In the Context of Metaverse: Text to Animated Axonometric Space of Ancient Cities Generation

Mengyao Li[1] , Zheng Zhang[2], Xin Yan[3,*], Keyang Tang[4], Jiaxin Ke[5]

Abstract: Given the rapid development of 5G transmission, information technology, and the evolution of people's usage habits, there has been a significant increase in demand for visual communication, primarily in the form of videos and animated graphics. Coupled with advancements in AR/VR devices, this trend is gradually transitioning towards the era of Internet 3.0 and the metaverse. The demand for mapping text into images/videos in cyberspace is consequently growing. Therefore, it is of great significance to explore a low-cost pathway for generating a textual mapping metaverse space. Drawing inspiration from the three-dimensional spatial axonometric drawing techniques in architecture, this paper addresses the lack of image-based historical records in the study of ancient cities. It proposes a method for generating vectorized, axonometric images from text, with clear scene relationships. Furthermore, by introducing the concept of "narrative", the generated scene graphs are extrapolated into animations, providing a technical pathway for lightweight generation of metaverse scenes from textual content.

Keywords: Metaverse; Text-to-image generation; Scene graph; Space form; Chinese ancient city

1 Introduction

1.1 Background

As Dr. Yanfeng Li points out in his doctoral thesis "*A Study on the Relationship between Tropes and Chinese Painting History*", there is a profound connection between Chinese painting and language, especially in literati painting, which is considered the main subject of traditional Chinese painting, reflecting the fusion of language and imagery in both content and form. In the field of text-to-image generation, a lot of research focuses on establishing an interpretable mapping from low-dimensional textual semantic information to high-dimensional visual information. This cross-modal task maturity represents a trend of interdisciplinary fusion between computer vision and natural language processing. Widely applied techniques include the Generative Adversarial Network (GAN) proposed (Reed et al., 2016), contrastive language-image pre-training (e.g., CLIP, represented by DALL-E) (Ramesh et al., 2021), diffusion model architectures (e.g., Stable diffusion) (Rombach et al., 2022), among others.

However, in the context of the metaverse, current text-to-image generation technologies face challenges in generating complex scene images with multiple interacting graphic objects. Models struggle to directly parse the interaction relationships and positional information between targets from the semantic features of the text.

[1] M. Li, Academy of Arts & Design, Tsinghua University, 100084 Beijing, China, e-mail: mengyao-21@mails.tsinghua.edu.cn

[2] Z. Zhang, School of Architecture, Tsinghua University, 100084 Beijing, China, e-mail: zhangzhe23@mails.tsinghua.edu.cn

[3] X. Yan (✉), School of Architecture and Urban Planning, Beijing University of Civil Engineering and Architecture, 100044, Beijing,, China, e-mail:yanxin@bucea.edu.cn

[4] K. Tang, Future Laboratory, Tsinghua University, 100084 Beijing, China, e-mail: tangkeyang@mail.tsinghua.edu.cn

[5] J. Ke, DUT School of software Technology & DUT-RU International School of Information and Software Engineering, Dalian University of Technology, 116000 Dalian, China, e-mail: kjx20000616@mail.dlut.edu.cn

Beyond conventional Natural Language Processing (NLP) semantic analysis, delving into the essence of narratives and studying the spatial scene characteristics of texts becomes the point of entry.

1.2 Introducing Narrative into the Metaverse for Scene Animation

Joseph Frank proposed the theory of "spatial form", which belongs to the category of narratology. It is observed that in certain literary texts, narratives primarily unfold around a relatively singular and fixed spatial context (Frank, 1945). This spatial situation often imitates or reproduces historical or real-life settings, aligning with people's logic of life, thus constituting a substantial and tangible "textual reality space". The selected text for this study, "A Record of Buddhist Monasteries in *Luoyang*", is precisely based on the real spatial foundation of *Luoyang* in *the Northern Wei* Dynasty (Yang & Zhou,2022). Similarly, in modern short stories such as The Furnished Room by O. Henry, narratives unfold around a perceivable real space(Henry, 2020). Furthermore, the "spatial form" defined by scholars like Joseph is an emotional space generated through the joint interaction of writers and readers, a perceptual space formed through the reader's recreation. Due to the limited description of the real space in the text, the scene layout generated by the text itself necessarily involves AIGC for computer vision supplementation and fitting; this constitutes a virtual space superimposed on the concrete perceptible "real space" of the textual story.

Joseph Frank argues that modernist literature is "spatial", whether in novels, poetries, or others, and the "simultaneity" of space replaces the "sequence" or series of time (Wu, 2008). It seems that what the novel presents is no longer narration, but a large number of fragmentary details. These detailed presentations manifest a spatial form (Wu, 2017), thereby exhibiting characteristics of temporal suspension and juxtaposed structure based on spatialized scenes.

In this sense, the differences between traditional animation-picture narrative and the spatial narrative that this study aims to investigate are obvious: traditional animation-picture books always tell a story, revealing one aspect and possibility of the narrative object, while animation-picture books based on spatial models tell many stories, revealing multiple aspects of the narrative object and endless possibilities. This model is particularly suitable for new media production and has strong interactivity between audience-readers.

1.3 Overview of Research

The subject of this study is to construct a visual ancient urban space based on historical imagination, aiming to propose a lightweight pathway for generating axonometric vector scene animations from text within the context of the metaverse. The technical methods involved in this research encompass axonometric image representation, computer vision, text semantic analysis, computational geometric modeling, etc. The research methods employed include literature review, interdisciplinary research, survey research, computational research, and so on.

