MULTIMODAL ARCHITECTURE: APPLICATIONS OF LANGUAGE IN A MACHINE LEARNING-AIDED DESIGN PROCESS

GEORGE GUIDA[1]
[1]*Harvard University Graduate School of Design*
[1]*gfguida@gmail.com, 0000-0002-2477-0064*

Abstract. Recent advances in Natural Language Processing (NLP) and Diffusion Models (DMs) are leading to a significant change in the way architecture is conceived. With capabilities that surpass those of current generative models, it is now possible to produce an unlimited number of high-quality images (Dhariwal and Nichol 2021). This opens up new opportunities for using synthetic images and marks a new phase in the creation of multimodal 3D forms, central to architectural concept design stages. Presented here are three methodologies of generation of meaningful 2D and 3D designs, merging text-to-image diffusion models Stable Diffusion, and DALL-E 2 with computational methods. These allow designers to intuitively navigate through a multimodal feedback loop of information originating from language and aided by artificial intelligence tools. This paper contributes to our understanding of machine-augmented design processes and the importance of intuitive user interfaces (UI) in enabling new dialogues between humans and machines. Through the creation of a prototype of an accessible UI, this exchange of information can empower designers, build trust in these tools, and increase control over the design process.

Keywords. Machine Learning, Diffusion Models, Concept Design, Semantics, User Interface, Design Agency

1. Introduction

In the 20th century, language came under deep suspicion in every visual arts domain, including architecture (Forty, 2004). Mies van der Rohe stated "build, don't talk" (Bonta, 1990) and Picasso that "a painter only has one language" - that of painting (Ashton, 1997), bringing forward an idea of language as an accessory or secondary within the creative process. While certain mediums find prevalence over others across history, architecture can never be reduced to a single mode of representation (Tschumi, 1993). From writing, drawing, model making, to building, language is ultimately at the center of a multimodal process. Architects navigate non-linearly between each of these modes, across the creative process. The convergence of text and image processing permits this multimodal process to increasingly be replicated or aided by machine learning models, giving new meaning to language within the creative process. With the development of new natural language and diffusion models, what therefore

emerges as the relationship between architecture and language? How will this change existing workflows and methodologies relating to 3D form generation?

Deep learning models are now capable of generating an infinite variety of high-quality images from natural language descriptions in short inference times from text prompts (Nichol et al., 2021). This increase in image quality and resolution is bringing forward a new phase of 3D synthesis that will widely affect the architecture engineering and construction (AEC) industry. Approaches and output resolutions of 3D models vary between generative design methods, depth map reconstructions of 2D images with monocular depth estimation (MiDaS) (Ranftl et al. 2020), and state-of-the-art-models that use pre-trained text-to-image diffusion models to perform text-to-3D synthesis with Neural Radiance Fields (Poole et al, 2022) (Lin et al, 2022). Each of these processes brings forward new opportunities in collaborative human-machine interactions, while fundamentally challenging our agency within emerging creative practices. This paper investigates the role of language in this process of multimodal 3D form generation using text-to-image and image-to-3D processes applied to an architectural competition of the MAXXI Museum in Rome, Italy. It explains a custom user interface and uses three hybrid machine learning-aided methodologies to unpack architects evolving agency and creative possibilities.

1.1. MULTIMODAL MODELS

State-of-the-art deep learning techniques are now able to generate high-quality images conditioned by text descriptions. These models typically use a combination of natural language processing (NLP) and computer vision techniques to generate synthetic images that reflect the input textual semantics. Some of the most effective models use Contrastive Language-Image Pre-Training (CLIP), a neural network trained on image-text pairs and guided diffusion models which use a process of denoising to generate images (Radford, 2021). Two models were used within this paper, including: Stable Diffusion v1, a latent diffusion model that uses OpenCLIP, an open-source version of CLIP and is trained on the LAION-5B dataset (Rombach et al. 2022), and OpenAI's DALL-E 2, a closed source model that uses image CLIP embeddings and runs these through a diffusion decoder with a 3.5 billion parameter GLIDE model (Ramesh et al. 2022). While Stable Diffusion was implemented locally on a Windows laptop, DALL-E was used through the beta image generation API within the visual programming interface, Grasshopper.

