DOI: 10.53469/jpce.2025.07(10).04

Carbon-Silicon Synergy: Architectural AI Design with Fused Text-Image Data

Yuxuan Zhu[1,*], Shu Wang[2]

[1]Gold Mantis School of Architecture, Soochow University, Suzhou 215009, China
[2]College of Architecture, Nanjing Tech University, Nanjing 211816, China
Correspondence Author, 480240649@ qq.com

Abstract: ***Architectural design increasingly operates across two interdependent environments: the material setting in which buildings are constructed and inhabited, and the computational setting in which requirements, images, simulations, and design alternatives are produced. This paper uses the term carbon-silicon synergy to describe a working relationship between these environments rather than a progression from the physical to the virtual. Its focus is the role of fused text-image data in architectural AI. Text can express programme, intention, regulation, and qualitative priorities, whereas images and drawings convey geometry, appearance, spatial hierarchy, and context. When these modalities are aligned, they can support a design process in which computational models translate between briefs, visual evidence, and candidate schemes while architects retain responsibility for interpretation and choice. The paper develops a conceptual framework organized around three requirements: complementary representation, traceable translation, and reversible human intervention. It then outlines a human-in-the-loop workflow extending from brief formulation and site interpretation to controlled generation, comparative evaluation, and documentation. Existing techniques in semantic segmentation, vision-language representation, diffusion-based synthesis, and graph-constrained layout generation are discussed as components of this workflow rather than as autonomous design agents. The analysis also identifies limits concerning constructability, cultural specificity, data provenance, authorship, and model error. The proposed framework is methodological; it does not report a trained system or empirical validation. Its contribution is to clarify how multimodal AI may be incorporated into architectural practice without reducing design to image production or transferring professional judgment to an opaque model.***

Keywords: carbon-silicon synergy; architectural AI; multimodal design; text-image data; human-AI collaboration; design workflow.

1. Introduction

Digital technologies have long mediated the relationship between architectural ideas and constructed form. Earlier systems primarily supported drawing, modelling, and fabrication; current machine-learning systems also classify visual material, retrieve precedents, generate images, and propose spatial configurations. This change matters because computational tools increasingly participate in how a design problem is framed, not only in how a resolved proposal is represented [1]. Architectural AI therefore needs to be examined as part of design reasoning rather than treated as a faster rendering service.

Architecture presents a difficult setting for such systems. A project brief combines measurable conditions with terms that remain open to interpretation: privacy, civic presence, domesticity, flexibility, or belonging. Site photographs and drawings record visible relations but do not, by themselves, state why those relations matter. Regulations and schedules are explicit in language yet rarely capture atmosphere, sequence, or material character. Design proceeds by moving among these forms of information, testing partial interpretations and revising them as consequences become visible.

Recent generative models make some of these translations technically possible. Vision-language learning can associate textual descriptions with visual features [5], while diffusion models can use language and spatial conditions to guide image synthesis [6,7]. Architectural studies have also demonstrated the use of deep learning for concept generation and the analysis of design images [2,3]. These advances are relevant, but a plausible image remains different from a defensible architectural proposal. It may ignore structure, climate, access, cost, regulation, or the social meaning of a place.

The central question is consequently not whether AI can produce architectural-looking images. It is how text and image data can be combined so that each modality corrects the limitations of the other, and how architects can inspect and revise the translations made by a model. This paper calls that relationship carbon-silicon synergy. The term refers to coordinated work between material conditions, human judgment, and computational representations; it does not imply that physical and digital environments are equivalent.

The paper makes two contributions. First, it defines carbon-silicon synergy through the principles of complementarity, traceability, and reversibility. Second, it translates those principles into a human-in-the-loop workflow for architectural AI. The framework connects brief formulation, visual analysis, controlled generation, comparative review, and design documentation. It is intended to organize future prototypes and evaluations, not to present an already validated software system.

The discussion is deliberately bounded. It concentrates on early and intermediate design decisions, where text-image systems are most useful for framing and comparing alternatives. Detailed engineering, statutory approval, and construction liability remain outside the competence of a general-purpose generative model. The following sections establish the conceptual position, examine multimodal information, describe the workflow, and consider its practical and ethical implications.