This research is based on the introduction of axonometric representation as a lightweight approach to reconstruct urban spaces and events. It begins by analyzing historical texts to extract spatial data and personal event information related to the city of *Luoyang* in *the Northern Wei* Dynasty. The former focuses on examining the composition, scale, and layout of individual buildings, the distribution of the urban road network, and the relationship between courtyards and streets, estimating and inferring the basic spatial organization of the ancient city. The latter focuses on integrating people and objects into a multi-element framework based on dynastic rituals and historical sources, placing them within the architectural space by considering their relative positions, spatial overlaps, and movements. These two approaches jointly depict the daily life of *Luoyang* in *the Northern Wei* Dynasty, producing a rich collection of animated films portraying the ancient city (see Fig. 1).

This study contributes to visually presenting the spatial relationships of ancient urban descriptions from historical texts, promoting historical research. It also aids in the low-cost generation of urban spatial narrative animations, enhancing cultural dissemination. Furthermore, it facilitates the provision of lightweight digital spatial scenes, enriching the development of the metaverse.

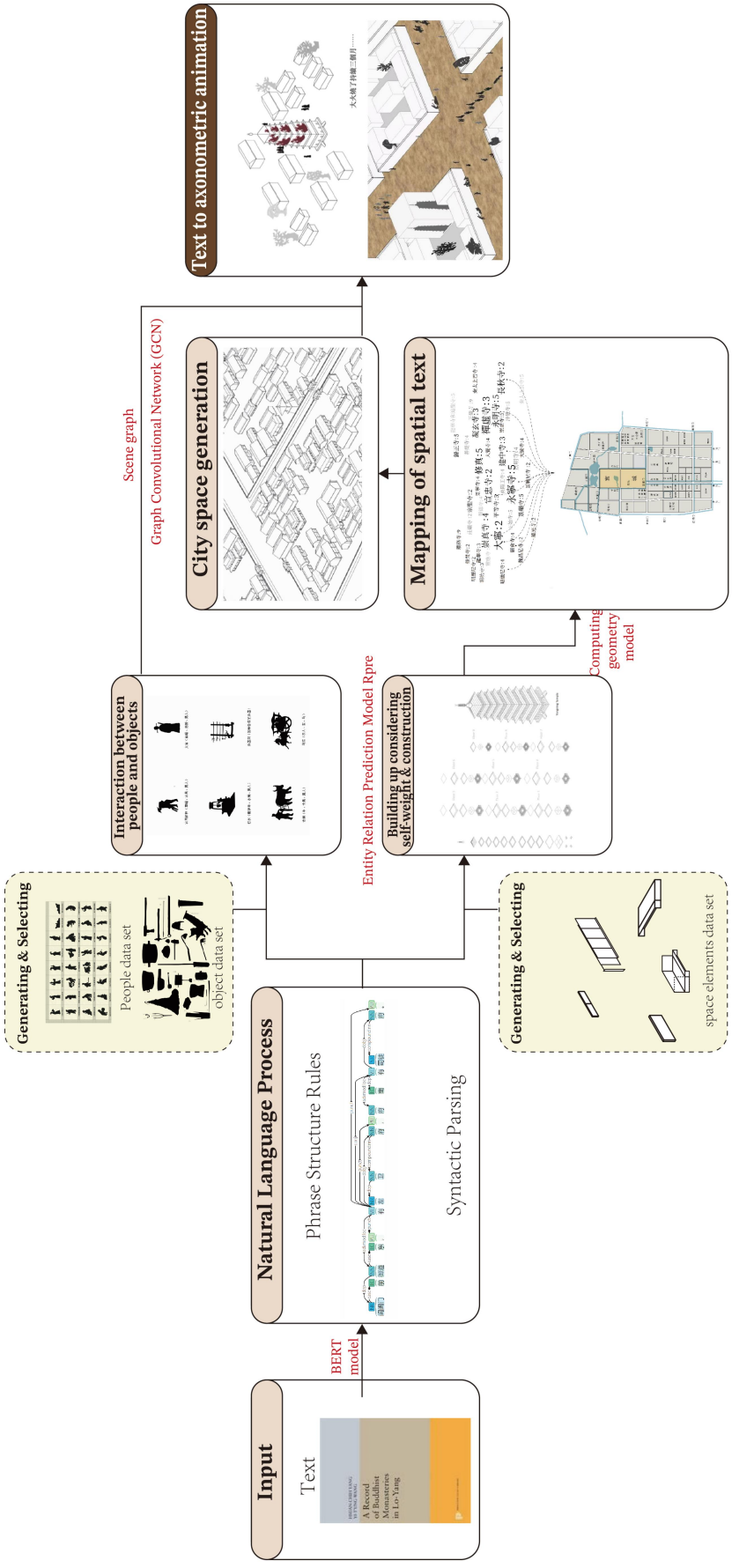


Fig. 1 Pathway of Text to Animated Axonometric Space.

2 A Lightweight Metaverse Form: Spatialization of Axonometric Vector Images

Vector graphics use geometric primitives such as points, lines, and simplified shapes to represent images. Their quality remains undiminished after scaling, as they are not resolution-dependent, and the files are easily editable with a smaller file size. According to Gestalt psychology, human perception prefers simplification, which means that the less information required to describe the organization of an object, the more likely it is to be perceived (Koffka, 1935). Simplified vector graphics are suitable for focusing or magnifying the story's setting in metaverse scenes according to narrative demands and can be arbitrarily scaled. The vector format employed in this study, inspired by axonometric drawings, is a "measurable" two-dimensional projection displaying three-dimensional space. It simultaneously reflects the three coordinate planes of an object on a projection plane, closely resembling human visual habits and providing a sense of depth (see Fig. 2). This method finds extensive applications in 3D modeling, artistic creation, and computer graphics interfaces.