1.2. MULTIMODAL MACHINE LEARNING IN ARCHITECTURE

Clip-guided diffusion models are increasingly present within architectural practice. However, they are still in their infancy in both 2D and 3D applications. Bolojan et al. outline the problematic nature of assigning the overall design complexity to AI models, offering complementary GAN models trained with architectural datasets to circumvent limitations of verbal representation for 2D images (Bolojan et al., 2022). ArchiTEXT reveals the possibility of 2D floor plan generation from semantics (Galanos and

Lastovich, 2021), and early 3D examples merge computational tools with early text-image models, AttnGAN, to generate abstract 3D massing from planimetric views (Campo, 2021). Words to Matter uses text-to-image models to generate patterns for fabrication into physical material designs (Yang and Buehler, 2021). Each of these processes reveal a dialogue between different modes, from text to 3D forms or construction elements. PlacemakingAI outlines the use of a ML-based UI to establish a missing communication between citizens or designers and urban stakeholders in the collaborative redesign of urban streets (Kim, Guida, and Garcia 2022). Using graphical user interfaces to aid emerging dialogues with these tools can be seen as a necessary bridge to navigate this exchange of information in early concept design stages.

Several problems unfold within this multimodal process, (1) the overall design complexity of architecture, whether this is a façade or internal condition inherits a reductionist simplification of a text prompt; (2) current tools accessible to the public rely on predefined datasets, therefore forcing use of non-architecture specific datasets; (3) due to dataset limitations the process of feature disentanglement is limited. These limitations however provide opportunities in merging or chaining different models together and creating a space for the interpretation of the resultant 2D images into 3D geometries.

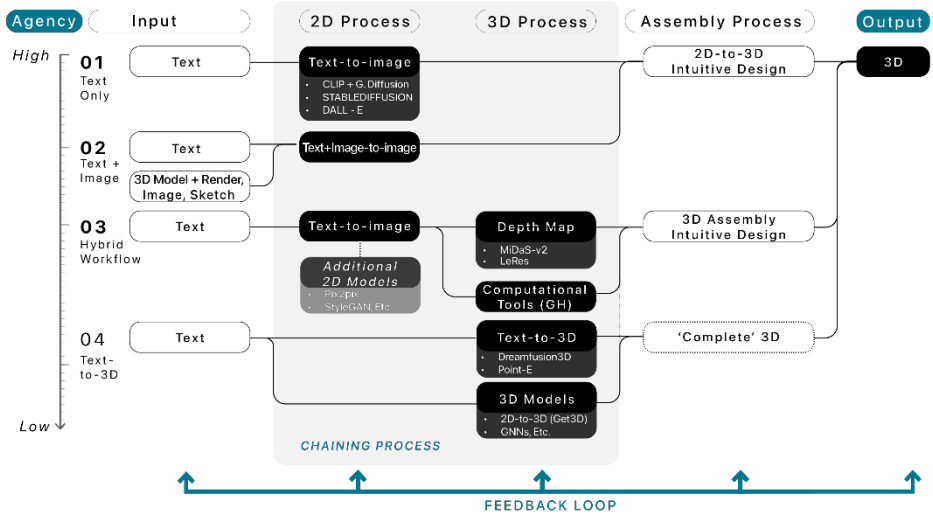


Figure 01. Multimodal Processes: (1) Methodology 1 - Semantic Assembly, (2) Methodology 2 - Semantic + Visual Assembly, (3) Methodology 3 - Hybrid 3D Assembly

To untangle the design process two principal steps were taken: (1) 3D form generation processes were investigated using two diffusion models applied to a design scenario (Figure 1). (2) an intuitive user interface was designed to make these tools more accessible and permit further methodologies. While the first step revealed a negotiation of agency and authorship, the second enabled a simplification of use within existing workflows. This revealed the emerging role of curation of the designer,

through datasets, image selection, and 3D form generation.

2. Methodology

The methodology will cover three text-to-image-to-3D processes applied to a proof-of-concept competition of the MAXXI Grande Extension in Rome, Italy. The process demonstrates how multimodal machine learning models can be integrated with varied degrees of agency within existing design workflows. As an open-call competition, the brief describes a building dedicated to artificial intelligence and the conservation of contemporary art across a site of 3,300m2 located adjacent to the existing MAXXI museum by Zaha Hadid Architects. This program includes a restoration laboratory, a technological laboratory, classrooms, storage, and appropriate circulation within a sustainable and accessible building.

