ACTA GRAPHICA 2026; 34 (2)
ISSN: 0353-4707

Original Scientific Paper

Generative AI Based Artwork: Self-Hosted Workflows for Images and Video Production

Dominique Cunin [1,*]

[1] École Supérieure d'Art et Design de Grenoble-Valence (ESAD•GV), 26903 Valence Cedex 9, France

* Correspondence: dominique.cunin@esad-gv.fr

Abstract: Generative AI technologies are at the centre of many debates. To better understand the challenges these technologies raise in the arts, we conducted a creative project based entirely on open-source, self-hosted models. This practice-based research approach in digital arts and technologies enabled us to co-produce *Pierrot Lunaire*, based on Arnold Schönberg's musical work, as part of a large-scale European project exploring AI and creativity. In this article, we present a detailed and comprehensive study of the workflows we discovered and implemented and suggest some of the consequences of introducing these technologies into our artistic practice.

Keywords: art; generative AI; self-hosted AI; practice-based research; AI-enhanced workflow; open-source creative technologies; contemporary music.

1. Introduction

In this paper, we examine in detail the production process of an artistic work created entirely using contemporary AI image-generation technologies, self-hosted on dedicated machines. We start by explaining the context of this work, *Pierrot Lunaire*, a practice-based research project in art and digital technology linked to the creation of educational content. After reviewing the fundamental principles of AI image generation technologies, we detail our workflows to demonstrate their novelty in graphic and visual creation.

The *Pierrot Lunaire* project is the result of a unique set of circumstances. As a professor specializing in interactivity and computer programming for the arts and design, as well as being involved in research and artistic creation, this project is discussed here from the perspective of the Valence School of Art and Design (ESAD•GV). ESAD•GV got the opportunity to participate in a large-scale project coordinated by the EconCult research laboratory at the University of Valencia in Spain, under the Digital Europe Programme (DIGITAL-2024-ADVANCED-DIGITAL-07): AI Supported and Enhanced Creativity for the Triple Transition (AI-SECRETT) [1].

This project aims to explore how AI can support creativity and innovation across key sectors of the European economy while contributing to the triple transition: the combined transformation toward digitalization, environmental sustainability, and inclusive social development. Its goal is to establish a joint Master's program among higher education institutions in Europe and to make online training modules available for lifelong learning to professionals seeking to acquire new skills related to the introduction of AI in many economic sectors.

For our contribution to this project, my initial concern was, at first glance, a trivial one: to create educational content about recent and still-emerging technologies, one must first be trained and have a deep understanding of these technologies, know them through hands-on experience to be able to contextualize them and construct relevant lessons, develop a critical mindset, and recommend toolkits that are valid for both experimentation and learning as well as in professional life.

That's why I decided to create an artwork (in collaboration) that would, through its very production, help identify the tools, methods, and workflows that a new generation of technologies based on machine learning via

artificial neural networks is rapidly bringing into our daily lives.

Regarding practice-based research in art and design, although this article is not the place for a discussion of the term "research-creation," we suggest referring to the definition provided by the Social Sciences and Humanities Research Council of Canada: "A research approach that combines creative and academic research practices and fosters the production of knowledge and innovation through artistic expression, scientific analysis, and experimentation. The creative process, which is an integral part of the research activity, enables the creation of substantial works across various art forms. [...]" [2].

The *Pierrot Lunaire* project emerged from a collaboration with LUX, Valence's national theatre, for a concert on May 12, 2026. The Conservatoire à Rayonnement Départemental de Valence performed Arnold Schönberg's 1912 musical work, enhanced by AI-generated images and videos projected during the performance. The project consisted of two phases: an exploratory phase to investigate AI's creative potential, followed by a production phase to create the visual accompaniment. For the production and delivery of our artistic proposal, I worked with Mayumi Okura, with whom we form the artistic duo Acronie[3].

Figure 1. View of the *Pierrot Lunaire* show on May 12 2026, LUX Valence, Acronie, Esad•GV. ©Acronie.

