(In)Visible Cities: Exploring generative artificial intelligence's creativity through the analysis of a conscious journey in latent space

Sheng-Yang Huang[1], Yuankai Wang[1], Qingrui Drolma Jiang[2]

[1] The Bartlett School of Architecture, University College London, London, UK
ucfnhua@ucl.ac.uk; ucbqy55@ucl.ac.uk
[2] Shanghai Research Institute for Intelligent Autonomous Systems, Tongji University, Shanghai, China
qingruijiangtsering@tongji.edu.cn

Abstract. The rise of generative AI has redefined architectural design by introducing latent space, challenging traditional methods. This paper aims to explore, structure, and analyse latent journeys, drawing from analytical design discourses. We construct journeys towards 'Isaura' from 'Invisible Cities' by Italo Calvino, bridging literature and visual narratives, utilising the text-image generating software, Midjourney. The objective is to identify spatial configurations that align with the designer's interpretation of the text, ensuring the accuracy of visual elements. Structured as a Markov (stochastic) process, the experiment encompasses four primary stages to offer a rational explanation for the journey and the role of each segment. Findings emphasise the potential of latent space in augmenting architectural design and underscore the necessity for analytical tools to avert the reduction of design to trivial formalism. The study's outcome suggests that understanding and leveraging the traits of latent space can nurture a more meaningful engagement with AI-driven design, presenting a novel approach to architectural creativity.

Keywords: Latent Space, Generative Artificial Intelligence, Text-to-image Generation, Architectural Creativity, Spatial Analysis.

1 Introduction

The rapid ascent of generative AI has unveiled novel horizons in architectural design, prompting a paradigm shift. This transformation departs from the conventional Cartesian framework, leading us into the uncharted terrain of latent space—a multidimensional realm unfurled by generative models. This shift not only challenges architectural norms but also ushers in innovative prospects. Traversing latent spaces holds architectural promise.

Designs dynamically emerge from 'journeys', embodying desired design attributes. Pioneers like Bolojan (Bolojan, 2019), Anadol (Anadol, 2019) and Meng (Meng, 2021) have showcased this potential.

However, latent spaces remain elusive due to a lack of robust analytical frameworks. This shortfall risks oversimplified connectionist design, undermining creative depth. Thus, the imperative lies in demystifying these spaces, unravelling mechanics, to tap into AI's full design potential. This study delves into latent space, intricately weaving structured journeys, unravelling spatial experiences. Inspired by past discourses, notably space syntax (Hillier, 2007; Hillier & Hanson, 1989), it constructs an analytical scaffold for emergent creative landscapes. Central to this quest is creating latent journeys converging on 'Isaura', an architectural entity vividly depicted in Italo Calvino's 'Invisible Cities' (Calvino, 1974). The aim is transmuting narratives into captivating architectural tales. Employing text-to-image tech and Midjourney software, we empower designers in the process.

The study aims to explore the creativity that the latent journey manifests. It perceives the journey as a Markovian (stochastic) process, delineating segment roles throughout. Foreseeing these insights deepening latent space comprehension, evolving AI design tools, stimulating novel design outcomes.

1.1 Latent Space

The concept of latent space (Abdal, Qin, & Wonka, 2019, 2020; Carranza, Huang, Besems, & Gao, 2021; Sainburg, Thielk, & Gentner, 2020) is commonly used in data science and relevant contexts refers to a compressed, often lower-dimensional space where representations of the data reside. These representations capture essential features or patterns of the data. In generative models, the latent space is the space from which we sample point vectors, often called latent codes (Abdal et al., 2020), to generate new data instances.

1.2 Hypothesis and Major Challenges

This study's hypothesis asserts that structured segmentation of text into keywords will lead to image generation, where keywords represent significant visual features within scenes of Isaura. The Midjourney blend algorithm is hypothesised to allow these features to be inherited by progeny images, resulting in unique blended outputs. The challenges of this research stem from three main areas. Firstly, the precision required in text segmentation to encapsulate essential keywords, which necessitates an analytical understanding of the text's structure. Secondly, the complex blending process, where control over the inheritance of visual features must be maintained. Lastly, the specific application of the Midjourney blend algorithm adds complexity due to its novel interpretation and prioritisation of visual features.