Two CLIP-guided diffusion models are described including Stable Diffusion, and DALL-E 2. Each method presents different ways of integrating these ML models in the process of 3D concept design (Figure 1), accessible through an intuitive user interface within Grasshopper. The DALL-E 2 image generation API was integrated through a custom component, permitting its use in Rhinoceros 3D modelling environment. These will lead to speculations on the limitations and opportunities of these models considering the recent developments in text-to-3D models.

2.1. METHODOLOGY 01: SEMANTIC ASSEMBLY

The first methodology positions text inputs and 3D outputs within an intuitive or 'manual' approach. This process presents a varied agency across several modes of representation. First, the text is created by the designer as a simplified or abstracted description of the design brief. Then, a series of synthetic architectural images is generated by the selected models trained on generic non-architectural datasets. Following an iterative process of image generation and curation, the resultant geometry is finally assembled, in this case, a series of exterior façades, leading to a feedback loop across each step. While this concedes the 2D image generation process to these ML models through a limited semantic representation of form, the process of interpretation of these images is fully constructed manually. Through this process of interpretation, the missing semantic embeddings within each image, for example in representing materiality, styles, proportions, or program is assigned to the designer.

Figure 02. Methodology 1 - Semantic Assembly using Stable Diffusion v1

2.2. METHODOLOGY 01: SEMANTIC + VISUAL ASSEMBLY

In this second methodology, a 3D massing model is first reconstructed and then used to condition the text-to-image process in the first methodology. Environmental, site-specific, and project-specific elements from the brief are translated into 3D forms, for example massing areas, building code, site constraints, and program distribution (indicated in color). While these processes could be automated through computational tools, this manual process of assembly permits a first instance of subjective interpretation of the brief into 3D form. This quantitative understanding of the architectural brief gives greater agency to the architect in defining the input text and images - in this case, renders of the massing (Figure 3).

With each of these inputs, both ML models were used to test out 10 different exterior and interior views, from entrances and glazed facades to circulation spaces and research labs. This permitted greater control in generating images based on specific elements of the building, to then be reassembled in 3D. Figure 3 compares each model, where the semantic specificity of each text generates varied results. While Stable Diffusion permits the adjustment of the model parameters and image settings such as image size, clip guidance scales, the image sampler used, and the seed, DALLE did not permit any adjustments.

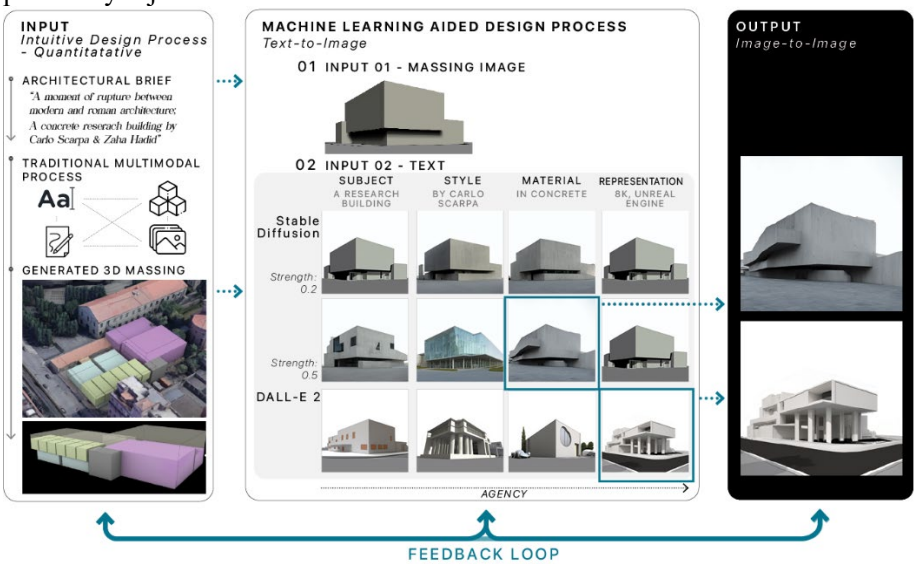


Figure 03. Methodology 2 - Semantic + Visual Assembly with Stable Diffusion and DALL-E 2