2. From Carbon-Silicon Dichotomy to Synergy

In this paper, the carbon domain denotes the embodied world

of sites, materials, labour, climate, institutions, and everyday use. The silicon domain denotes digital records and operations: models, databases, simulations, learned representations, and synthetic media. The distinction is analytical. Contemporary architectural practice already moves continuously between the two, since a physical building is negotiated through surveys, codes, models, messages, drawings, and images before it is constructed.

A dichotomous account places these domains in competition and encourages exaggerated predictions about replacement. A synergistic account asks a narrower question: which decisions are improved when computational representation is connected to situated professional knowledge? This shift avoids treating virtuality as an autonomous destination. Digital output gains architectural relevance only when it remains answerable to a site, a programme, and the people affected by the proposal.

Material conditions supply resistance that computational systems cannot remove. Dimensions conflict, budgets constrain choices, existing fabric carries historical value, and the same spatial arrangement can be experienced differently by different communities. These conditions are not noise to be optimized away. They are part of the design problem and must be represented in forms that can interrupt or redirect generation.

The silicon domain offers a different set of capacities. It can search large collections, expose recurring visual features, compare many alternatives, and maintain links between verbal requirements and graphic representations. Architectural research has used such capacities to generate conceptual compositions, analyse design corpora, and explore floor-plan configurations [2,3,8]. Their value lies less in numerical volume than in making relations visible that a designer may then question.

Three principles define synergy in operational terms. Complementarity requires text, images, geometry, and performance data to contribute distinct information rather than repeat one another. Traceability requires a designer to recover the prompt, source material, constraints, model version, and selection decisions behind an output. Reversibility requires every computational proposal to remain editable and rejectable; the workflow must permit a return to earlier assumptions without treating model output as an irreversible solution.

These principles also clarify authorship. The model does not possess a stake in the building or assume responsibility for its consequences. Architectural judgment remains with the people who formulate the brief, select data, interpret alternatives, and authorize decisions. Carbon-silicon synergy is therefore a distribution of tasks, not a transfer of agency from practitioner to system.

3. Fused Text-Image Data as a Design Interface

Text-image fusion is often described as if two files were simply placed in the same dataset. For design work, the more important issue is functional correspondence. A sentence about daylight should be connected to openings, orientation, section, and measured performance; a request for privacy should be connected to visibility, access, and thresholds. Fusion becomes useful when relationships among such statements and representations can be inspected across iterations.

3.1 Complementary Information in Text and Images

Text is effective for stating programme, priorities, exclusions, and reasons. It can distinguish a fixed requirement from a preference and record whose interest a decision serves. Images and drawings are effective for showing proportion, adjacency, material appearance, and spatial sequence. Neither mode is complete. A text-only brief leaves many spatial relations unresolved, while an image may conceal ambiguity behind visual coherence. A multimodal record should preserve both the statement and the visual consequence.

3.2 Cross-Modal Representation

Vision-language models demonstrate that paired captions and images can be embedded in a shared representational space, enabling retrieval and comparison across modalities [5]. In an architectural workflow, that capacity can support precedent search, prompt checking, and the grouping of alternatives by stated intention. It should not be interpreted as evidence that a model understands architectural meaning in the professional sense. Similarity in an embedding space is a useful signal, but it does not establish regulatory compliance, technical feasibility, or social value.

3.3 Visual Analysis as Situated Evidence

Before generation, images can be used analytically. Semantic segmentation classifies pixels into categories such as road, vegetation, building, vehicle, or pedestrian, making visible patterns easier to compare across a site record [4]. The resulting map is not an evaluation of the place. It is an intermediate description whose relevance depends on sampling, image date, camera position, category definitions, and the question being asked.

Figure 1 illustrates this type of scene decomposition. Within the proposed framework, such an image would be paired with textual metadata describing location, time, viewpoint, uncertainty, and design relevance. A designer could then use the pairing to formulate constraints—for example, preserving a sightline or improving pedestrian continuity—without treating the segmentation output as a complete account of lived experience.