2 Research Method

2.1 Isaura

The research utilises an English text excerpted from 'Isaura', a segment of the novel 'Invisible Cities', originally authored by the Italian writer Italo Calvino and translated into English by William Weaver (Calvino, 1974). Isaura was selected due to its diverse scale of spatial descriptions in the text and its richness in detailed architectural and landscape elements. This selection is expected to aid in the validation of the results.

"Isaura, city of the thousand wells, is said to rise over a deep, subterranean lake. On all sides, wherever the inhabitants dig long vertical holes in the ground, they succeed in drawing up water, as far as the city extends, and no farther. Its green border repeats the dark outline of the buried lake; an invisible landscape conditions the visible one; everything that moves in the sunlight is driven by the lapping wave enclosed beneath the rock's calcareous sky. Consequently two forms of religion exist in Isaura. The city's gods, according to some people, live in the depths, in the black lake that feeds the underground streams. According to others, the gods live in the buckets that rise, suspended from a cable, as they appear over the edge of the wells, in the revolving pulleys, in the windlasses of the norias, in the pump handles, in the blades of the windmills that draw the water up from the drillings, in the trestles that support the twisting probes, in the reservoirs perched on stilts over the roofs, in the slender arches of the aqueducts , in all the columns of water, the vertical pipes, the plungers, the drains, all the way up to the weathercocks that surmount the airy scaffoldings of Isaura, a city that moves entirely upward." (Calvino, 1974)

The paragraph contains 217 English words and covers 1250 characters, inclusive of spaces. Within this, there are 31 nouns, 16 verbs, 15 adjectives, 3 adverbs, 2 pronouns, 16 prepositions, 2 conjunctions, 8 articles, and 1 number. No interjections were identified.

2.2 Data Preprocessing

We performed a quantitative analysis of the text, segmenting it into descriptors (keywords) extraction and syntactic tree extraction using the NLTK algorithm and the relevant dataset, which includes components such as 'punkt', 'wordnet', 'stopwords', and 'averaged_perceptron_tagger' (Aarsen et al., 2023). During preprocessing, stop words were removed, and lemmatisation was conducted. For descriptor extraction and weight determination, we employed the YAKE! method (Campos et al., 2018) , as detailed in Table 1.

For syntactic dependency tree construction, we extracted the text, applied a language model, catalogued words with their part-of-speech (POS) tags, and visualised the dependency tree using the spacy package, as seen in Figure 1a.

Our methodology aimed to identify descriptor-POS tag relationships and enable efficient text segmentation. Segmentation criteria stipulated that each segment should contain at least one descriptor and be supported by three or more syntactic tree structures, as demonstrated in Figure 1b. In Midjourney, valid prompts have an approximate soft cap of about 60 words, although this is not rigidly enforced (Wilson, 2023). Given this constraint, we limit the scale of each text segment to a maximum of three sentences.

Table1. Descriptors (keywords) extracted from the Level 2 journey.

Descriptors	Descriptor weight
'subterranean lake'	0.016277156
'inhabitants dig long'	0.03280391
'dig long vertical'	0.03280391
'long vertical holes'	0.03280391
'city extends'	0.044917079
'rock calcareous sky'	0.070361465
'buried lake'	0.070560677
'lsaura'	0.077492703
'green border repeats'	0.080849655
'invisible landscape conditions'	0.080849655
'lapping wave enclosed'	0.080849655
'wave enclosed beneath'	0.080849655
'city'	0.082378136
'drawing up water'	0.083640456
'lake'	0.08679621
'deep'	0.092264704
'subterranean'	0.092264704
'inhabitants dig'	0.095909494
'dig long'	0.095909494
'long vertical'	0.095909494

Notes: A lower score indicates an increased importance of using YAKE! for descriptor determination.

2.3 Orientation

The initial text is partitioned into 21 segments, each of which is treated as a prompt to generate corresponding segmental images through the Midjourney process. Figures 1a - c illustrate how the first seven segments operate in this process.