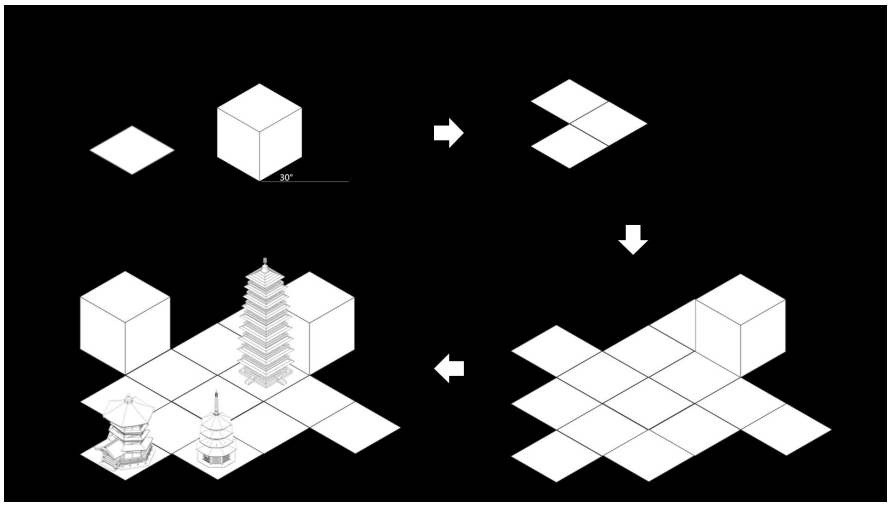


Fig. 2 Spatialization of Axonometric Vector Images.

The article begins with ancient urban texts that have restoration and experiential requirements, transforming the architectural blocks analyzed from the text into combinations of vector lines with computable lengths and locatable coordinates. The lines adopt a simplified drawing style, emphasizing their different spatial logic relationships, such as progression, overlap, symmetry, etc. Gestalt psychology suggests that closed and meaningful forms are more likely to form shapes (Koffka, 1935). Each textual element is broken down into axonometric representations, expressed in closed vector line drafts, resulting in individual components such as "roofs, beams, columns, walls, bases, steps, ramps", and other architectural elements (see Fig. 3). With a predefined baseline, the material's weight and the physical reality effect of solid construction can be presented through height-based stacking.

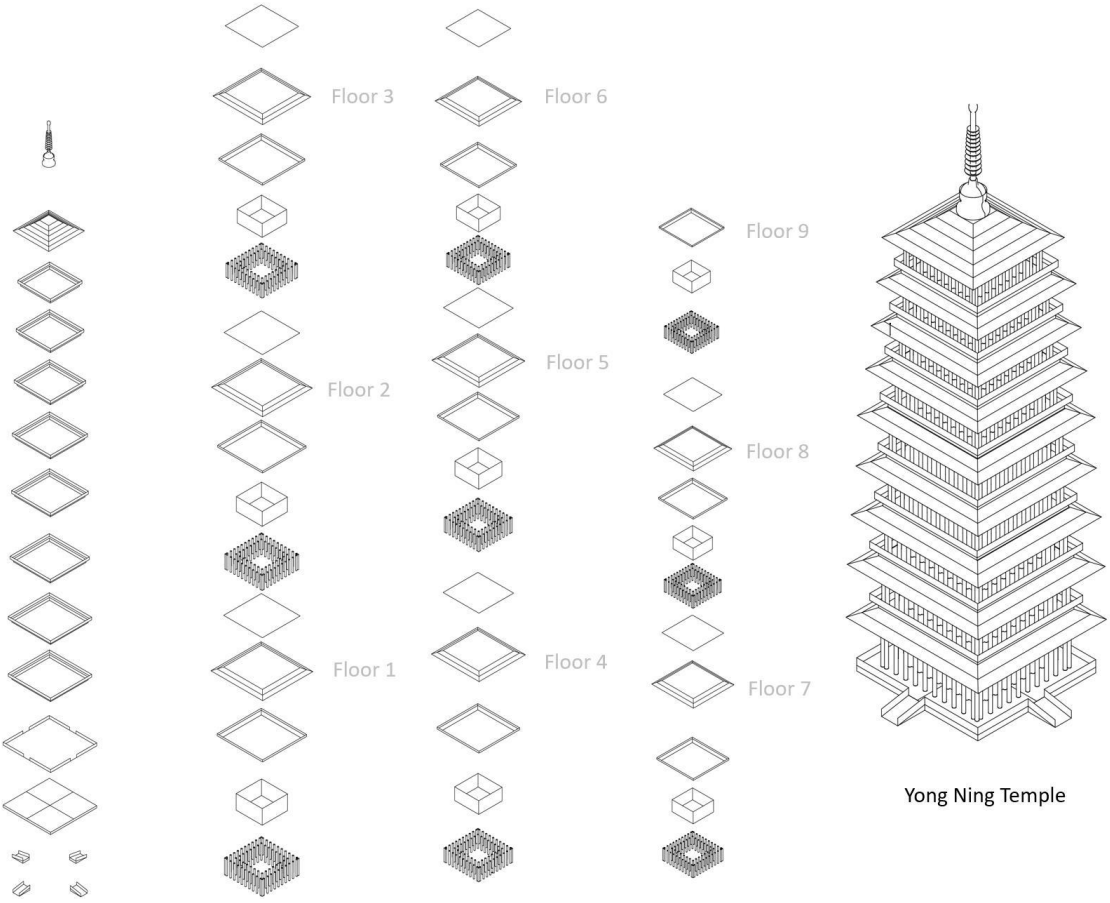


Fig. 3 Model of Yongning temple.