3.4 Controlled Generation

Text-conditioned diffusion models provide a flexible means of exploring visual alternatives because language can guide synthesis through learned cross-attention [6]. Architectural use nevertheless requires additional control. Edges, depth maps, semantic layouts, and other spatial conditions can reduce the distance between a verbal intention and a geometrically organized image [7]. These controls are valuable for iteration, but they do not turn a generated image into a coordinated building model.

A responsible workflow separates exploration from

verification. During exploration, a model may vary massing, facade rhythm, material character, or atmosphere. During verification, selected alternatives are reconstructed in appropriate geometric and analytical tools, where dimensions, circulation, environmental performance, and technical systems can be checked. The distinction prevents visual plausibility from being mistaken for project resolution.

3.5 From Elements to Spatial Candidates

Architectural information often contains relations that are clearer in a graph than in a prompt: one room must adjoin another, a service zone must connect vertically, or a public route must remain separate from private circulation. Graph-constrained generation has shown how room types and adjacency relations can guide floor-plan production and iterative refinement [8]. This approach is relevant to text-image fusion because linguistic requirements can be translated into explicit relational constraints before visual candidates are generated.

Figure 2 can be read as an illustration of this intermediate layer. Textual requirements are organized into entities and relations; a network processes these constraints; candidate configurations are then visualized for comparison. The figure should not be read as evidence that the present study trained or evaluated the displayed network. Its role is explanatory: it shows why an auditable representation between brief and image is necessary.

A fused dataset for architectural work should therefore contain more than prompts and final pictures. At minimum, it should connect source images, descriptive text, spatial entities, relationships, constraints, revisions, and the reasons for selection or rejection. This record makes disagreement visible and allows later reviewers to distinguish what was observed, what was generated, and what was decided.

Figure 1: Example of semantic scene decomposition used to translate street-view images into machine-readable spatial categories

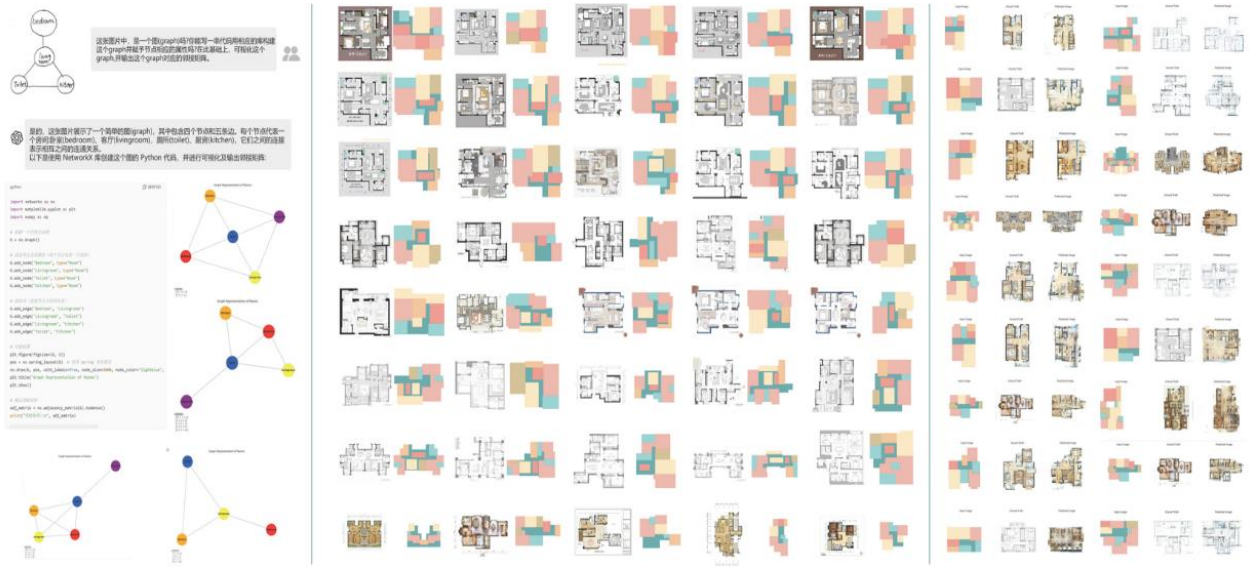

Figure 2: Text-and-graph representation linking residential elements, relations, and candidate layouts