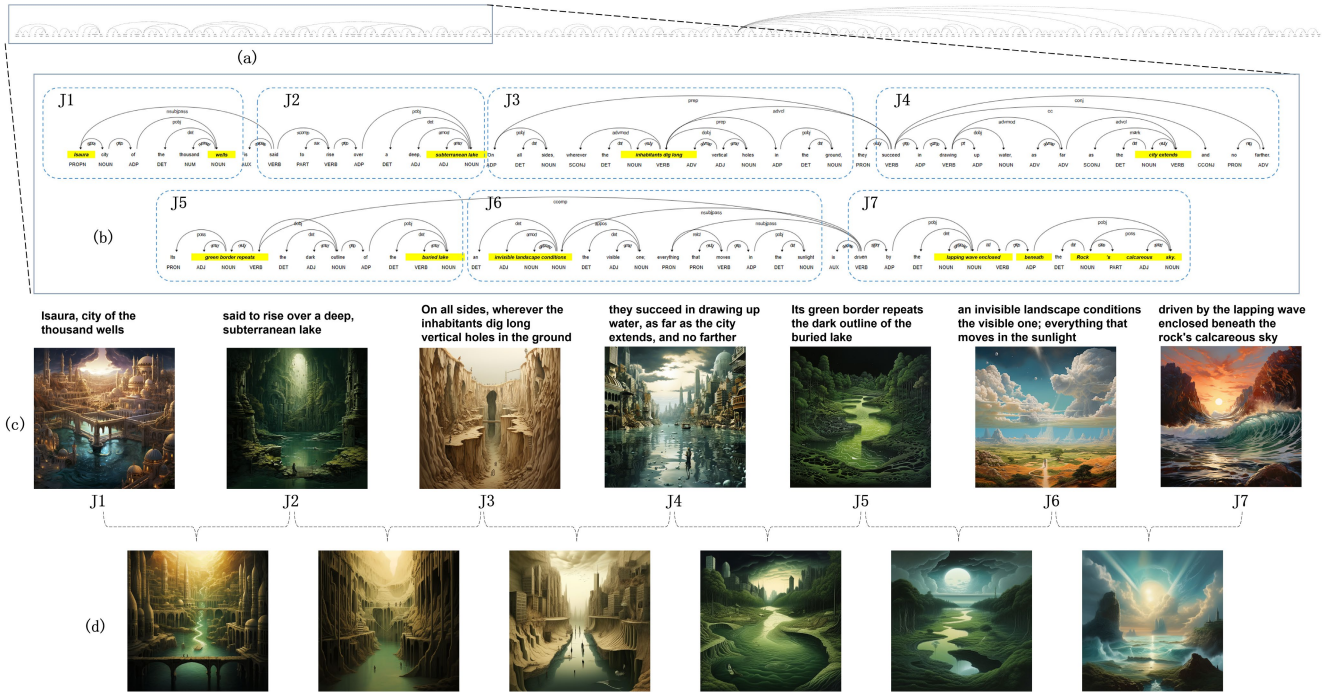

Figure 1. The process from text segmentation to journey generation, (a)The complete syntactic structure of the Isaura narratives, (b) Enlargement of part of the structures, (c) Generating Level 1 segment images, (d) blending segment images to proceed with the journey (to Level 2).

For each segment, we iteratively generate four candidates for selection. We employ the CLIP analysis technique to evaluate the semantic information contained within the generated images, supplemented with our own visual assessment. Specifically, we utilise the laion2b_s13b_b82k_augreg model within the OpenCLIP framework, a pre-trained model with 179,385,345 parameters. This scoring process (as depicted in Figure 2) enables us to filter and select segment images for journey synthesis. In other words, the chosen segment images not only achieve high similarity scores but also closely align with the authors' perception of the textual descriptions.