In painting, when several objects meet locally on a projection plane, they are perceived as having a distinct separation relationship of foreground and background depth. However, if the overlapping parts are attributed to the two objects, the phenomenon of "Transparence", commonly seen in modernism, emerges—introducing a new spatial dimension beyond the physical order of reality (Rowe & Slutzky, 1963). Compared to graphics with transparency, axonometric drawings allow for the intuitive understanding of the relative positions between buildings through occlusion relationships (from the size distribution of neighborhood positions to the arrangement of buildings within each block), delineating the cognitive five elements of city image—roads, regions, landmarks, boundaries, and nodes (Lynch, 1964). This facilitates the restoration of the virtual metaverse space of ancient cities.

3 Deconstruction of Text and Correspondence with All Spatial Elements

Gestalt psychology posits that human perception is inherently organized as a coherent whole entity, which means that there are more significances in the whole and the whole entity is not simply a sum of its parts. Humans naturally distinguish between Figure and ground (Koffka, 1935). Therefore, the decomposition and consideration of graphic elements are particularly important, as two-dimensional images are composed of points, lines, shapes, and their combinations. Characters and objects are also essential components of the urban metaverse space. Based on textual records and ancient images, the actions and gestures of people from different dynasties are exhaustively classified (such as hand gestures, sitting postures, bowing, etc.). Objects include key props relevant to the plot and the characteristic images of artifacts from the corresponding dynasties, classified based on categories such as clothing, food, shelter, and transportation. Technically, this study utilizes HanLP model (He & Choi, 2021) to achieve employing part-of-speech tagging and syntactic

analysis tools to parse the text structure, construct semantic trees, and extract the entities and their interaction relationships implied in the text(see Fig. 4).

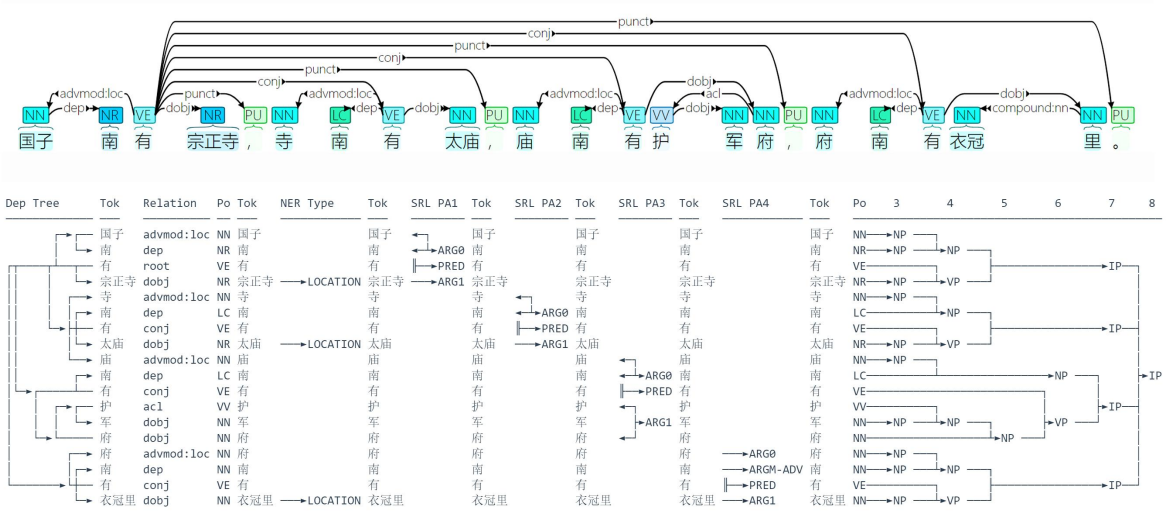


Fig. 4 Constituency Parsing by HanLP_restful.

For example, through Abstract Meaning Representation (AMR) analysis of the phrase "*Guo Zi Nan You Zong Zheng Si Si Nan You Tai Miao Miao Nan You Hu Jun Fu Fu Nan You Yi Guan Li*" (see Fig. 5), we observe that "*Guo Zi*", "*Zong Zheng Si*", and "*Tai Miao*" are correctly identified as nouns (NN), and "*Zong Zheng Si*" forms a "domain" relationship with the subsequent "*Si*". However, "*Hu Jun Fu*" is split into "*Hu/Jun/Fu*" due to its verb-object structure, and "*Li*" in "*Yi Guan Li*" is sometimes identified as a spatial relationship and sometimes as a proper noun, depending on the model. Positional terms like "*Nan*(south)" are consistently recognized as location indicators. This series of relative relationships between buildings reveals that "*Zong Zheng Si*" is located to the north of "*Yi Guan Li*", "*Hu Jun Fu*", and "Tai Miao", with each building positioned successively northward and south of "Guo Zi". Based on this, the study expands the historical texts related to *the Northern Wei* Dynasty, supplementing proper nouns and creating a specialized dictionary to improve the accuracy of urban spatial text recognition.