2. Materials and Methods

2.1. *Pierrot Lunaire* and Its Era

Pierrot Lunaire is a musical masterpiece composed by Arnold Schönberg in 1912. It is inspired by the eponymous collection of poems, *Pierrot Lunaire*. *Rondels bergamasques*, by the Belgian poet Albert Giraud, published in French in 1884. The poems are associated with the Symbolist artistic movement, which influenced all artistic practices in the 1880s and 1890s (novels, painting, sculpture, music, theater, etc.). Here, Pierrot is indeed the theatrical persona from the commedia dell'arte, known in popular culture as a clown with a white face and costume—naive and sad, madly in love with the beautiful Colombine, who, for her part, cares only for Arlequin, a colorful, optimistic yet mischievous character. Cassandra, another archetypal character from Italian theater—an old man destined to be mocked and sometimes portrayed as Colombine's father—also appears in several of Giraud's poems.

Albertine Zehme, a cabaret performer who commissioned the work from the composer, was in search of "absolute freedom of sound," which led Schönberg to write a most unique work: spoken-sung —Sprechgesang, the use of twelve tones (dodecaphonism), and a cyclical structure of 3 x 7 poems—are all hallmarks of this pioneering and radically new work, which the Symbolist poems, freely translated into German by Otto Erich Hartleben in 1893, imbued with a dark, nightmarish atmosphere. The arrangement of the poems, which does not follow the order of the original collection, is divided into three cycles: Pierrot, first consumed by romantic fantasy, then by blasphemy in a nightmarish world, and finally by nostalgia for a fabulous past to which he dreams of returning.

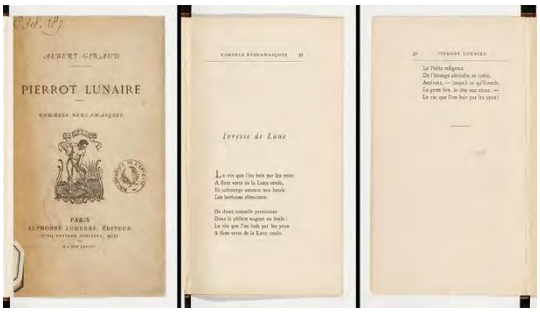

Figure 2. The first pages of *Pierrot Lunaire. Rondels bergamasques* by Albert Giraud, 1884, as found on https://gallica.bnf.fr

Let us briefly review the socio-artistic context of this work by Schönberg. Before 1912, several technological breakthroughs had already taken place, and their impact on the arts and society can still be felt today. Between 1826 and 1839, photography was discovered, explored, and consolidated as a technology for capturing images of reality in the form of

images on paper. The practice of photography as an art form developed and asserted itself over other visual art forms: painting, in particular, was profoundly disrupted by this new imaging technology. Indeed, the accurate reproduction of reality—which had been one of painting's functions—is now mechanically achieved in a very short time and with great precision by the camera, as the development of negatives on photographic paper allows for the reproduction of what the camera "saw." A new image industry emerges, compelling painting to rethink itself. It is well known that Impressionism and Pointillism in painting are a direct consequence of the birth and adoption of photography by society. Symbolism, Surrealism, Expressionism, and Abstraction in art are well-known developments linked to this technological revolution, as well as to another invention.

1895 marks the historic birth of cinematography, following a series of technical experiments that all attempted to add a third dimension to photographic images: the flow of time. The moving image gave rise to a new art form; animation opened up new perspectives on the act of representation and marked a radical shift in the relationship to images for both creators and audiences.

It is also worth recalling the economic and political context here: tensions between countries were rising dangerously; the colonial system was causing trauma and fostering competition between countries that were increasingly arming themselves; empires were fragmenting; and alliances made provocations from one government to another highly volatile. In 1914, the situation shifted irrevocably, sparking a bloody and destructive armed conflict in Europe: The Great War of 1914–1918 began two years after the premiere of *Pierrot Lunaire*. We see just how much this work emerged from a turbulent context—both politically and technologically [4]—and opened the world to "modernity."

We will explore the similarities between our current era and the context in which *Pierrot Lunaire* was created, and how this reality has served as a source of inspiration and a point of reference for our artistic project.