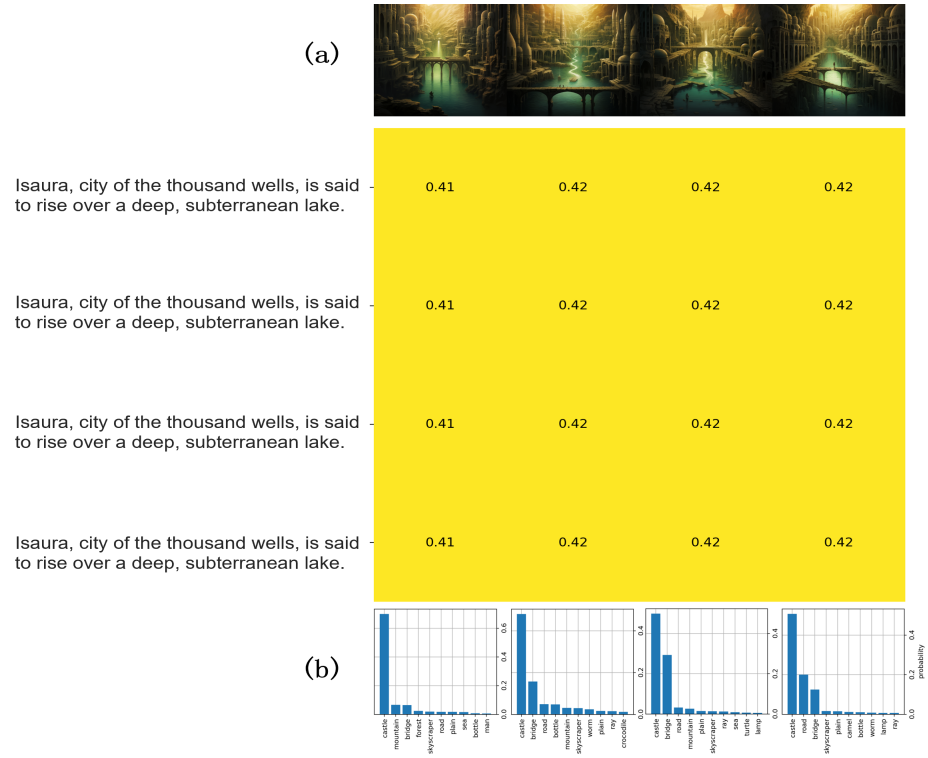

Figure 2. Assessing the text-image correlations at Level 1 journey.

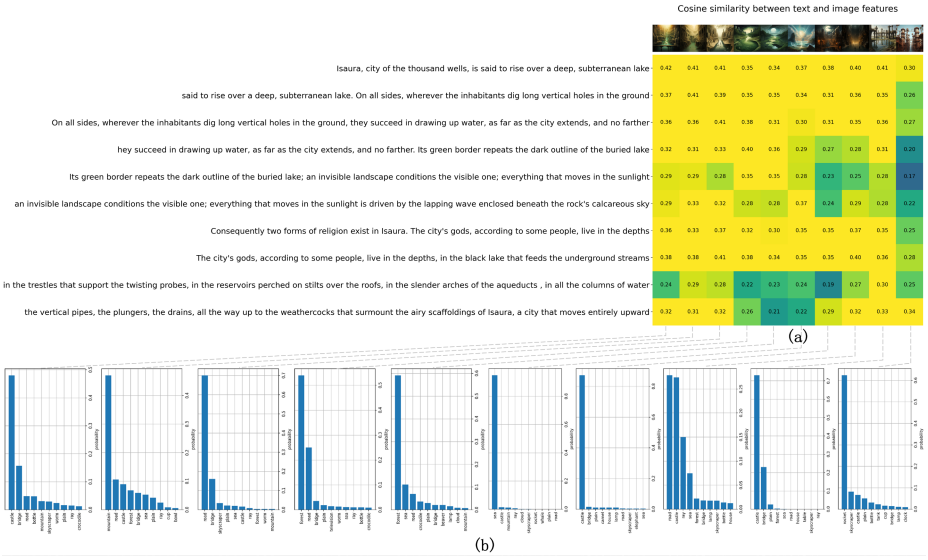

Figure 3. (a) Assessing the text-image correlations at Level 2 journey. (b) Semantic analysis of each segment at Level 2 journey.

2.4 Blending

Upon advancing to the next level, the segments will merge in pairs. This merging implies that segment images are paired according to their order in the original text, and the merged figures are output using the blending algorithm via Midjourney. Each execution produces four merged images for selection.