2.3. METHODOLOGY 03: HYBRID 3D ASSEMBLY

The third methodology uses 2D feature extraction through depth map models in a

process of assembly to combine plans as spatial logic with exterior and interior 3D spaces to create a concept design. Here, the first two methodologies are hybridized, with the massing logic sketched by hand in plan to constrain the overall massing volume and used as an image input to generate a plan logic. This is then combined with exterior images generated by prompts such as "*A research building by Carlo Scarpa & Zaha Hadid in concrete with a grand entrance, hyper-detailed painting by William Turner*", and interior images depicting "*sequences of open voids*", research spaces, and classrooms. What emerged was a process of assembly where depth map reconstructions of textured 2D images were combined with a new internal diagram of the building (Figure 4).

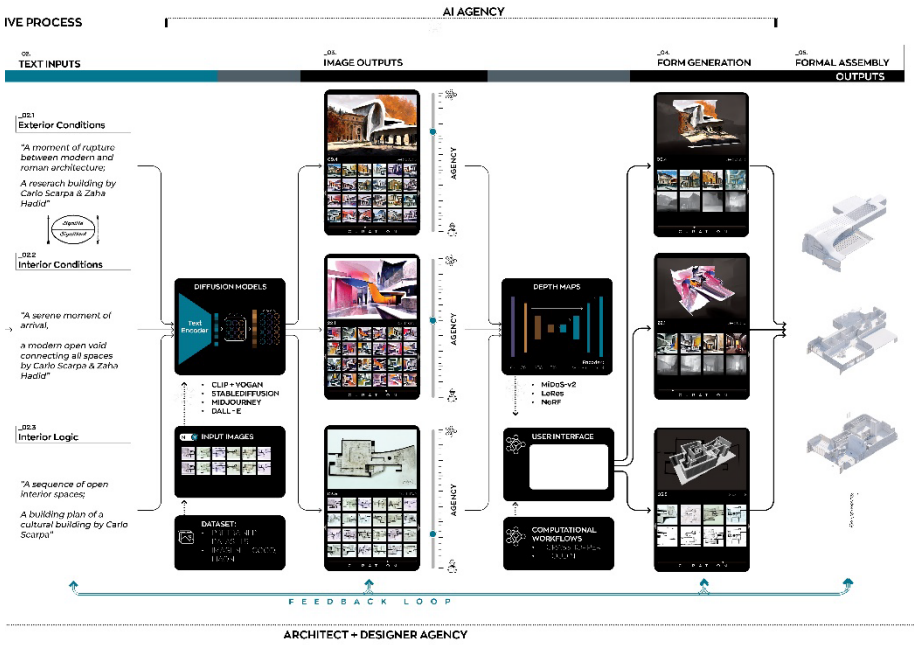
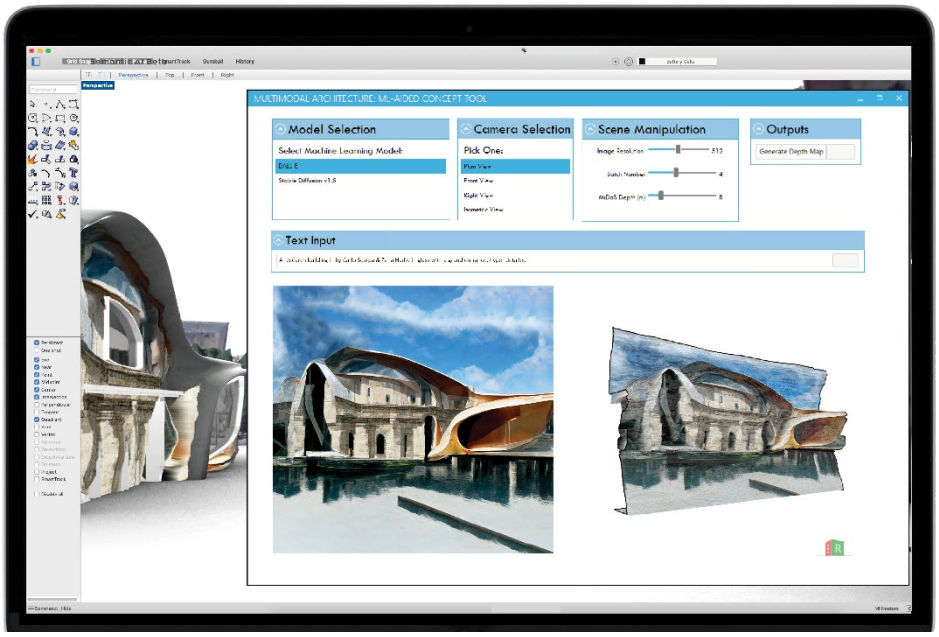