4. A Human-in-the-Loop Architectural AI Framework

The proposed workflow has four linked stages. Each stage produces material that can be reviewed independently, and each ends with a decision made by a responsible participant. The arrangement follows established human-AI interaction guidance: systems should communicate their capabilities, support correction, and make the effects of user feedback visible [9].

4.1 Brief Translation and Data Provenance

The first stage converts the project brief into a structured record without discarding the original language. Requirements are tagged by source, priority, scale, and degree of certainty. Site images and precedents are accompanied by provenance information, including creator, date, location, permission, and any transformation already applied. Ambiguous terms are retained as questions for the design team rather than silently converted into fixed numerical targets.

4.2 Multimodal Exploration

The second stage establishes links among verbal requirements, visual references, spatial relations, and initial constraints. The model may retrieve precedents, identify conflicts in the brief, or generate candidate images and layouts. Each output is stored with its prompt, control inputs, model settings, and source references. The purpose is to create a field of alternatives that can be compared, not to rank one image as an automatic optimum.

Architect intervention is expected at this stage. A designer may reject an association, revise the vocabulary, lock a geometric relation, or introduce a counterexample that the dataset does not represent. These actions should be recorded because they reveal how professional judgment modifies the computational search. Iteration is meaningful only when it changes the problem definition as well as the appearance of the output.

4.3 Comparative Evaluation

The third stage compares selected alternatives through criteria appropriate to the project. Some criteria may be computed, such as area, adjacency, daylight access, or travel distance. Others require situated review, including cultural fit, dignity, spatial legibility, or the distribution of benefit among users. The framework does not collapse these judgments into a single score. Instead, it keeps the criteria visible and records trade-offs.

Evaluation should also test the behaviour of the AI component. Relevant questions include whether small changes in wording cause disproportionate changes, whether spatial controls are respected, whether certain typologies dominate the output, and whether sources can be traced. A limited prototype may be useful if its domain is stated clearly. Generalization to another building type, region, or population should be demonstrated rather than assumed.

4.4 Documentation and Feedback

The final stage transfers accepted decisions into conventional project documentation and records which computational outputs influenced them. Generated images remain references unless their geometry and specification have been reconstructed and checked. Post-design review should capture failures as well as successful uses, so that later projects can refine prompts, constraints, datasets, and review procedures. This feedback loop is the practical mechanism through which carbon-silicon synergy develops over time.

5. Discussion

The framework changes the emphasis of architectural AI from output production to accountable translation. Its value is not measured by the number of images generated but by whether different forms of information remain connected as a proposal develops. This position has consequences for agency, technical performance, governance, and research evaluation.

5.1 Architectural Agency and Authorship

Generative systems distribute authorship across data collection, model development, prompt formulation, curation, and subsequent design work. A single image rarely reveals these contributions. Project teams should therefore document the provenance of training and reference material, the role of model output, and the decisions made by named practitioners. Such documentation is necessary for professional accountability even where legal definitions of authorship remain unsettled.

Keeping the architect in the loop does not mean approving a result at the end of an automated sequence. It means retaining control over how the problem is represented, which alternatives are excluded, and which criteria govern selection. Intervention at these points is more consequential than cosmetic editing of a generated image.

5.2 Performance, Transfer, and Cultural Specificity

Model performance is conditional on the data and task used for evaluation. Architectural datasets often overrepresent particular building types, regions, photographic conventions, and professional styles. Outputs may therefore appear coherent while reproducing a narrow visual norm. Dataset curation can influence both fidelity and diversity [3], but curation itself is a design decision that requires disclosure and review.