2.5 Journey Network Analysis

Michael Batty's perspective on cities as networks and flow systems (Batty, 2013) underscores the importance of analysing urban dynamics beyond local considerations. To capture the stochastic and complex journey interactions, centralities emerge as justifiable for main metrics. Centralities (e.g., degree, closeness, betweenness) reveal pivotal nodes and pathways facilitating information, resource, and influence flows in journey networks. This aligns with Batty's vision of cities as interconnected networks, rendering centralities apt for quantifying component significance and contribution.

2.6 Interpretive Latent Space

Since the images were generated using the non-open-source software Midjourney, the original latent space information is inaccessible. As a result, we employed a combination of Principal Component Analysis (PCA) and t-distributed Stochastic Neighbour Embedding (t-SNE) to create a three-dimensional alternative latent space. Therefore, the coordinates of the vector points corresponding to each image instance are not the same as their coordinates in the original latent space. Nonetheless, the topology of the image instance network remains unaffected.

3 Results

Presented in Figure 4, the 'Level 4' image outputs denote potential destinations of the study's examined journey. These images are not only visually congruent but also consistently symmetrical. At the heart of this symmetry is a linear water body, which anchors the composition and is bordered by a series of tower-like structures tinted in earthy tones. These scenes are illustrative of an urban settlement closely linked with a low-lying, dark water expanse. Notably, they accentuate various vertical elements that mirror specific components from the foundational text. While the original text delineates a layer of vertical pipes connecting the edifices to the water, the images unexpectedly omit key features. Elements such as windmills, revolving pulleys, pump handles, and most conspicuously, the vertical apertures, are absent.

Figure 4. Four final Isaura landscapes, generated after a complete journey. All were produced in the same batch, transitioning from Level 3 to 4, representing potential destinations of the journey.

4 Discussions

4.1 Destinations upon Various Description Granularity

As illustrated in Figure 5 (a), the entire journey comprises four levels. The final journey is obtained by merging the results of the third level. In contrast, Figure 5 (b) depicts a 'skip journey', wherein the journey concludes directly after merging all segmental images from the second level. Figure 5 (c) bypasses levels 2 and 3, generating the endpoint scene directly from all the segments in level 1. It presents an aerial view of the entire cityscape, contrasting with the partial urban scenes in Figure 5(a) and (b). Here, we can observe a correlation between the number of skips and the scope of the scene presented.

This phenomenon suggests a metric, termed as 'descriptive granularity', which can exert specific influences on the generated spatial images. To further elucidate, let us consider a journey level that encompasses *N* segments. If each of these segments is characterised by *D* descriptors, then the descriptive granularity at this level, denoted as G_d, can be mathematically expressed as:

$$\mathrm{G_d} = \sum_{\mathrm{i=1}}^{\mathrm{N}} \mathrm{D_i}/\mathrm{N} \quad (1)$$

The current spatial depth of journey is inversely proportional to the descriptive granularity. In a comparative sense, Figure 5(b), which has only undergone a single blending process, does not intuitively appear to be superior to Figure 5(a), which is situated at a greater spatial depth.