2.2. AI and Contemporary Disruptions

Artificial Intelligence technologies have become part of everyday life in just a few years. Based on machine learning techniques, they work by leveraging a vast amount of annotated data—a vast collection of information organized in relation to one another within an abstract digital space using learning algorithms and condensed into a "model." This space, referred to as latent, can be viewed as a reservoir of potential that can be queried to generate new forms based on those contained within the model. When discussing the use of these models, the term "inference" is employed. In the case of language models, the reservoir is built from texts whose components (letters, words, sentences, paragraphs, etc.) are analysed statistically: the greater or lesser probability that one word follows another is encoded as a digital value (a weight). But it is also with words and sentences that a human user will be able to "query" the model to have it generate sentences in response. It is therefore the relationship between the words and sentences provided as input to the system that will influence how the weights are deployed in the model, and thus what it produces as output—that is, a response in the form of text.

When it comes to images, the technical principle is similar: millions (and millions) of images are analysed, segmented, and annotated, and then an algorithm is trained to reproduce these images from noise (an image made up of randomly coloured pixels). Encoded into a model that, here again, contains all the parameters (the weights) needed to reproduce all the images with which it was trained, a particularly sophisticated form of mixture can be produced when the model is queried using text. Starting from a noise image, the model will select arrangements among the data present in its latent space to compose a proposed image. Translated into pixels, this latent image becomes a digital image like any other.

It seems to us that the emergence of these technologies and their rapid adoption are part of a widespread crisis of representation. It is probably no coincidence that ChatGPT was unveiled to the general public in 2022. One might see a convenient date coincidence here with *Pierrot Lunaire* (1912–2022 = 110 years), but it is rather the emergence from the

COVID-19 crisis that raises questions for us here: after more than two years of lockdowns organized clumsily and brutally imposed on populations worldwide, arbitrarily curtailing individual freedoms in a climate of fear of an unknown and supposedly deadly disease, the trust that citizens placed in their leaders—already largely eroded—has been further diminished in a context of widespread doubt. The very fabric of our industrialized societies has been thrown into confusion: the ability of science to respond to a global health crisis has been questioned, with medicine and research never moving fast enough to satisfy politicians and the public. How can we trust the information disseminated by both mainstream and independent media if science itself is unsure of itself? As recipients of this information, what means do we have to verify and authentify the statements circulating on official and unofficial channels?

Add to this the increasingly alarming reports on the state of our planet's climate, which compel us to envision the inevitable destruction of our living environment and, consequently, the obligation of current generations to prepare for a difficult and painful struggle for survival in a future that is drawing ever closer—a future where it will be too hot, too cold, and where food and water will be in short supply— the resurgence of major armed conflicts marked by Russia's invasion of Ukraine in 2022 or the current conflict between the U.S. and Iran, as well as the ongoing collapse of globalization pushing the major world powers into an increasingly aggressive, widespread economic war, we find ourselves in a context strangely similar to that preceding the Great War. In our view, what AI enables and provokes is fully part of this crisis of representation, even giving it a philosophical dimension: we are once again asking ourselves about the very nature of the act of creation.

2.3. Working on Projects with GenAI

In addition to helping popularize these technologies, the press and news media also amplify numerous arguments regarding these recently emerged technologies: jobs could be replaced by AI, and our use of these tools is already said to be having a disastrous impact on our cognitive abilities, leading to cognitive atrophy [5]. Artistic and creative professions in general are particularly affected by these fears of replacement: creating a photographic image of a person in a specific environment to showcase a commercial product now hinges on a single prompt sent with a click from a program instructing an image-generation model to assemble the image, rendering the work of the photographer and designer seemingly obsolete.

But artists can also view these technologies differently. Thanks to the availability of open-source models, it is already possible to set up creative workflows using AI on personal computers. Better yet, it is possible to modify these models and add "skills" to them by fine-tuning their training using specific text or image datasets. Exploring latent spaces then becomes an aesthetic and technical endeavour: there is no longer any need to rely on the services of a private company with questionable ethics regarding the data used for training, and the safeguards (censorship) it imposes no longer apply.