Transfer across contexts should be treated cautiously. A model trained on residential floor plans may not support healthcare circulation; a facade corpus may not represent local construction practice; a visual preference model may encode the judgments of a limited group. Carbon-silicon synergy depends on maintaining contact with the specific material and cultural setting, not on assuming that scale produces universality.

5.3 Risk, Transparency, and Data Governance

Multimodal workflows introduce risks involving privacy, copyright, confidential project information, bias, and misleading output. Site photographs can contain identifiable

people or sensitive locations. Reference images may carry unclear permissions. Prompts and model logs may expose client requirements. These matters need a documented governance process rather than informal reliance on platform defaults.

The NIST AI Risk Management Framework organizes responsible practice around governance, mapping, measurement, and management across the system lifecycle [10]. Applied to architectural design, this suggests recording intended use, affected stakeholders, data provenance, known limitations, review responsibilities, and the conditions under which model output must not be used. The level of documentation should be proportional to the consequence of the decision.

Transparency also has a communicative dimension. Clients and collaborators should be able to distinguish survey evidence, authored design work, model-generated material, and speculative visualization. Clear labels do not reduce creative value; they prevent uncertain material from acquiring authority merely through graphic polish.

5.4 Research Implications

Future work should test the framework through bounded case studies. Useful evaluations would compare unimodal and multimodal workflows, measure how consistently spatial controls are followed, document the kinds of corrections architects make, and examine whether decisions can be reconstructed from the project record. Such studies should report datasets, model versions, participant roles, and failure cases. Without this evidence, claims about improved creativity, efficiency, or sustainability remain hypotheses.

6. Conclusion

Carbon-silicon synergy provides a way to discuss architectural AI without opposing physical practice and digital computation or treating one as a substitute for the other. Text-image systems can connect briefs, visual evidence, and design alternatives, but their architectural value depends on complementary representation, traceable translation, and reversible human intervention. The proposed workflow places those requirements at each stage, from data preparation to documentation. It also draws a firm boundary around the present contribution: the paper offers a conceptual and methodological framework, not a validated design system. Subsequent research must test the framework with declared datasets, explicit criteria, and accountable participants. Until then, fused text-image data is best understood as an interface for disciplined inquiry—one that can widen architectural exploration while leaving responsibility for interpretation and decision with the design team.

References

[1] Carpo, M. The Second Digital Turn: Design Beyond Intelligence. Cambridge, MA: MIT Press, 2017.

[2] As, I., Pal, S., and Basu, P. “Artificial Intelligence in Architecture: Generating Conceptual Design via Deep Learning.” International Journal of Architectural Computing, 16(4), 2018, pp. 306-327. https://doi.org/10.1177/1478077118800982.

[3] Newton, D. “Generative Deep Learning in Architectural Design.” Technology|Architecture + Design, 3(2), 2019, pp. 176-189. https://doi.org/10.1080/24751448.2019.1640536.

[4] Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. “DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs.” IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(4), 2018, pp. 834-848. https://doi.org/10.1109/TPAMI.2017.2699184.

[5] Radford, A., Kim, J. W., Hallacy, C., et al. “Learning Transferable Visual Models from Natural Language Supervision.” Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 2021, pp. 8748-8763.

[6] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. “High-Resolution Image Synthesis with Latent Diffusion Models.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10684-10695.

[7] Zhang, L., Rao, A., and Agrawala, M. “Adding Conditional Control to Text-to-Image Diffusion Models.” Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 3836-3847.

[8] Nauata, N., Hosseini, S., Chang, K.-H., Chu, H., Cheng, C.-Y., and Furukawa, Y. “House-GAN++: Generative Adversarial Layout Refinement Network towards Intelligent Computational Agent for Professional Architects.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 13632-13641.

[9] Amershi, S., Weld, D., Vorvoreanu, M., et al. “Guidelines for Human-AI Interaction.” Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Paper 3, 2019, pp. 1-13. https://doi.org/10.1145/3290605.3300233.

[10] Tabassi, E. Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. Gaithersburg, MD: National Institute of Standards and Technology, 2023. https://doi.org/10.6028/NIST.AI.100-1.