

Chapter 11

Building Simulations with Generative Artificial Intelligence



Jon McCormack and Mick Grierson

Abstract In this chapter, we explore the possibilities of generative artificial intelligence (AI) technologies for building realistic simulations of real-world scenarios, such as preparedness for extreme climate events. Our focus is on immersive simulation and narrative rather than scientific simulation for modelling and prediction. Such simulations allow us to experience the impact and effect of dangerous scenarios in relative safety, allowing for planning and preparedness in critical situations before they occur. We examine the current state of the art in generative AI models and look at what future advancements will be necessary to develop realistic simulations.

Keywords Augmented reality · Diffusion models · Generative artificial intelligence · Immersion · Machine learning · Simulation · Virtual reality · Visualisation

11.1 Introduction: A Scenario

It is mid-summer in a small rural town in eastern Australia. You are standing on the spacious wooden veranda of your beautiful home, surveying an expansive vista of wild flora nestled in the valley formed by two distant mountains. It is a clear, sunny day. You feel a hot, dry, gusty wind on your face. All around you, the landscape is parched—heavy rains earlier in the year allowed the surrounding landscape to

J. McCormack (✉)
SensiLab, Monash University, Caulfield East, VIC, Australia
e-mail: jon.mccormack@monash.edu

M. Grierson
Institute for Creative Computing, University of the Arts London, London, UK
e-mail: m.grierson@arts.ac.uk

blossom, and it filled with tall grasses and a rich variety of tinder. But a few months ago, the rains gave way to continuous weeks of well-above-average temperatures. Rain has not fallen in months, and now everything is a faded hue of off-yellow and completely dry.

As you turn to go inside, you hear a far-off roar that sounds like the rumble of thunder. The sound seems to be getting louder. As the wind picks up, you notice an orange glow on the horizon. Within the space of just a few minutes, it becomes quite clear that a massive bushfire is bearing down on you. Thousands of tiny burning embers are blowing around you, mixed with an acrid, grey smoke that chokes the atmosphere, soon making it difficult to see more than a few metres in any direction. The pleasant blue sky has quickly turned pitch black, creating an eerie sense of disquiet.

As the wall of super-heated flames—a minute ago just a pretty deep orange glow in the distance—are now closing in on you and your property at over 100 km/hour, you have just a few seconds left before the wall of fire will consume you. You see several animals running just ahead of the fire front, but less than a second later, they are swallowed by the fire and disappear. A sense of panic and dread kicks in as it feels as though the blood is draining from your body. You rush out to the back of the property to locate the large steel door that leads down to an emergency fire bunker. You pull the vegetation that has grown around the door away as the roar becomes unimaginably loud; the thick black smoke has turned day into the darkest night. You scramble to get into the bunker, pulling the heavy steel door shut, just as an enormous wave of fire engulfs your home and land. You sit inside the dark bunker, the intense roar of the fire still audible, and a wave of super-heat can be felt above. You notice your heart is racing and pounding against the wall of your chest. You take deep breaths, telling yourself to stay calm while trying to convince yourself that you will survive.

A few moments later, the attendant removes the virtual reality headset and bodysuit you were wearing, letting you know that “the simulation is over”, reassuring you that “you are safe now and there’s nothing to worry about”. You suspect that you must look very frightened and distressed as you are asked if you need a few minutes to take in what you have just experienced, before completing your disaster readiness training and heading back home. You are in your town’s local community hall, and it is only a short drive back to your property—the same one that you just experienced burning to the ground with a realism so visceral you are now sweating profusely and in a mild state of shock. This is not an experience you will quickly forget.

11.1.1 Building Simulations

While the scenario just presented is currently largely in the realm of speculative fiction, the technology to produce such a simulation has made rapid advances over the last few years, suggesting that it may shift from speculation to practical realisation

in the next decade. Further, in addition to being able to provide virtual reality (VR) simulations of familiar environments for training purposes, it is potentially possible that such experiences could even be simulated with generative AI using augmented reality (AR) in situ. Currently, however, building realistic simulations for immersive technologies—such as VR and AR—is a complex and time-consuming process.