This concept of self-hosted AI (or local AI, referring to technical terms in computer networking—such as wide area network and local area network—as well as to standalone applications running on a computer without the support of a remote server) enables scalable exploration of latent spaces while remaining at the cutting edge of technology. Furthermore, several types of AI can be interconnected, with the entire system running on the same computer. To describe these configurations, the term "Agentic AI" has recently come into use. What is emerging is thus a form of technological art that treats AI as a raw material, just like computer programming, but with a power never seen before. Indeed, latent spaces, as reservoirs of more or less vast potential, place artists in an unprecedented situation, as it is now possible to describe the texts, images, and music we would like to see or hear, and AI models can produce an infinite number of variations thereof. It is therefore up to the creators to select the images relevant to a given artistic project and to set up complementary processing chains as needed. We are therefore facing a profound paradigm shift in which the selection of a model and the textual description used to activate it have become the means by which images are generated. But how are the

prompts used to generate images and text interpreted?

A term has emerged with the widespread adoption of AI: "hallucination." This refers to the moment when the inference settings cause the model to produce results that fall outside the scope of the training data. In language models, this produces "false" knowledge, resulting from a mix of data that generates new but undesirable information (non-existent URLs, historical events that never happened, etc.). For images, this produces unexpected, unwanted, monstrous, unsettling, or even disturbing forms, which may lead us to believe we are witnessing a kind of daydream—or rather, a nightmare. Creating a nightmare machine. This is the project we decided to pursue as part of *Pierrot Lunaire*. Our initial goal was to generate, in real time during the concert, images and short videos depicting the 21 poems by Albert Giraud. The plan would have been as follows:

Figure 3. Original workflow scheme of the project (NB: realtime generation couldn't be done).

After some preliminary experimentation, we quickly understood that it would be difficult to achieve real-time performance for our project, although solutions are now beginning to emerge. Before detailing the workflows we used and drawing conclusions about the relationship between artistic creation and generative AI, let's review how a diffusion model works to generate images.

2.4. GenAI?

This article does not aim to delve into the specific technical details of what generative AI is. However, it is still essential to fully understand its principles in order to assess its implications for artistic creation.

Machine learning is a long-established field within computer science, and deep learning came to the forefront with the emergence of the first AI chatbots, known as LLMs. The biological metaphor of neural networks is used to create mathematical data structures.

An artificial neuron represents a digital calculation function that can transform data passing through it. When organized and arranged in parallel with one another, neurons form a layer. When data is sent through the neurons, a transformation is applied to it, producing new data. If the function values of each neuron are modified, the result will be different. The configuration of the neurons therefore determines how the input data is processed to produce a specific result.

By stacking layers of neurons one after another, we construct a network of functions, a transformation matrix. The shape of the network—the set of function values for all the neurons (known as "weights")—produces a specific configuration. Deep Learning uses networks with a large number of layers (deep networks) to increase the number of possible ways to process input data. The challenge is to prepare enough possible configurations of the neural network to obtain different results when input data is sent to it. The network must therefore be trained using a large number of known results that we wish to obtain so that it can configure itself through repetition. When it comes to image generation, the basic principle is as follows.

We define the input as an image that we want a network with several million parameters (or weights) to reproduce. At the network's output, we place an image of the same dimensions but composed of pixels with random colours—visual noise. The goal is that, as the image passes through the network, the weights of each neuron change slightly in value. At the end of the calculation, if the image produced does not match the expected result (noise similar to that provided as input), a calculation of the differences between the result and the target is performed. This is a "loss" value, which

will be used to adjust the weights of the neurons (slightly more, slightly less). The lower the loss value, the closer the neural network gets to a configuration that can reproduce the desired output. If the number of neurons (or parameters) is sufficient, such training can be performed with multiple possible outputs—in this case, several noise images associated with the same input image. This approach enhances the neural network's capability, which can be summarized as a specific noise generation function for a reference image.

This is referred to as a "diffusion" principle, inspired by thermal diffusion [6]. If this process is repeated with a dataset consisting of millions of images [7], each paired with different noise patterns, a vast set of configurations becomes embedded in the neural network; this is then referred to as learning, by analogy with human memory. A form of abstraction is then produced: a gigantic digital space containing billions of possible configurations, made up of image fragments. The neural network is then called a "model," since it models the process of generating an image. The space thus modeled is called latent space. It is a space awaiting activation to perform the task for which it was designed: generating images by assembling configurations. By exploring the model's latent space, one can therefore potentially generate an infinite number of combinations from the images in the training dataset.

Figure 4. Noising process for diffusion model training representation (by the author).