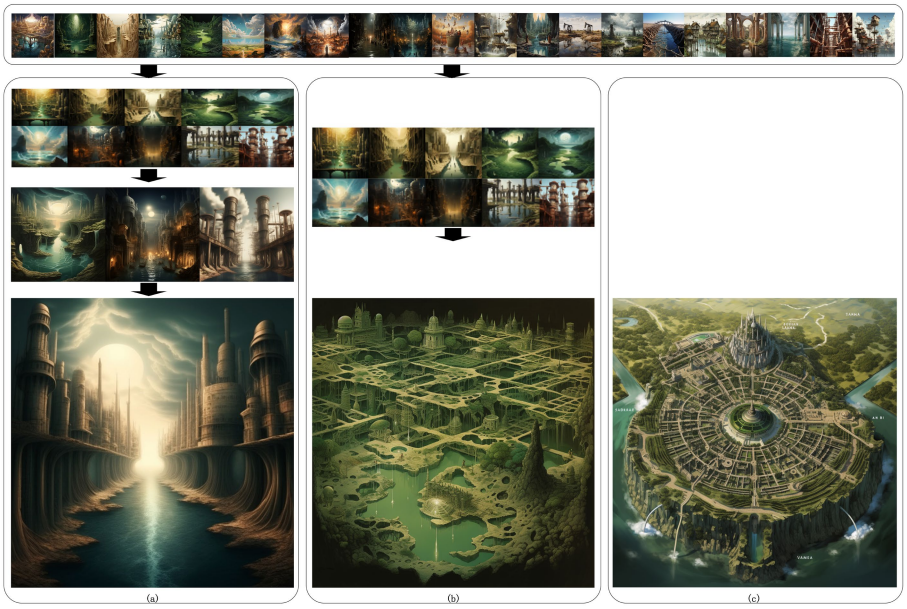

Figure 5. (a) The destination after progressing through Levels 1 to 4, a complete journey, (b) The destination achieved by directly blending Level 2 segment images, (c) The destination produced using all descriptors in an unstructured format, without undergoing other blending.

4.2 Latent Journey Analysis

Segment image (scene) 1 has significantly contributed to the overall journey, as evident from its strong degree centrality values of 7. Our findings indicate that segment scenes 3 and 12 exhibit weak connectivity with others, as revealed through image dimensionality reduction analysis and Markov chain transition probabilities (Figure 6a), as well as the projection of the latent space in the matrix post image t-SNE dimensionality reduction (Figure 6d). This phenomenon might arise due to inadequate precision in extracting critical descriptors such as "dig long vertical holes" and "in the revolving pulleys" (Figure 5f). These descriptors encapsulate dynamic processes, and unidirectionally generated images might not fully capture these processes. This interpretation is further supported by generating prompts in reverse using images through the Clip model and comparing their similarity with the original segmented text (Figure 6e).

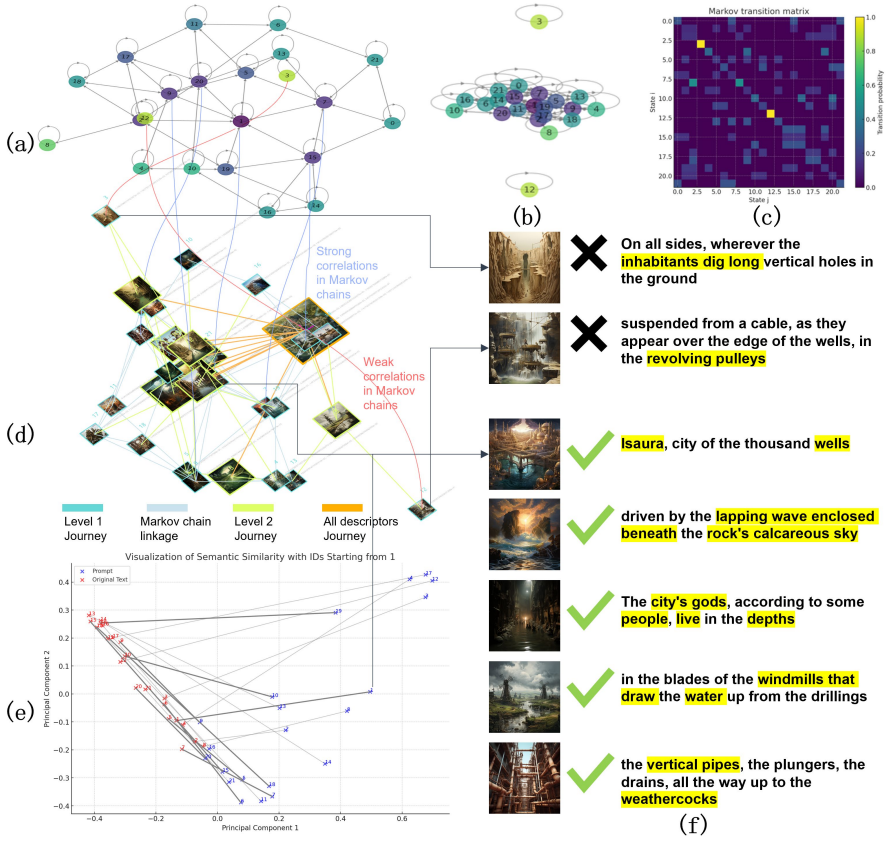

Figure 6. (a) Visualisation of the Markov chain representing the complete journey, (b) The Fruchterman-Reingold layout highlighting the limited involvement of segments 3 and 12, (c) The Markov transition matrix, (d) The latent space illustrating cross connections among level1, level2, and All descriptors, (e) Mapping of the descriptors in a two-dimensional latent space, (f) Semantic assessment of the segments.