Three-dimensional (3D) simulations usually begin with modelling the geometry and textures of every object that will appear in the simulation. Despite advances in 3D modelling techniques and the wide availability of existing models online, this is a highly specialised and time-consuming task. Beyond the modelling of physical form, a simulation also needs to model behaviour. For the simulation to be realistic, that behaviour must be accurate to reality or at least be plausible to reality. Lastly, for the simulation to be immersive, it must provoke a strong sense of presence, convincing any participants that what is happening is “real”. This typically involves high-fidelity images, sound, haptics, kinaesthetics and beyond.

Our speculative scenario made use of sensory experience beyond mere visual simulation, incorporating sound, proprioception, kinaesthetic, haptic and even olfactory simulation (e.g. the smell of smoke). Not all of these sensory modalities are currently well synthesised by AI, and a heavy commercial focus on visual and audio synthesis currently dominates the well-known foundational AI models (Bommasani et al., 2021).

11.2 Simulation of Extreme Event Scenarios

In this section, we look at the current state of the art in generative AI and how it might be usefully purposed for immersive simulation.

As we noted elsewhere, “over the last decade, a [number] of innovations in generative machine learning (ML) models have allowed the generation of photo-realistic images of [nonexistent] people (Karras et al., 2018), coherent paragraphs of text (Vaswani et al., 2017), conversion of text directly to [runnable] computer code and, [more] recently, from text descriptions to images (Ramesh et al., 2022), video (Singer et al., 2022; Blattmann et al., 2024), and 3D models (Gao et al., 2022)” (McCormack et al., 2023). Neural radiance fields (NeRFs) (Mildenhall et al., 2020) can synthesise 3D scenes from novel viewpoints using sparse 2D images as input and guided by text descriptions (Zhang et al., 2023). Tools such as these are already being offered to creators through platforms such as NVIDIA’s Open USD-based Omniverse.

These tools are increasingly used in audio-visual production, combining a range of generative AI techniques, including diffusion models (Yang et al., 2023), specialised generative adversarial networks (Iglesias et al., 2023), autoencoders and image-to-image systems (Wang et al., 2018). Initially popular for their potential for still image generation, they have more recently become surprisingly usable for video and 3D scene generation. As noted elsewhere, “systems such as DALL-E 2, MidJourney and Stable Diffusion allow the generation of detailed and complex

imagery from short text descriptions. These text-to-image (TTI) systems allow anyone to write a brief description [(a ‘prompt’)] and have the system respond with a series of images that depict the scene described in the text, typically within 5 [to 30] seconds” (McCormack et al., 2023). More recently, diffusion model-based text-to-image systems have demonstrated rapid advances in both quality and popularity. At the time of writing, these systems can produce high-quality imagery as fast as a person can type in a prompt (Stability.ai, 2023). They can also facilitate image editing and manipulation.

As my team noted, “the obvious source of these systems’ popularity is that they offer something new: being able to generate an image, [video sequence or 3D render] just by describing it, without having to go to the trouble of learning a skill—such as [illustration,] painting, photography, [cinematography or 3D modelling]—to actually make it. And importantly, the quality and complexity of the [media] generated is often [comparable] to what an experienced [professional] human creator could produce, [at least at the surface level. Moreover, generative AI] systems demonstrate a semantic [interpretation] of the input text and can convert those semantics so that (in some cases) they are more-or-less coherently represented in the generated images. This new-found capability has inspired many useful image generation and [manipulation possibilities,] such as ‘outpainting’, where a pre-existing image can have its edges [extended] with coherent and plausible content, or as an ‘ideation generator’, where new versions of a set of input images are generated” (McCormack et al., 2023).

11.2.1 Use in Visual Simulation

The idea that through new generative AI technologies we can construct high-fidelity simulations of real-world events presents a step change in developing simulation systems. Rather than labouring over detailed 3D models, building complex simulations by hand or using digital media such as cinema or photomedia to construct a rich simulation experience, generative AI potentially presents the opportunity to deliver high-fidelity simulations simply by describing them in language.