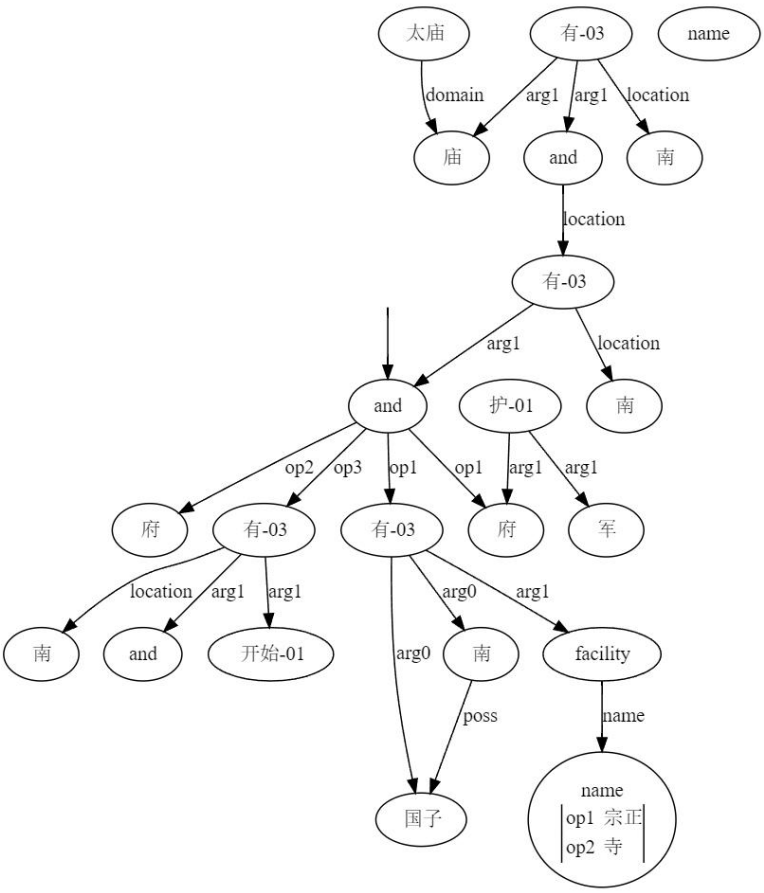


Fig. 5 Abstract Meaning Representation by HanLP 2.0.

As described above for "*Zong Zheng Si (Temple)*", "A Record of Buddhist Monasteries in *Luoyang*" mentions more than 70 temples, including "*Yongning Temple, Jianzhong Temple, Changqiu Temple, et al.*" This study cross-references multiple texts and uses semantic analysis at the urban scale to verify the relative positions of 48 temples, mapping them to form a basic road network, water system, and temple distribution. With the increasing detail of archaeological findings and site locations for *Luoyang* in *the Northern Wei* Dynasty, the study further refines the reconstructed urban space using modern Geographic Information Systems (GIS) satellite data.

Moreover, semantic analysis is applied at the block scale to refine the layout of courtyard buildings. Descriptions such as "*Fu Tu You Si Mian, Mian You San Hu Liu Chuang* (the pagoda has four sides, each with three doors and six windows)" and "*Shi Er Men Er Shi Si Shan* (twelve gates and twenty-four doors)" provide module information corresponding to the construction hierarchy of temples, enabling estimations of building sizes and typologies. The types are then matched to archetypes in a typological database, such as linear, L-shaped, C-shaped, or rectangular forms. Descriptions like "*Fo Dian Yi Suo, Xing Ru Tai Ji Dian* (a Buddha hall resembling the Taiji Hall)" and "*Seng Fang Lou Guan Yi Qian Yu Jian*(more than a thousand monk rooms and towers)" depict the type and structure of individual buildings. Additionally, the layout of courtyards and streets is outlined, for example, with "*Si Mian Ge Kai Yi Men. Nan Men Lou San Chong, Tong San Dao, Qu Di Er Shi Zhang, Xing Zhi Shi Jin Duan Men* (each side has a gate, the southern gate has three stories, with three passages, 20 zhang above the ground, resembling the present-day Duan Gate)". Through the analysis of the above historical texts, deductions can be made to estimate the basic spatial landscape of the ancient city of *Luoyang* in *the Northern Wei* Dynasty (see Fig. 6).