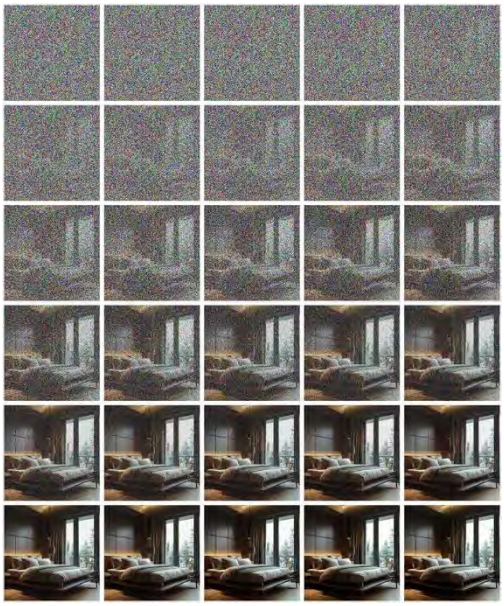

Figure 5. Denoising process durring diffusion model inference representation (by the author).

If we reverse the direction of this process—that is, if we feed a noise image into the model—the computations result in the generation of an image similar to the one used during the training phase: this is called denoising. This method of using a neural network that has been pre-trained to generate new data is called inference. Starting from random noise, the model is asked to make a kind of prediction to generate an image, by denoising the initial random image iteratively, in several successive steps. Fragment by fragment, pixel block by pixel block, the goal is to predict the most likely neighbours to arrive at the desired result, by leveraging all possible known configurations of the network to produce the desired result. It is inference that enables the exploration of a model's latent space. The final element missing here is a technique to guide the inference. This is because the latent space contained within diffusion models is immense, and the larger the training dataset, the larger the model (and thus the larger the file size, and the more computationally intensive it is to run during inference). There must therefore be a way to guide the denoising process in order to produce an image that is not generated entirely at random, without any sense of structure or visual composition. This is where Contrastive Language–Image Pretraining (CLIP) [8] technology has been a game-changer, making it possible to use

descriptive text to guide the inference of diffusion models.

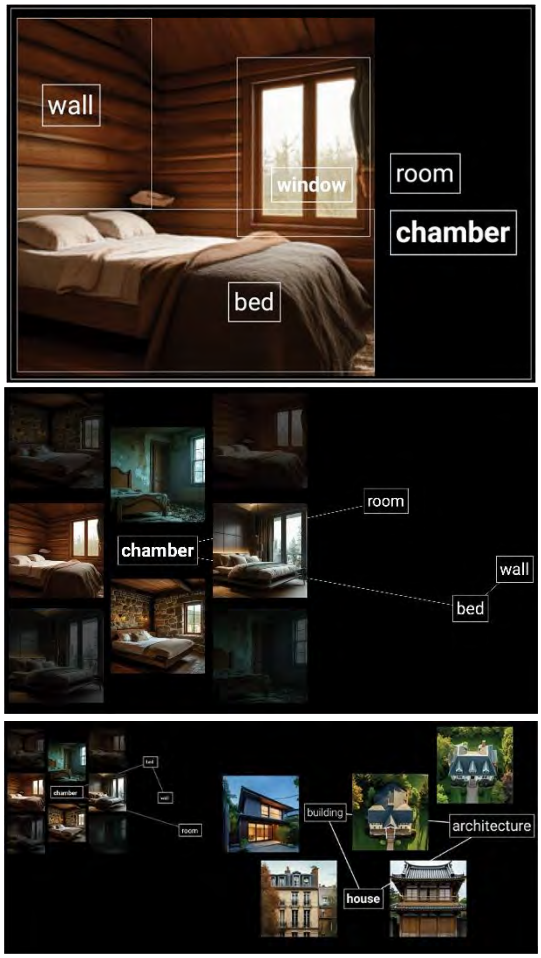

Figure 6,7,8. Visual reprensentation of CLIP technical concept (by the author).