Current text-to-image (TTI) systems rely on diffusion models. These models are trained by adding noise to a training set, forcing the model to learn how to convincingly reconstruct image representations. This approach has significant advantages in image generation quality over previous methods such as generative adversarial networks (GANs) (Goodfellow et al., 2020). The fundamental innovation of TTI systems lies in the integration of two different approaches—a language transformer model that accepts image descriptions as text and an image generator that synthesises the image. The transformer is usually based on CLIP (Contrastive Language–Image Building Simulations with Generative AI 5 Pre-training) models, a neural network that learns visual concepts from natural language supervision (Radford et al., 2021). This is a significant improvement over previous models such as convolution neural networks (CNNs), which excelled at basic classification of objects in



Figs. 11.1 and 11.2 Images of bushfire created using *Stable Diffusion*

an image but could not recognise more salient concepts such as style, context or semantics. The image generator uses a multi-step process that operates in the image latent space, using a UNet neural network and scheduler. The output of this “diffusion” process is an image tensor that is decoded into an image by an autoencoder.

To illustrate the potential of these systems for simulation, we used an open-source version of Stable Diffusion. Figures 11.1 and 11.2 show two sample AI-generated images created using Stable Diffusion. The prompts used were “national geographic photo of an Australian bushfire, landscape, trees” (left) and “national geographic photo of fire-fighters with a hose fighting a large bushfire” (right). As can be seen, the prompts generate quite “realistic” images that would typically be associated with Australian bushfires. Using phrases such as “national geographic photo” pushes the system into producing high-quality, documentary-like images, as would typically be associated with *National Geographic* (we could have specified “old black and white daguerreotype” or “Banana Fish anime” to completely change the aesthetic style of the image).

This simple example highlights some of the issues with creating prompts: that one needs to be quite specific in the prompt about details such as surface aesthetics, style, context, etc. Such a requirement leads to much of the prompt language containing references to the visual aesthetics of the image: including style, lighting, level of detail, even descriptions of camera lens focal lengths, angle or position of the shot and other various cinematic conventions. The necessity of providing such detail on surface aesthetics, composition, etc. mirrors CLIP’s ability to capture these image qualities as general image features irrespective of the objects depicted in them.

Figure 11.3 shows another example of generative AI simulation of flooding events. To create these images, the following prompts were used: “national geographic photo of people piling sandbags in an Australian town after flooding” (left) and “national geographic photo of an Australian town after mild flooding” (right). In this example, the prompts are interpreted correctly, but only to a point. The way the sandbags are being piled does not really make practical sense (they would be unlikely to mitigate the effects of rising water), the “people” depicted do not have



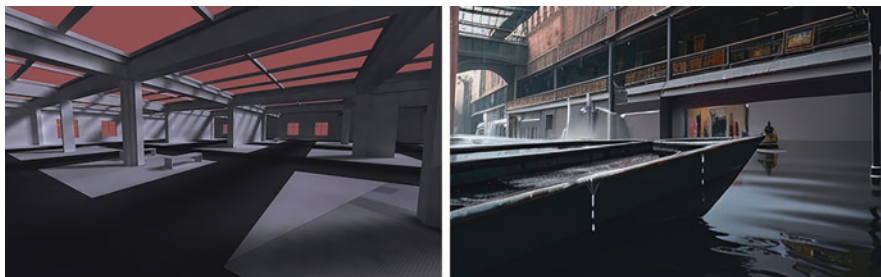
Figs. 11.3 and 11.4 Images of flooding created using Stable Diffusion

natural poses or body parts, the layout of the town is impossible, and so on. An obvious solution to these issues would be to provide more specific directions.

However, providing too much detail on the objects and their relations in the scene does not necessarily translate into the generated image. Using Fig. 11.1 as an example, if we were to modify the prompt used in the second image to be more specific about the exact number of firefighters, specifics of their individual poses, details of their uniforms or their specific location in the image, the results are unlikely to directly match the prompt. For example, if we specified “five firefighters”, we may get any number between two and ten or more. This is due to the way CLIP converts the input prompt into a latent embedding.