Conversely, a scene like "Isaura, city of the thousand wells" generated scenario 1. Thanks to its distinctive spatial form description and fixed state, it also demonstrates a high level of similarity in cross-validation. Moreover, owing to its relatively abstract description that does not focus on specific minor elements, it transitions seamlessly with other scenes (Figure 6a)

4.3 Research Limitation

The outputs from generative models exhibit stochasticity due to their probabilistic nature in computing. It must be acknowledged that a single test cannot represent the entirety of all possible outcomes. Furthermore, the bias introduced by the choice of algorithmic tools should not be overlooked. In addition, due to the utilisation of the interpretive latent space, it is imperative to recognise that information pertaining to distances or any coordinate-based inquiries should be understood as mapped outcomes within this alternative latent domain.

5 Conclusion

In conclusion, this study delved into generative AI's potential for architectural creativity through latent journeys. The structured segmentation of text into keywords facilitated image generation, demonstrating the interaction between textual descriptions and visual outcomes. The use of interpretive latent space, PCA, and t-SNE allowed for insightful analysis despite the limitations of the Midjourney software. The proposed 'descriptive granularity' metric emerged as a valuable tool for evaluating text-image generation processes. The results underscore the significance of controlled blending in creating captivating architectural scenes. Moving forward, incorporating parent image weights and human design input presents promising avenues for enhancing generative design. By bridging the gap between text and imagery, this research has presented an attempt in advancing AI-driven architectural creativity and analytical methodologies.

Acknowledgements. Author 1 and Author 2 have contributed equally to this research and should both be considered as co-first authors of this paper.

References

Aarsen, T., Nothman, J., Bird, S., Dimitradis, A., Sepler, D., Milajevs, D., . . . Tan, L. (2023). Natural Language Toolkit. Retrieved from https://www.nltk.org/

Abdal, R., Qin, Y., & Wonka, P. (2019). *Image2StyleGAN: How to Embed Images into the Stylegan Latent Space?* Paper presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision.

Abdal, R., Qin, Y., & Wonka, P. (2020). *Image2StyleGAN++: How to Edit the Embedded Images?* Paper presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Anadol, R. (2019). Machine Hallucination. Retrieved from https://refikanadol.com/works/machine-hallucination/

Batty, M. (2013). *The New Science of Cities*. Cambridge, MA: MIT Press.

Bolojan, D. (2019). Deep Himmelblau. Retrieved from https://coop-himmelblau.at/method/deep-himmelblau/

Calvino, I. (1974). *Invisible Cities*. Orlando, FL: Harcourt Brace & Company.

Campos, R., Mangaravite, V., Pasquali, A., Jorge, A. M., Nunes, C., & Jatowt, A. (2018). Yet Another Keyword Extractor (Yake). Retrieved from https://pypi.org/project/yake/

Carranza, E. R., Huang, S.-Y., Besems, J., & Gao, W. (2021). *(in)Visible Cities: What Generative Algorithms Tell Us About Our Collective Memory Schema*. Paper presented at the 28th International Conference of the Association for Computer-Aided Architectural Design Research in Asia (CAADRIA) Ahmedabad.

Hillier, B. (2007). *Space Is the Machine: A Configurational Theory of Architecture* (e-edition ed.). London: Space Syntax.

Hillier, B., & Hanson, J. (1989). *The Social Logic of Space*: Cambridge University Press.

Meng, S. (2021). *Exploring in the Latent Space of Design: A Method of Plausible Building Facades Images Generation, Properties Control and Model Explanation Base on Stylegan2*. Paper presented at the The International Conference on Computational Design and Robotic Fabrication.

Sainburg, T., Thielk, M., & Gentner, T. Q. (2020). Latent Space Visualization, Characterization, and Generation of Diverse Vocal Communication Signals. *bioRxiv*, 870311.

Wilson, A. (2023). Midjourney Keywords: Top Prompt Terms Guide. Retrieved from https://approachableai.com/midjourney-keywords/