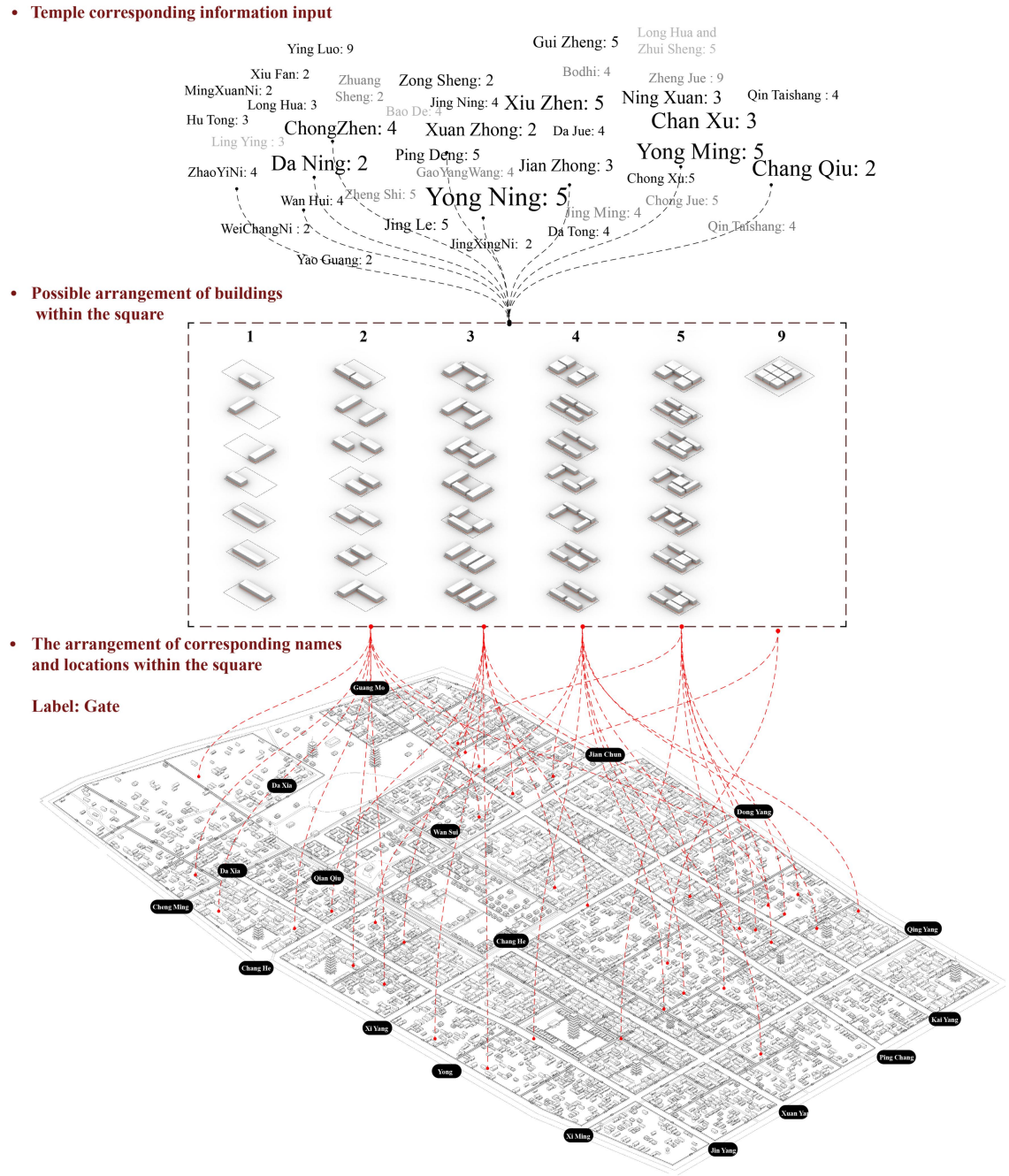
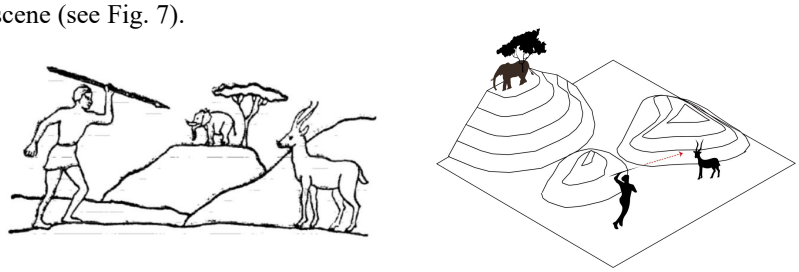


Fig. 6 Phrase-driven generative framework.

4 Coupling of Image Elements with Scene Space

Gestalt psychology posits that perception tends towards controlling multiple stimuli to form organic entities—this tendency for stimuli to be perceived as a unified whole is referred to as the principle of grouping (Koffka, 1935). Constructivist theory suggests that people from different cultural backgrounds have significant differences in their perceptual biases toward depth information in perspective views. Some Africans perceive that the spears point towards the elephant, which appears closer in two dimensions, while those with modern education are aware that, under the effect of perspective, the elephant is located on a distant mountain, with the spear pointing towards the antelope (Krech et al., 1974). However, when the same elements are redrawn using axonometric projection, perception can discern that each element in the image is

constrained by a set of axonometric reference lines at 120°angles (Thomas, 2018), allowing for a clear differentiation of the size, distance, orientation, and other interactive relationships among the elements in the scene (see Fig. 7).



(a) (b)

Fig. 7 The discrepancy in spatial representation between perspective views and axonometric projections. (Figure source: 4-a, reference 13; 4-b, author-drawn)

This study designs a semantic feature extraction model for the ancient text "A Record of Buddhist Monasteries in *Luoyang*" and selects the datasets for text-to-image generation. Characters and artifacts from *Bei Wei* Dynasty exhibit distinctive characteristics of the era. Accordingly, this study similarly depicts their clothing, hairstyles, and everyday utensils according to the logic of the real world, including wide outer garments, the upper garments and trousers, belts, high-toothed wooden clogs, '*Guo Yi*' (a ceremonial dress in the *Wei* and *Jin* dynasties), agricultural tools, carts, walking sticks, inkstones, bamboo slips, and dyeing utensils(Chen, 2012)(see Fig. 8). All of these form a system of pictorial elements that can flexibly complement those above architectural spatial context.

Fig. 8 Typical elements of daily ancient life and combination of people actions.

Axonometric drawings, devoid of perspective distortion, represent a relatively high-dimensional modality of scene space information based on two-dimensional low-dimensional images. When the viewpoint shifts, the lines of each block remain undistorted, allowing for the creation of a shifting lens effect through the movement of the frame, as well as the ability to move elements such as graphical blocks to advance the development of animated plots. Axonometric urban spatial scenes intuitively embody the "spatial form" of texts by presenting juxtaposed structures, echoing the "scattered point perspective" technique seen in classical Chinese paintings such as the "*Qing Ming Shang He Tu*" and panoramic bird's-eye views. In addition to simple undistorted camera translations, spatial meanings can be generated through techniques such as "merging", "interlacing", and "crossing" of characters and objects in planar combinations (see Fig. 9), creating spatial narrative effects such as placing items at specific elevations and people ascending stairs.