Another potential approach is to use an image-to-image (ITI) method. These are largely similar to TTI systems, and most popular diffusion-based TTI platforms offer an ITI mode. Using this approach, imagery can be adapted using a text prompt to make modifications and transformations while preserving important characteristics of the scene. There are a few potential methods for generating realistic scenes more easily using ITI. For example, simple 3D environments can be created using existing 3D models and then modified using generative AI to render more realistic environments for simulation. This is demonstrated in Fig. 11.3 using Stable Diffusion. A basic 3D environment is transformed into a more complex scene through the use of a text prompt while preserving the overall structure and characteristics of the scene. There are a few potential problems with this approach. For instance, it can be a challenge to control the content that might appear in the generated scene due to the diversity of images used to train the model. In Fig. 11.3, the generated image features a boat in place of a shadow, which appears in the input image. This can be mitigated by adjusting the strength of the transformation. Another approach is to use fine-tuning methods, including the creation of custom low-rank adaptors (Hu et al., 2021) to guide content generation more explicitly with examples and custom embeddings. Despite offering considerably more control, problems with content consistency still cannot be entirely avoided with approaches such as these (Figs. 11.5 and 11.6).

Aside from adjusting and controlling content with prompting, there are potentially other, more direct content control methods that may be more practical. One



Figs. 11.5 and 11.6 A low-quality 3D render is shown on the left, with an image-to-image version on the right generated by Stable Diffusion. The ITI prompt was simply “A flooded art gallery”

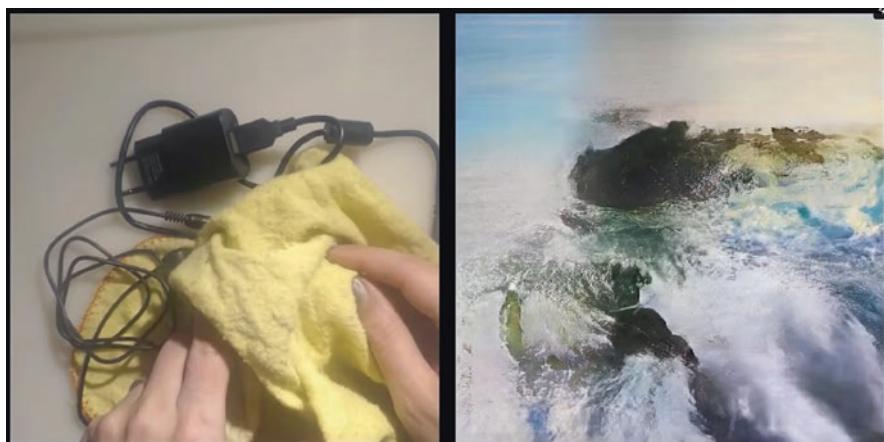


Fig. 11.7 Image from *Learning to see: Gloomy Sunday, 2017*. A simple ITI approach is used to create a real-time transformation of webcam input. The input image on the left-hand side is transformed into imagery of the sea and rendered in under 20 milliseconds, allowing for interactive control of the generative AI system

could use specifically curated models, such as fire and/or water generators. These models could be specifically trained to add photo-real fire and water effects on input images using ITI methods, similar to the approach taken in Memo Akten’s *Learning to See: Gloomy Sunday* (Akten et al., 2019), as shown in Fig. 11.7. Here, models were created from datasets of open water, fire, flowers and other categories of images, and these were then used for generation.

Using this approach, models need not understand a wide range of different kinds of imagery, have knowledge of context, nor draw on complex text prompts in order to guide generative image models. Instead, input images could be adapted by models with highly limited yet detailed and carefully engineered outputs. These kinds of models are far simpler than large, monolithic text-guided diffusion systems and as a result can be run in real time in high definition on modern hardware. Furthermore, as these models could be specifically tailored to the problem of disaster simulation,

they could also contain information on behaviour, for example, through the application of fluid dynamics models within the content generation pipeline.

This approach could be applied to the generation of content for simulations in AR environments. Focused, bespoke generative AI models such as those based on ITI approaches described above could quite easily be deployed in order to transform real-time stereoscopic image feeds from AR headset cameras, with the transformed output rendered directly to the headset in real time. This creates the opportunity for generating disaster simulations in real-world environments, where participants experience photo-real 3D generative AI simulations superimposed on the world as it exists. This requires less computing power as only the transformed elements need rendering. In addition, it could provide the opportunity to conduct disaster readiness training in situ with greater fidelity, allowing firefighters and other emergency services personnel to model scenarios in specific locations where there are known challenges, such as in public buildings and city centres and with potentially vulnerable communities.