CLIP is a model that differs from fusion models. It enables the linking of text fragments with image fragments. Since it is a model that uses two different modalities—textual and visual—it is referred to as a multimodal model. Here too, there is a training phase that has enabled the creation of relationships between billions of text/image parameters. The latent space produced by CLIP is therefore dual-purpose; it combines text encoding (text encoder) with image encoding (Vision Transformer, or ViT). During the inference phase, this allows us to use text describing the image we wish to generate as input; this is the "prompt," a term derived from the graphical interface of the classic computer terminal: a command prompt where the user types the commands, they want the computer to execute. Once the prompt is projected into CLIP's latent space to take the form of a set of vectors (an embedding), it becomes possible to guide the diffusion model's inference to attempt to force the model to produce an image containing the elements described in the prompt. But CLIP quickly reaches its limits, even when combined with other more precise and broader text encoders, such as T5 (Text-to-Text Transfer Transformer) [9], to capture the nuances of the natural language prompt more accurately.

Now that we've laid the foundations for a workflow using GenAI and its key principles, we'll outline the main workflows we used to create the images and videos for the visual environment of *Pierrot Lunaire*.

3. Results

Figure 9. A Rehearsal with the musicians team in the auditorium of Valence's Music Conservatory (photo by Pierre Bassery).

3.1. Technical Implementation

Remember that our goal here was to explore self-hosted technical AI systems. Our refusal to use online generative AI stems from several factors. For one, training datasets are not always accessible on online tools like Midjourney, so it is impossible to verify their sources or anticipate their biases. This poses real intellectual property issues, as it is impossible to know the terms under which the images used for training were acquired. Additionally, the results are very quickly limited in quantity in free plans, pushing users to subscribe to paid plans or purchase "tokens" that can be sent to inference servers. Given that a prompt never yields the ideal image on the first try—but requires dozens of attempts to stabilize a result—the "paywall" becomes a major obstacle. And finally, since the computations are performed on

remote servers, users have no assurance regarding the protection of their private data: when a prompt asks to combine one person's face with another's body—for which photos have been submitted—how and by whom will the original photos and the generated images be used? Here, responsibility is shared between the hosting company—which can appropriate user data without any transparency—and the ethics of users, who are, in principle, free to misuse the deepfakes created in this way.

For artistic exploration with no limits on time or the volume of images and videos generated, using a dedicated computer onto which the entire toolchain and models are downloaded allows you to disconnect from the internet and maintain complete control over your data and processing time. Within the framework of AI-SECRETT, exploring such a local workflow with an open-source software suite was a natural choice for us, as our philosophy is simple: we must move beyond the computer screen, not merely content ourselves with being users or consumers, but become producers and creators who not only use technologies but also contribute to their evolution. The development of a critical discourse on these technologies is only possible through genuine practice that can support informed reasoning and critical distance.

In practice, implementing image-generation technologies using diffusion models relies on computer programming, most commonly using Python scripts. Although general-purpose LLMs like Claude and agent-based AI systems now make it possible for anyone to write and run such programs, this process remains difficult for non-specialists and does not easily allow artists to experiment, as the traditional write-run-debug loop remains somewhat inflexible. The release of Stable Diffusion [10] models under open-source licenses starting in 2022 paved the way for local image generation practices, as the models can be downloaded to a computer for use in inference or fine-tuning. Developer communities then seized this opportunity and developed user interfaces for Stable Diffusion models (SD1.5, SDXL, etc.), making a form of "GenAI art" accessible to a wider audience.

The most well-known tools are AUTOMATIC1111 Stable Diffusion Web UI [11], a web interface that allows users to easily configure an image generation pipeline using open diffusion models downloaded from dedicated sites like Hugging Face [12], and ComfyUI [13], a visual programming environment based on nodes representing functions connected to one another (also known as dataflow programming [14]) that enables the creation of custom workflows based on both local and online models.

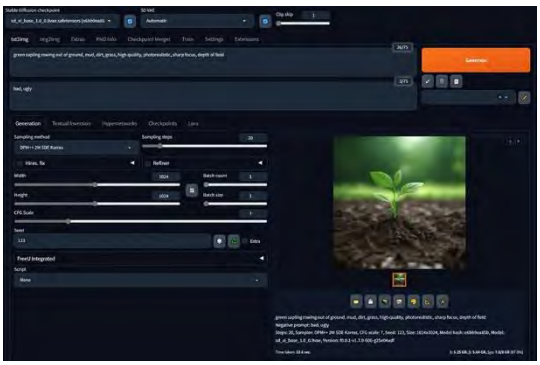

Figure 10. A1111 SD Web UI, screenshot form easywithai.com

Because it offers much greater flexibility, is more akin to an integrated development environment, allows users to create their own nodes, is actively maintained by the developer community, and has fostered a strong community of practice, ComfyUI immediately became our tool of choice.