Figure 04. Methodology 2 - Semantic + Visual Assembly of (1) Exterior, (2) Interior, (3) Plan

2.4. USER INTERFACE

A prototype for a user interface was developed to improve the navigation across methods and gain greater agency within the process. Hosted in Grasshopper, this permitted the use of the DALL-E 2 image generation API through a custom component. This included three functionalities, (1) creating images from text prompts, (2) creating variations of images of massing models from the 3D modelling environment (3) generating MiDaS depth maps of output images for a 3D reconstruction. At each step each image or 3D model can be downloaded or modified manually in Rhinoceros, accelerating the design process, and improving the feedback loop of information or modes.

Created using GH_CPython (Rahman 2017), this permitted the use of Python 3.9 with locally hosted libraries, to process the API requests and process the output base 64 strings from JSON to image formats. Inputs include the API key for authentication, the text prompt, the image input, the image resolution size, the model, and output batch count. The user interface was made with the Grasshopper plugin HumanUI, where



users can navigate interchangeably between the Rhinoceros modelling interface allowing users to manipulate generated 3D geometries from depth maps images.

Figure 05. HumanUI User interface to generate 3d outputs

3. Results

This process reveals how we are in a similar moment to how text-to-image models were over a year ago when AttnGAN, or CLIP+VQGAN models were emerging. Early projects reveal how the process of interpretation of generated images was central to architectural proposals (Campo, 2021). Whether through manual image reconstruction or computational workflows, these relied on a distinct process of subjective interpretation due to image qualities. We are increasingly entering into early phases with 3D form-making where lower quality outputs imply increased designer agency. These intuitive processes establish a new dialogue with the support of a user interface and depth map reconstructions, underlining the opportunities for interpretation and curation in these new processes. Figure 6 demonstrates examples of varied agency in which 3D geometry for the MAXXI Museum competition was generated, from an intuitive process resembling traditional multimodal processes - where 3D forms emerge from text, sketches, and images.