Another advantage of specifically engineering models for simulation is that they can be more transparently usable and explainable than currently popular contemporary text-guided diffusion systems. For example, it can be challenging for users of contemporary generative TTI systems to understand precisely which aspects of their prompt may be having the greatest impact.

As illustrated in Table 11.1, human prompt writers often tend to over-equate the complexity, poetics and quality of the prompt with that of the resultant generated

Table 11.1 An example of differences between human-generated prompts and machine-based description


Prompt	Generated image	Description
Imagine a dream-like scene where reality blurs and the boundaries between woman and peacock dissolve. Sketch a woman's body full of delicate vulnerability, her features soft and poetic. Let the peacock's head emerge, seamlessly integrating with its essence, symbolising the deep connection with the world of colours of the peacock's tail. Use the impasto technique to add a tactile quality, allowing the viewer to visually feel the texture of the artwork. Set against a deep, velvety canvas of dark blue on a black background, this ethereal combination creates a sense of enchantment, encouraging viewers to explore the depths of their imagination		Painting of a woman with peacock feathers on her head

image. The table shows a relatively long human-authored prompt (left) and the resultant image generated by Midjourney (middle). We ran this image through a state-of-the-art BLIP (Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation) image captioning model (Li et al., 2022), which can generate a descriptive caption for any image. The captioning model gives an overall description of the image, much like the way a human would, as the model recognises not only the objects in a scene but also the relationships and basic surface aesthetic properties. Table 11.1 (right) shows the results of this “machine eye” perception of the image. As can be observed, this is a far more direct and literal description of the image than that of the original prompt.

11.3 Accuracy and Ethics

In addition to issues of transparency and explainability in contemporary TTI systems, there are considerable problems with both the accuracy and also the ethical grounding of many large foundation model-based generative AI tools. Such systems are not being developed within the constraints of well-defined use cases, nor with domain building simulations with generative AI-specific requirements in mind. Evidence points to the need for the development of generative AI that is specifically tailored to the problems of simulation.

For example, careful examination of the images shown in Figs. 11.1, 11.2, 11.3 and 11.4 reveals problems of accuracy and bias. Due to both the data used for training and the nature of diffusion systems, references such as “Australian landscape” tend to get translated into cliched representations based on statistical averages in the training data. A number of studies have analysed AI-generated images, identifying a wide range of biases, such as under-representing certain race groups (Bansal et al. 2022; Naik & Nushi, 2023), cultural gaps (e.g. over-representing specific nations (Naik & Nushi, 2023), or the reinforcement of stereotypes (e.g. “a photo of a lawyer” consistently showing a white male) (Bianchi et al., 2023).

A recent analysis of 3000 images generated by Midjourney using prompts to depict national identities also highlighted tendencies towards bias and stereotypes prevalent in generative AI systems. For example, prompting an image of “New Delhi’s streets” generated images that were mostly portrayed as polluted and littered (Turk, 2023). This perpetuates cultural norms that are prevalent in training datasets while under-representing less stereotypical and non-Western aspects of culture, society and landscape. Although some researchers have proposed ways to mitigate these effects, such as adding specific phrases, e.g. “irrespective of gender” (Bansal et al., 2022), or through the use of more specific prompts to mitigate bias, these mitigation strategies are often ineffective (e.g. despite explicitly mentioning words such as “white”, “wealthy” or “mansion”, Bianchi et al. (2023) report that Stable Diffusion continues to associate poverty with people of colour).

There is emerging research exploring ways that AI systems can be potentially better designed through the inclusion of those with domain expertise in fields where models will be deployed. Co-production of ML systems is a developing international field that attempts to respond to risks including those mentioned above. Recent work (e.g. Grabe et al., 2022) indicates existing research on AI system design does not adequately address design challenges posed by AI. They propose a method for understanding the potential complexities of design through two specific features: uncertainty regarding system capability, as exemplified by the lack of system transparency highlighted above, and output complexity, which, as we have described, is a fundamental problem for TTI approaches using foundation models. Other work (e.g. Mucha et al., 2020) highlights the importance of creating AI interfaces tailored to users' needs and of gaining feedback from users early in the design process, supporting the fundamental principle that generative AI systems for simulation should be specifically designed through collaboration with domain experts and that this approach is vastly preferable to the use of existing general-purpose TTI systems in the context of simulation design.

