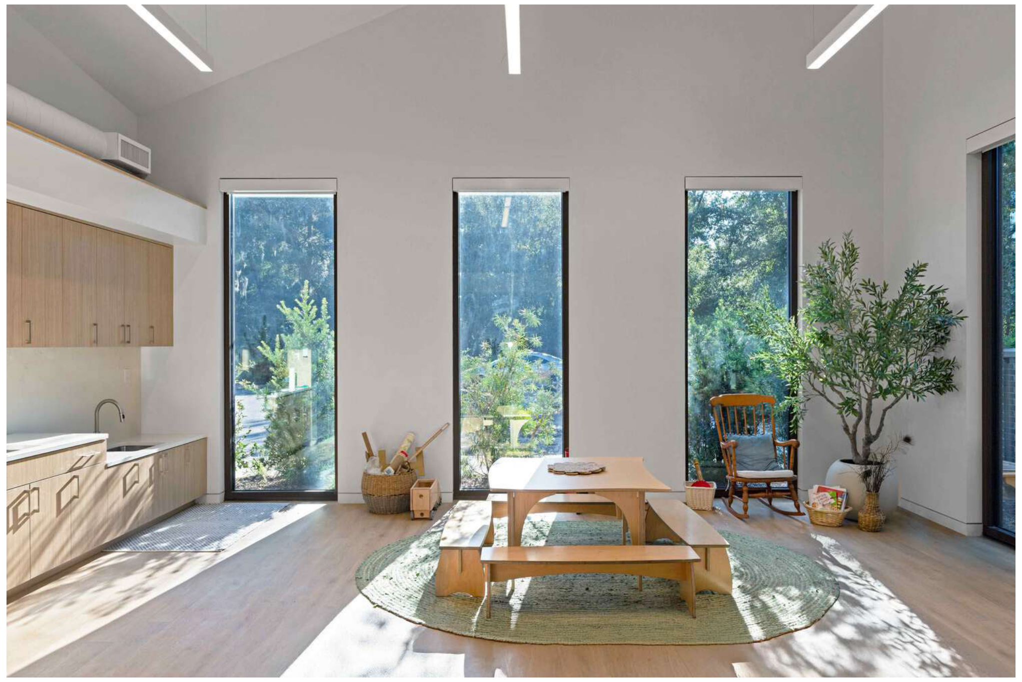


Figure 41. Preschool of the arts, by Boyd Architects

List of Figures

Bibliography

Abd-Alhamid, F., Kent, M., & Wu, Y. (2022b). Quantifying window view quality: A review on view perception assessment and representation methods. *Building and Environment*, *227*, 109742. https://doi.org/10.1016/j.buildenv.2022.109742

Ananthaswamy, A. (2024). *Why machines learn: The Elegant Math Behind Modern AI*. Penguin.

As, I., & Basu, P. (2021). *The Routledge companion to artificial intelligence in architecture*.

Bachelard, G. (2014). *The Poetics of Space*. Penguin.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., Sydney, V. A., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., . . . Liang, P. (2021, August 16). *On the Opportunities and Risks of Foundation Models*. arXiv.org. https://arxiv.org/abs/2108.07258

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021, April 29). *Emerging Properties in Self-Supervised Vision Transformers*. arXiv.org. https://arxiv.org/abs/2104.14294

Carta, S. (2022). *Machine learning and the city: Applications in Architecture and Urban Design*. John Wiley & Sons.

Doersch, C., Singh, S., Gupta, A., Sivic, J., & Efros, A. A. (2015). What makes Paris look like Paris? *Communications of the ACM*, *58*(12), 103–110. https://doi.org/10.1145/2830541

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020, October 22). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. arXiv.org. https://arxiv.org/abs/2010.11929?ref=labelbox.ghost.io

Ghojogh, B., Ghodsi, A., Karray, F., & Crowley, M. (2020, September 17). *Multidimensional scaling, Sammon Mapping, and ISOMap: tutorial and survey*. arXiv.org. https://arxiv.org/abs/2009.08136

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017, April 17). *MobileNets: efficient convolutional neural networks for mobile vision applications*. arXiv.org. https://arxiv.org/abs/1704.04861

Khademi, S., Shi, X., Mager, T., Siebes, R., Hein, C., De Boer, V., & Van Gemert, J. (2018). *Sight-Seeing in the eyes of deep neural networks: Vol. abs 1511 7247* (pp. 407–408). https://doi.org/10.1109/escience.2018.00125

McInnes, L., Healy, J., & Melville, J. (2018, February 9). *UMAP: uniform manifold approximation and projection for dimension reduction*. arXiv.org. https://arxiv.org/abs/1802.03426

Pallasmaa, J. (2005). *The eyes of the skin: Architecture and the Senses*. Academy Press.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021, February 26). *Learning transferable visual models from natural language supervision*. arXiv.org. https://arxiv.org/abs/2103.00020

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by Generative Pre-Training*. https://www.semanticscholar.org/paper/Improving-Language-Understanding-by-Generative-Radford-Narasimhan/cd18800a0fe0b668a1cc19f2ec95b5003d0a5035#citing-papers

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., & Sutskever, I. (2021, February 24). *Zero-Shot Text-to-Image Generation*. arXiv.org. https://arxiv.org/abs/2102.12092

Reasons for the sensational in architecture. (2023, January 1). Domus. https://www.domusweb.it/en/architecture/2022/10/26/reasons-for-the-sensational.html

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2015, June 8). *You only look once: Unified, Real-Time Object Detection*. arXiv.org. https://arxiv.org/abs/1506.02640

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2021, December 20). *High-Resolution Image Synthesis with Latent Diffusion Models*. arXiv.org. https://arxiv.org/abs/2112.10752

Sengupta, S., Chen, J., Castillo, C., Patel, V. M., Chellappa, R., & Jacobs, D. W. (2016). *Frontal to profile face verification in the wild*. https://doi.org/10.1109/wacv.2016.7477558

Tan, M., & Le, Q., V. (2019, May 28). *EfficientNet: Rethinking model scaling for convolutional neural networks*. arXiv.org. https://arxiv.org/abs/1905.11946

The DINOv2 Team. (2023, April 17). DINOv2: State-of-the-art computer vision models with self-supervised learning. *Meta AI*. https://ai.meta.com/blog/dino-v2-computer-vision-self-supervised-learning/

Van Der Maaten, L., & Hinton, G. (2008). *Visualizing Data using t-SNE*. https://jmlr.org/papers/v9/vandermaaten08a.html

Van Dooren, E. (2020). anchoring the design process: A framework to make the designerly way of thinking explicit in architectural design education. *Architecture and the Built Environment*, *17*, 176. https://doi.org/10.7480/abe.2020.17.5351

Zhang, L., Song, J., Gao, A., Chen, J., Bao, C., & Ma, K. (2019, May 17). *Be your own teacher: Improve the performance of convolutional neural networks via self distillation*. arXiv.org. https://arxiv.org/abs/1905.08094

Zumthor, P. (2006). *Atmospheres: Architectural Environments, Surrounding Objects*. Birkhaüser.

Zumthor, P. (2010). *Thinking architecture*. Birkhauser.