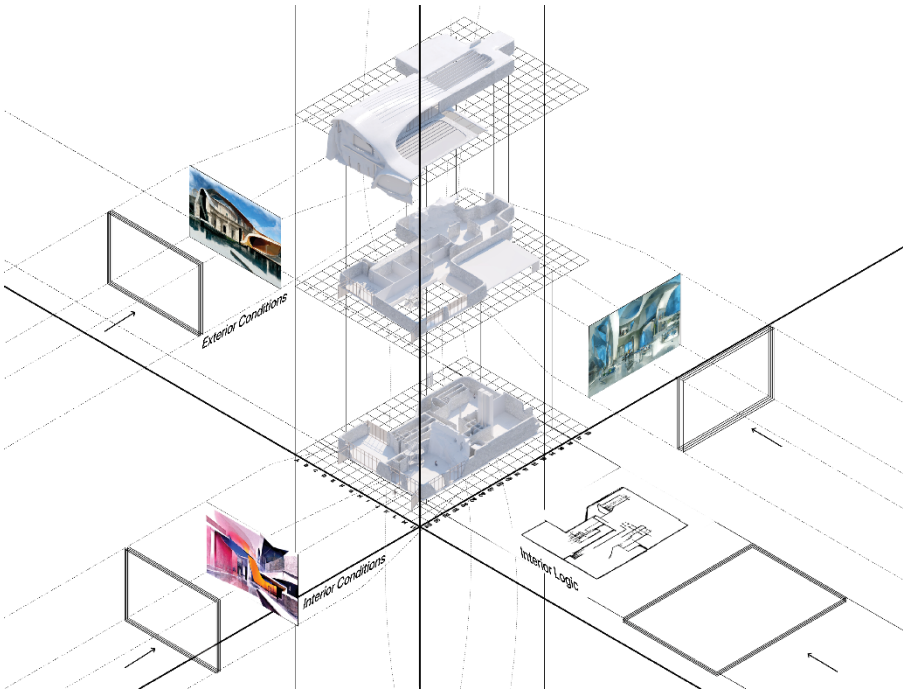


Figure 06. Intuitive to machinic text - to - 3D process examples, each process requires a varied level of interpretation and therefore agency in the process of generating form.

The proposed methodologies intentionally rely on a process of interpretation of the architectural brief. While the internal planimetric logic of rooms and spatial relations can be conditioned through sketch or image inputs, future work will incorporate emergent semantic models that can be described through text. The use of user interfaces will help designers navigate across representational modes, models, and output 3D resolutions. While these methods relate to early design stages, this will increasingly extend across different design stages, embedding generative design tools or performative models leading to a future where we can generate construction documentation from a text prompt alone. This, therefore, changes the role of language and the architectural brief within this multimodal process. Koolhaas states how only at the moment that a concept, ambition, or theme is put into words, we can begin to proceed to think about architecture (Koolhaas, 1993).

The next steps of text-to-3D, such as Dreamfusion (Poole et al., 2022) and Point-E (Nichol et al. 2022), or 2D-to-3D processes, such as NVIDIA Get3D (Jun et al., 2022), demonstrate this increased role of ML in the process of form making with greater resolution and semantic accuracy. These present improvements on the amount of embedded semantic information we can extract from current depth map mesh models, however, these also demonstrate the risks of a reduced agency. Current issues of intellectual property could be overcome through dataset personalization and hybrid intuitive approaches. The use of design-centric user interfaces will nonetheless present

new opportunities for increased feedback loops where designers can navigate or chain models across multiple modes with the necessary disciplinary knowledge.

4. Conclusion

While sentiment around these tools has disseminated fear about the displacement of the architect, this paper demonstrates how the creative possibilities in chaining, or merging traditional multimodal practices with text-to-image and image-to-3D ML models are broad, opening a new creative space of hybrid methodologies. As 2D and 3D CLIP-guided diffusion models develop together with new foundational approaches, it is becoming evident that our agency will shift increasingly towards roles as curators, of trained datasets, input language and images, 3D outputs, and most importantly of the chosen multimodal process connecting intuition and AI tools. Lower resolution and less defined models enable a process of interpretation and assembly, between plans, exterior, and interior conditions.

User interfaces and integrated computational tools permit the democratization of these methods into a 3D modelling software through an image generation API. This promotes a view of artificial intelligence as a tool rather than the commonly related anthropomorphic view of AI as a competitive agent (Eptein et al. 2020). With these new design methods, text and verbal speech find new relevance as a key medium of design, directly implying materiality and form. What emerges is a visual and textual interplay where humans and machines work iteratively through a feedback loop of information. As Tom Markus states, "language is at the core of making, using, and understanding buildings" (Forty, 2004). Ultimately these tools and hybrid processes have the potential to bring forward a collaborative process with machine intelligence, leading to benefits in the built environment and human-centric design solutions.

Acknowledgements

I would like to thank Andrew Witt and Jose Luis Garcia del Castillo Lopez as my advisors during the development of this Spring 2022 Master's Thesis at the Harvard Graduate School of Design. Additional thanks to Tatjana Crossley and Ana Gabriela Loayza for the support along the way.

References

Bolojan, D., Vermisso, E., & Yousif, S. (2022). *Is Language All We Need? A Query Into Architectural Semantics Using a Multimodal Generative Workflow.* 353–362. https://doi.org/10.52842/conf.caadria.2022.1.353

Bonta, J. P. (1990). *Reading and Writing about Architecture*. Design Book Review, 18, 13.

Campo, M. (2021). *Architecture, Language and AI*. Association for Computer-Aided Architectural Design Research in Asia (CAADRIA), 1, 211–220.