11.3.1 Data Laundering

As we noted elsewhere, “one of the key factors that contributes to the capability of TTI models is their access to [very large] datasets used for training and validation. Achieving the visual quality and diversity that they are capable of reproducing requires a [vast] corpus of human-created imagery, which is typically scraped from the internet, in a practice that has been dubbed ‘data laundering’. Scraped datasets—which [may] include copyrighted media—rely on special exemptions for ‘academic use’ to avoid any legal barriers preventing their use, or for copyright owners to claim against (Baio, 2022). For [example], Stability AI (the creators of Stable Diffusion) funded the Machine Vision & Learning research group at the Ludwig Maximilian University of Munich to [undertake] the model training and a small [not-for-profit] organisation, LAION, to create the training dataset of approximately 5.85 billion images, many of which are [copyrighted], and in general appropriated for this purpose without the image [creators’] direct permission” (McCormack et al., 2023).

We further noted that artists have “raised [concerns] about the ethical and moral implications of their work being used in such systems. These concerns include the appropriation of an [individual] artist’s ‘style’, mimicry, and even the replacement of a [specialist] human artist or illustrator. Furthermore, there is [currently] no easy way to be excluded or removed from such datasets” (McCormack et al., 2023), and any mechanisms are generally “opt-out”, meaning that unless you take action to prevent your own data from being excluded, it is considered fair game for scraping. The use of copyrighted material in AI training data is currently being tested legally in several different countries. Governments may need to draft new legislation to deal with these issues, as has already happened in the European Union.

11.3.2 *Copyright Issues*

The use of copyrighted images in datasets highlights the question of whether training models on copyrighted data should be considered plagiarism or a form of copyright infringement. As we pointed out elsewhere, “being able to easily generate an image in [a specific] artist’s or [house] style [(e.g. ‘National Geographic’)] without paying for that artist to create it (or paying any royalties or licensing fees), allows users of such technology to bypass the traditional economic, legal and moral frameworks that have supported artists and businesses traditionally. Generating copyright-free images immediately for commercial use without the cost or time involved in securing copyright from a [human] artist may become an attractive proposition, raising the interesting legal [question of who would be the defendant in any copyright infringement case brought about by this scenario]” (McCormack et al., 2023).

11.3.3 *Making AI “Safe”*

Beyond ethical questions involving the sources of data and representational bias are the mechanisms by which many large companies try to ensure that generative AIs are “safe”. Many models are augmented with what is known as “Reinforcement Learning from Human Feedback” (RLHF), where outsourced workers in developing economies are paid to “sit for several hours every day watching videos of harmful content and analyzing textual descriptions of hate speech, sexual violence, bestiality, and violence” (Ngila, 2023). This human tagging or classifying of unsafe content is used to train additional AIs that filter results to prevent the underlying generative system from showing harmful content.

Some people have already developed psychological dependencies or been prompted to take real-world action following advice from generative AI systems, with both positive and negative results, including suicide, divorce or self-harm. As models become even more sophisticated, we are likely to see new forms of human–AI relationships with potentially dangerous results. In the context of simulation, there are a number of important considerations for the simulation to be credible. Generative AI suffers from what is euphemistically referred to as “hallucinations”—factually incorrect or erroneous results. The implications for a generative AI “hallucinating” in a simulation context can range from benign to catastrophic, depending on context and situation. For example, the “Australian town” depicted in Fig. 11.4 does not exist, and no real town would be structured in the way it is depicted. Simulations may be speculative, allowing us to ask, “what if...”, but if the answer is based on factual inaccuracies, the value of the simulation may be worthless.

11.4 Conclusion

These issues, when considered in the round, support the overall conclusion that specific, tailored bespoke generative AI models could offer significant advantages over large, monolithic generative AI tools in the context of disaster simulation. They are more controllable in terms of content, as the training process can effectively constrain their output to a known selection of imagery labels. They are more efficient, being able to transform multiple high-resolution video streams in real time on a single modern computer having a relatively modest capacity. They could be used to augment existing, low-quality 3D scenes with photo-realistic real-time generative AI incorporating relevant and plausible behaviour. They could also be used to render stereoscopic photo-real AR experiences for emergency readiness training in real-world environments. They are potentially more transparent, being trained on known data that could be specifically selected by domain experts. As a result, they are less likely to generate out-of-domain hallucinations, potentially offering greater accuracy, avoiding potential copyright infringement and mental health risk to those working with them.