3.2. Workflows

To create the images for *Pierrot Lunaire*, we developed a method similar to that used in animated films, based on the creation of a storyboard that follows the progression of the poems and Schönberg's musical composition. Each poem is structured in the same way: two stanzas of four lines and a final quintet, with repeated phrases creating a distinctive rhythm. Albert Giraud's Symbolist poems are particularly grotesque and bloody; they depict images of horror that capture the hardships of his era and the crises society was facing at the time. Tuberculosis features prominently in the poems, as this disease was then spreading death in a devastating manner across the peoples of Europe, particularly among the socially excluded and the poor. The figure of the poet (likely the author himself) is frequently invoked, and the precarious position of artists within the bourgeois society that surrounds and exploits them is a recurring theme in

Giraud's work. Like all forms of Symbolist art of that era, the figures used in the work refer to elements and situations from the author's lived reality. Alongside this layer of literary interpretation that unfolded in front of us, Schönberg's work—incredibly dense and profound, in total rupture with its time—had to be taken into consideration.

For each of the 21 poems in *Pierrot Lunaire*, divided into three main sections of seven poems each, we decided to create a kind of visual tableau composed of images and videos. Our division within the poems was sometimes based on the stanzas themselves, and at other times we opted for a more personal approach, based on our own interpretation of the poem, its relationship to the music, and its duration. We thus created a sort of miniature film for each poem, which also allowed us to conceive of a gradual evolution in the aesthetics of the images throughout the three parts of the musical work. The first part uses the aesthetics of early photography and the daguerreotype—images that are sometimes slightly blurry, dusty, in black and white tinted with sepia—a nostalgic and sombre aesthetic. The second part introduces some colour to evoke both the evolution of photographic technique and expressionist cinema. The final part adopts a more contemporary and colourful aesthetic, drawing on visual references ranging from surrealism to contemporary art photography, and concluding with a nod to Baroque-Rococo painting.