Dhariwal, P., & Nichol, A. (2021). *Diffusion Models Beat GANs on Image Synthesis* (arXiv:2105.05233). arXiv. https://doi.org/10.48550/arXiv.2105.05233

Eptein, Z., Levine, S., G. Rand, D., & Rahwan, I. (2020). *Who Gets Credit for AI-Generated Art?* IScience.

Forty, A. (2004). *Words and Buildings: A Vocabulary of Modern Architecture. Thames & Hudson.*

Gallanos, T., & Lastovich, T. (2021). *Architext.* Retrieved December 12, 2022, from http://architext.design/

Gao, J., Shen, T., Wang, Z., Chen, W., Yin, K., Li, D., Litany, O., Gojcic, Z., & Fidler, S. (2022). *GET3D: A Generative Model of High Quality 3D Textured Shapes Learned from Images* (arXiv:2209.11163). arXiv. https://doi.org/10.48550/arXiv.2209.11163

Heumann, A. (2016, January 31). *Human UI* [Text]. Food4Rhino. https://www.food4rhino.com/en/app/human-ui

Huang, J., Johanes, M., Kim, F. C., Doumpioti, C., & Holz, G.-C. (2021). *On GANs, NLP and Architecture: Combining Human and Machine Intelligences for the Generation and Evaluation of Meaningful Designs.* Technology|Architecture + Design, 5(2), 207–224. https://doi.org/10.1080/24751448.2021.1967060

Kim, D., Guida, G., & García del Castillo y López, J. L. (2022). *PlacemakingAI: Participatory Urban Design with Generative Adversarial Networks*. 485–494. https://doi.org/10.52842/conf.caadria.2022.2.485

Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., & Chen, M. (2022). *GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models* (arXiv:2112.10741). arXiv. https://doi.org/10.48550/arXiv.2112.10741

Nichol, A., Jun, H., Dhariwal, P., Mishkin, P., & Chen, M. (2022). *Point-E: A System for Generating 3D Point Clouds from Complex Prompts* (arXiv:2212.08751). arXiv. https://doi.org/10.48550/arXiv.2212.08751

OpenAI. (n.d.). *OpenAI Image Generation API*. Retrieved February 6, 2023, from https://platform.openai.com

Poole, B., Jain, A., Barron, J. T., & Mildenhall, B. (2022). *DreamFusion: Text-to-3D using 2D Diffusion* (arXiv:2209.14988). arXiv. https://doi.org/10.48550/arXiv.2209.14988

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). *Learning Transferable Visual Models From Natural Language Supervision* (arXiv:2103.00020). arXiv. https://doi.org/10.48550/arXiv.2103.00020

Rahman, M. A. (2017, August 1). *GH_CPython* [Text]. Food4Rhino. https://www.food4rhino.com/en/app/ghcpython

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). *Hierarchical Text-Conditional Image Generation with CLIP Latents* (arXiv:2204.06125). arXiv. http://arxiv.org/abs/2204.06125

Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., & Koltun, V. (2020). *Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-shot Cross-dataset Transfer* (arXiv:1907.01341; Version 3). arXiv. https://doi.org/10.48550/arXiv.1907.01341

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). *High-Resolution Image Synthesis with Latent Diffusion Models* (arXiv:2112.10752). arXiv. https://doi.org/10.48550/arXiv.2112.10752

Bernard, T. (1993). *Modes of Inscription*. ANY: Architecture New York, Anyone Corporation. www.jstor.org/stable/41845559

Yang, Z., & Buehler, M. J. (2021). *Words to Matter: De novo Architected Materials Design Using Transformer Neural Networks.* Frontiers in Materials, 8. https://www.frontiersin.org/articles/10.3389/fmats.2021.740754

Zhou, Y., & Park, H.-J. (2021). *Sketch with Artificial Intelligence (AI)—A Multimodal AI Approach for Conceptual Design*. 201–210. https://doi.org/10.52842/conf.caadria.2021.1.201

Zhuang, X. (2022). *Rendering Sketches—Interactive rendering generation from sketches using conditional generative adversarial neural network*. 517–524. https://doi.org/10.52842/conf.ecaade.2022.1.517