11.4.1 *Limitations: Multimodal AI*

In this chapter, our main focus has been on exploring the use of AI to generate realistic imagery for visual simulation. As discussed, many contemporary systems use text prompts to generate output—e.g. text to image, text to video and text to 3D model—and in these ways are not too distinct from a text-based search. We have also explored how more bespoke generative AI systems can potentially play a significant role in the future. However, an obvious current limitation of this analysis is that the interaction with such generative AI systems is uni-modal.

However, as our simple scenario in Sect. 1 demonstrates, an immersive simulation is a multimodal experience, encompassing multiple senses and ways of interacting. Multimodal interaction has been well studied from a human–computer interaction perspective (see, e.g. McCormack et al., 2018 for an overview). Recently, multimodal generative AI systems have been gaining traction. These systems consider multiple modes of input and output (e.g. text, image, video) allowing cross-modalities to be considered. For example, Google DeepMind recently announced a new AI platform, which they call “Gemini” that allows “reasoning seamlessly across text, images, video, audio, and code” (Google DeepMind, 2023). While still in development, multimodal AI systems have the potential to analyse scenes or environments and to then ask questions that would require expertise (“how safe would this exit be in a fire?”, “where is the safest place to go if this area is under imminent threat of flooding?”). It may be that these multimodal systems are better able to generalise scenarios as a result of constructing representations from a greater number of dimensions, for example, combinations of sound, image and text. The

capability and power of such systems is potentially enormous, and more research is needed in order to understand how they might one day be deployed for the purposes of simulation.

References

- Akten, M., Fiebrink, R., & Grierson, M. (2019). *You are what you see*. SIGGRAPH.
- Baio, A. (2022). AI data laundering: How academic and nonprofit researchers shield tech companies from accountability. *Waxy.org*. <https://t1p.de/lssgi>. Accessed 15 Dec 2023.
- Bansal, H., Yin, D., Monajatipoor, M., & Chang, K. W. (2022). How well can text-to-image generative models understand ethical natural language interventions? In Y. Goldberg, Z. Kozareva, & Y. Zhang (Eds.), *Proceedings of Conference on Empirical methods in natural language processing* (pp. 1358–1370). ACL.
- Bianchi, F., Kalluri, P., Durmus, E., Ladhak, F., et al. (2023). Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proceedings of ACM Conference on Fairness, accountability and transparency* (pp. 1493–1504). ACM.
- Blattmann, A., Dockhorn, T., Kulal, S., ..., & Rombach, R. (2024). *Stable video diffusion: Scaling latent video diffusion models to large datasets*. <https://t1p.de/mj6wf>. Accessed 15 Dec 2023.
- Bommasani, R., Hudson, D., Adeli, E., Altman, R., Arora, S., ..., & Liang, P. (2021). *On the opportunities and risks of foundation models*. *Arxiv.org*. <https://t1p.de/gcl9r>. Accessed 15 Dec 2023.
- Gao, J., Shen, T., Wang, Z., Chen, W., et al. (2022). Get3d: A generative model of high quality 3D textured shapes learned from images. *Advances in Neural Information Processing Systems*. *Arxiv.org*. <https://t1p.de/ji55g>. Accessed 15 Dec 2023
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., et al. (2020). Generative adversarial networks. *Communications of the ACM*, 63(11), 139–144.
- Google. (2023). Welcome to the Gemini era. Google. <https://t1p.de/06u8v>. Accessed 15 Dec 2023.
- Grabe, I., Duque, M., Risi, S., & Zhu, L. (2022). Towards a framework for human-AI interaction patterns in co-creative GAN applications. In *IUI Workshops* (pp. 92–102) *SemanticScholar.org*. <https://t1p.de/21dkt>. Accessed 15 Dec 2023
- Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., ..., & Chen, W. (2021). *Lora: Low-rank adaptation of large language models*. *Arxiv.org*. <https://t1p.de/berdx>. Accessed 15 Dec 2023.
- Iglesias, G., Talavera, E., & Ivarez, A. (2023). A survey on GANs for computer vision: Recent research, analysis and taxonomy. *Computer Science Review*, 48, 100553.
- Karras, T., Laine, S., & Aila, T. (2018). *A style-based generator architecture for generative adversarial networks*. *Arxiv.org*. <https://t1p.de/cjn6s>. Accessed 15 Dec 2023.
- Li, J., Li, D., Xiong, C., & Hoi, S. (2022). *Bootstrapping language-image pre-training for unified vision-language understanding*. <https://t1p.de/jvybx>. Accessed 15 Dec 2023.
- McCormack, J., Roberts, J., Bach, B., Freitas, C., et al. (2018). Multisensory immersive analytics. In K. Marriott, F. Schreiber, T. Dwyer, K. Klein, et al. (Eds.), *Immersive analytics* (pp. 57–94). Springer.
- McCormack, J., Cruz Gambardella, C., Rajcic, N., Krol, S. J., et al. (2023). Is writing prompts really making art? In C. Johnson, N. Rodríguez-Fernández, & S. M. Rebelo (Eds.), *AI in music, sound, art and design* (pp. 196–211). Springer.
- Mildenhall, B., Srinivasan, P., Tancik, M., ..., & Ng, R. (2020). *Nerf: Representing scenes as neural radiance fields for view synthesis*. *Arxiv.org*. <https://t1p.de/zmsgb>. Accessed 15 Dec 2023.
- Mucha, H., Robert, S., Breitschwerdt, R., & Fellmann, M. (2020). *Towards participatory design spaces for explainable AI interfaces in expert domains*. Fraunhofer Institut. <https://t1p.de/62e5k>. Accessed 15 Dec 2023
- Naik, R., & Nushi, B. (2023). Social biases through the text-to-image generation lens. In *Proceedings of AAAI/ACM Conference on AI, ethics & society* (pp. 786–808). ACM.