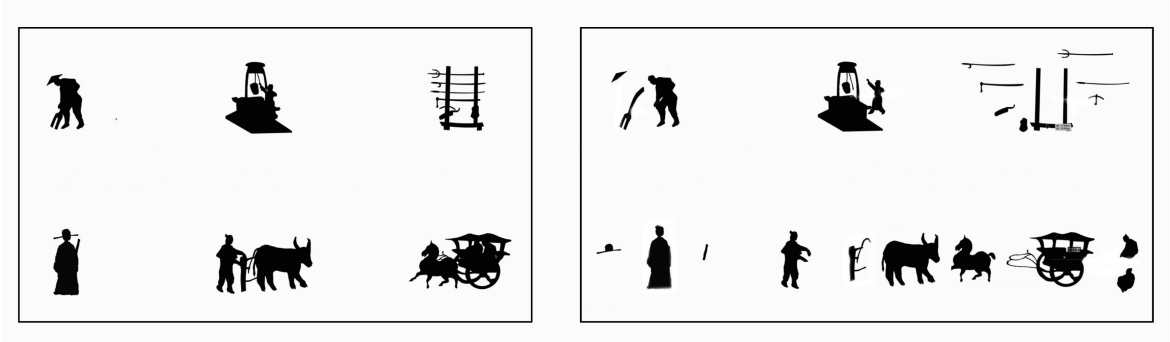


Fig. 9 Movable elements such as graphical blocks.

Generating complex scenes with multiple interacting objects in images is a high-level computer graphics task. It often involves using semantic layouts as one of the steps in the image generation process, predicting the bounding box positions and categories of each entity in the target image beforehand, and utilizing techniques like "graph convolution" to generate scene layouts (Hong et al., 2018). This study proposes a phrase-driven generative framework that encodes all words describing the same object in the input text into phrase features. It establishes correspondences between phrases and objects and models relative positional relationships between objects based on phrases (Johnson et al., 2018). With the aid of scene layout graphs, it is possible to intuitively model various entity objects and their relationships, thereby helping to bridge the semantic gap between text modality and image modality.

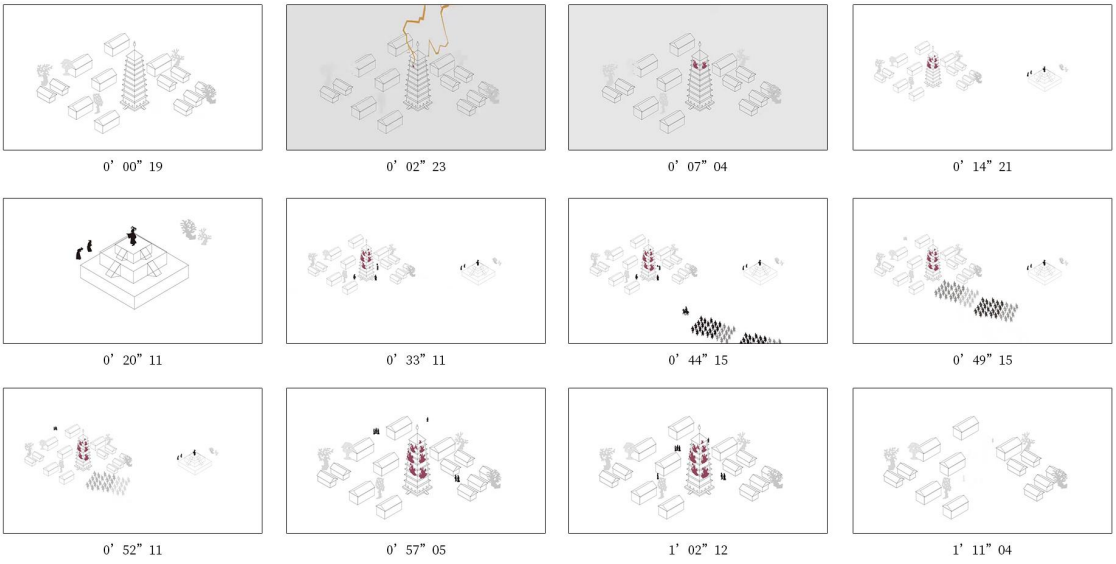


Fig. 10 Animation generation about Yongning temple on fire in Luoyang.

If the static axonometric scene graph places spatial elements refer to the spatial formal characteristic of "simultaneity", then when the elements of the axonometric scene are moved and coordinated with camera translation and scaling, similar to storyboard, a sequential depth sequence in time can be generated, thereby obtaining a multi-faceted panoramic spatial narrative (Furniss, 2008). Taking human figures as an example, while ensuring minimal loss of semantic transmission, dynamic actions of individuals are efficiently displayed using as few frames as possible. By accumulating a large library of elements (buildings, characters, objects, etc.) in the scene, even with occlusion relationships, all elements remain independently complete, forming definite relationships. For example, "opening a door" may lead to "the appearance of light (inside the room)", facilitating the future invocation of these elements in simple programming to achieve a one-change-all effect (see Fig. 10). This allows the ancient urban space to conform to basic architectural logic and initial image intelligence, enabling the narration of architectural and spatial stories. Due to the avoidance of perspective viewpoints and light and texture rendering in three-dimensional axonometry, computational

complexity and file size can be greatly reduced, further enhancing the production efficiency of mapping text to lightweight metaverse spaces (see Fig. 11).