- Ngila, F. (2023). OpenAI underpaid 200 Kenyans to perfect ChatGPT—Then sacked them. *Quartz*. <https://t1p.de/ftdpc>. Accessed 15 Dec 2023
- Radford, A., Kim, J., Hallacy, C., ..., & Sutskever, I. (2021). *Learning transferable visual models from natural language supervision*. *Arxiv.org*. <https://t1p.de/z25q7>. Accessed 15 Dec 2023.
- Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). *Hierarchical text-conditional image generation with clip latents*. *Arxiv.org*. <https://t1p.de/12kic>. Accessed 15 Dec 2023.
- Singer, U., Polyak, A., Hayes, T., Yin, X., ..., & Gafni, O. (2022). *Make-a-video: Text-to-video generation without text-video data*. *Arxiv.org*. <https://t1p.de/bex23>. Accessed 15 Dec 2023.
- Stability.ai. (2023). Introducing sdxl turbo: A real-time text-to-image generation model. *Stability AI*. <https://t1p.de/k1xqx>. Accessed 15 Dec 2023.
- Turk, V. (2023). *How AI reduces the world to stereotypes*. *Restofworld.org*. <https://t1p.de/9rm40>. Accessed 15 Dec 2023.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30. *Nips.cc*. <https://t1p.de/70qm4>. Accessed 15 Dec 2023
- Wang, T., Liu, M., Zhu, J., Liu, G., et al. (2018). Video-to-video synthesis. *Advances in Neural Information Processing Systems*. <https://t1p.de/mm6ua>. Accessed 15 Dec 2023
- Yang, L., Zhang, Z., Song, Y., Hong, S., ..., & Yang, M. (2023). *Diffusion models: A comprehensive survey of methods and applications*. *Arxiv.org*. <https://t1p.de/zx5sp>. Accessed 15 Dec 2023.
- Zhang, J., Li, X., Wan, Z., Wang, C., & Liao, J. (2023). *Text2nerf: Text-driven 3D scene generation with neural radiance fields*. *Arxiv.org*. <https://t1p.de/339xt>. Accessed 15 Dec 2023.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