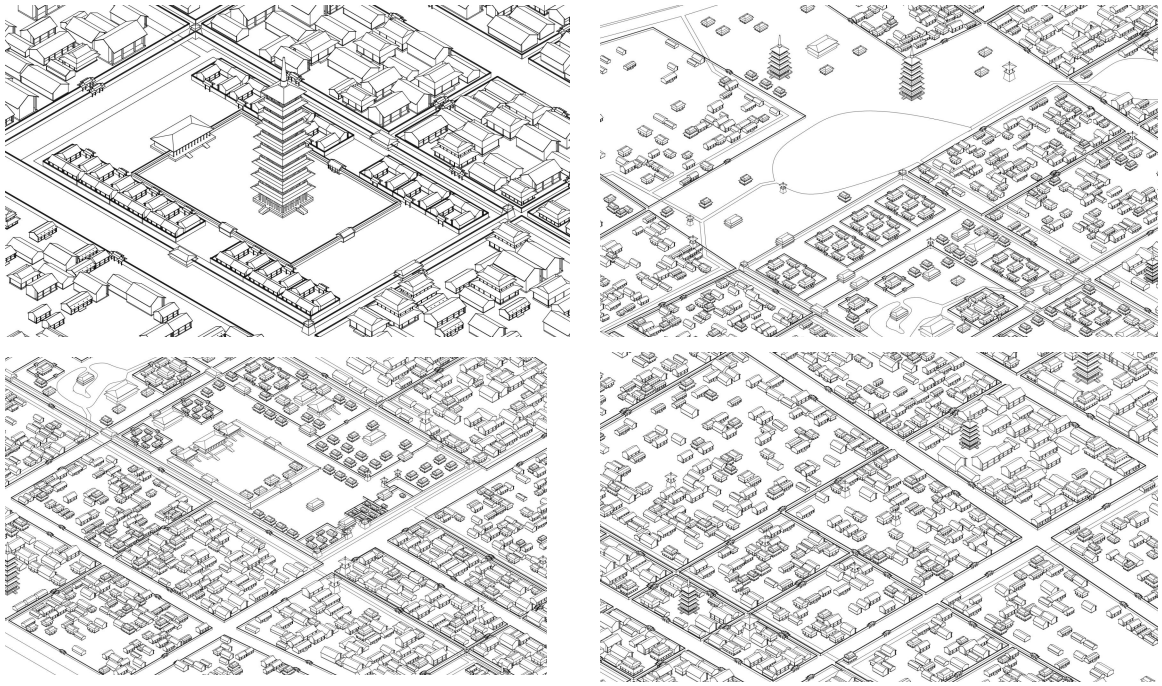


Fig. 11 Generation of ancient urban space.

5 Conclusion

This study introduces the concept of "spatial form" from narratology and draws upon visual perception theories such as Gestalt psychology and structural psychology. The research focuses on vectorized axonometric representation and takes ancient urban space as a case study to explore a pathway for generating spatial animated metaverse. Technically, the study utilizes the HanLP model and employs part-of-speech and syntactic analysis tools to parse the text structure, constructing a semantic tree to extract implicit entity objects and their interaction relationships from the text. Additionally, the research selects and expands the dataset for text-to-image generation to build a suitable text generation image dataset. Through a phrase-driven generative framework, relative positional relationships between objects are modeled based on phrases, thereby obtaining scene layout diagrams with embedded vectors to bridge the semantic gap between text and image modalities.

By constructing and updating buildings, the rise and fall of cities can be depicted (Kostof, 1991), and by depicting changes in characters and objects, humanistic scenes in the city can be deduced. Overlaying movements and scaling of scenes, image-animation based on spatial models can juxtapose multiple narratives, revealing multiple facets and infinite possibilities of the narrative objects, ultimately constituting a lightweight urban metaverse space.

Acknowledge

The authors gratefully acknowledge the financial support provided by the China Postdoctoral Science Foundation (2023M741955).

References

1. Chen, W. (2012). On the accurate restoration of historical costumes in ancient costume animated films. Decorations, (6), 64-65.
2. Frank, J. (1945). Spatial form in modern literature: An essay in two parts. The Sewanee Review, 53(2), 221-240.
3. Furniss. (2008). The animation bible: A guide to everything, from flipbooks to flash. Laurence King.
4. He, H., & Choi, J. D. (2021). The stem cell hypothesis: Dilemma behind multi-task learning with transformer encoders. arXiv preprint arXiv:2109.06939.
5. Henry, O. (2020). The furnished room. Lindhardt og Ringhof.
6. Hong, S., Yang, D., Choi, J., & Lee, H. (2018). Inferring semantic layout for hierarchical text-to-image synthesis. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7986-7994).
7. Johnson, J., Gupta, A., & Fei-Fei, L. (2018). Image generation from scene graphs. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1219-1228).
8. Koffka, K. (1935). Principles of gestalt psychology (First ed.). New York: Harcourt, Brace.
9. Kostof, S. (1991). The city shaped (pp. 9-39). Little, Brown and Company.
10. Krech, D., Crutchfield, R. S., & Livson, N. (1974). Elements of psychology. Alfred A. Knopf.
11. Lynch, K. (1964). The image of the city. MIT Press.
12. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., ... & Sutskever, I. (2021, July). Zero-shot text-to-image generation. In International Conference on Machine Learning (pp. 8821-8831). PMLR.
13. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., & Lee, H. (2016, June). Generative adversarial text to image synthesis. In International Conference on Machine Learning (pp. 1060-1069). PMLR.
14. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10684-10695).
15. Rowe, C., & Slutzky, R. (1963). Transparency: Literal and phenomenal. Perspecta, 45-54.
16. Thomas, H. (2018). Drawing architecture. Phaidon.
17. Wu, X. (2017). From Kafka to Kundera: Novels and novelists of the 20th century. Sanlian Bookstore, Life • Reading • New Knowledge.
18. Wu, Y. (2008). Reproduction of spatial theory and literature. Gansu People's Publishing House.
19. Yang, X., & Zhou, Z. (2022). Luoyang Jialan Records. Zhonghua Book Company.