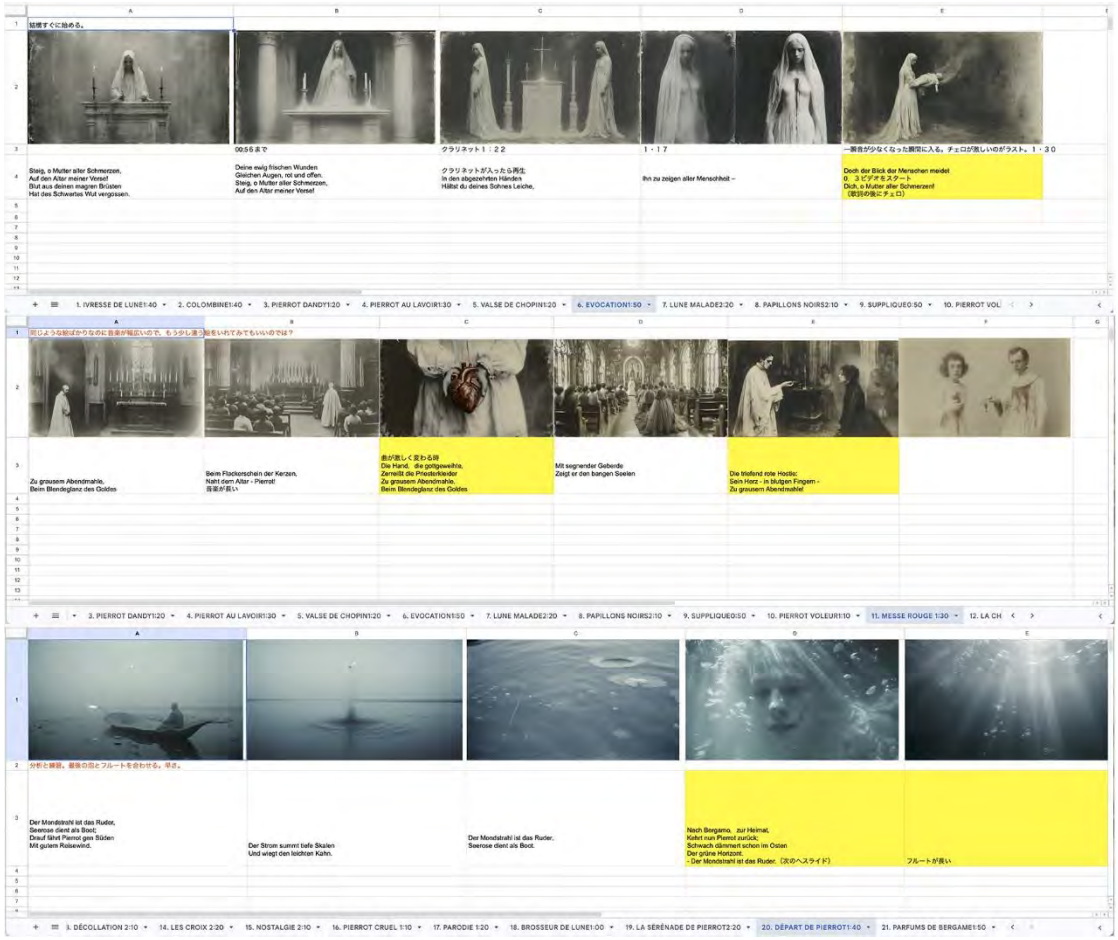

Figure 11. An excerpt of the organization sheet used during our visual poems creation (©Acronie).

And while we weren't able to generate these images in real time from the poems, we created all the images and videos in advance so we could "mix" them live during the concert, sitting among the musicians on stage and, to some extent, following the conductor's cues: we played the images, much like contemporary VJs, using crossfades or fades to black. To do this, we developed a custom software suite, and in particular a main controller similar to a mixing

console, built entirely using web technologies, to control how the images and videos are displayed and sequenced in the projection.

This also involved preparatory work based on recordings of the musical piece, which we had to rehearse to synchronize the projection of images and videos with the music, sometimes leading to adjustments (recreating an image, adding a video, changing the order of appearance to better convey our artistic intent in relation to the poem). This flexibility was essential, as the musicians from the Valence Conservatory, under the conductor's direction, did not play at the same tempo as the recordings available online or on CD.

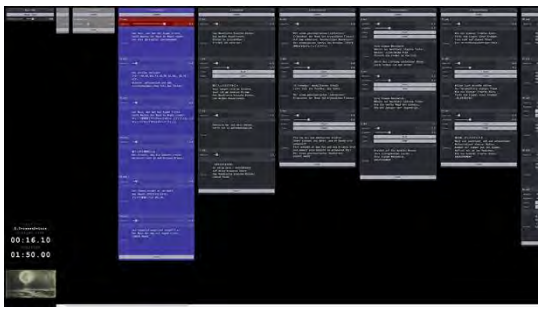

Figure 12. The main controller of our live-oriented-visual mixing tool, giving access to each poem, listing all medias and offering precise opacity and framerate controls (©Acronie).

3.2.1. txt2img

Generating images from text prompts was our primary workflow throughout the design and production of *Pierrot Lunaire*. The first phase of our work involved exploring and gaining a deep understanding of the techniques and the ComfyUI tool. Many technical concepts emerged and demystifying them was not always easy. While the principles of training and inference explained above are clearly recognizable in a ComfyUI workflow, there are parameters that influence the generation of the final image which are relatively opaque and for which it is difficult to find a clear and understandable explanation for someone who is not a specialist in AI engineering.

The fundamental component for any image generation is a sampling function, the Sampler. Its role is to denoise an input image based on a positive and negative prompt from the latent space of a given model. There are several parameters in the ComfyUI KSampler object, the values of which can have a significant impact on image generation. The "seed" field corresponds to the initial value used to fill the input latent image with noise [15]. The number of steps corresponds to the number of denoising iterations that will be performed. A small number of steps will produce an image close to noise; a large number will refine the image; too many may introduce unwanted artifacts (oversampling, overstepping).

The type of sampler determines the denoising algorithm used to move from one step to the next; it is the mathematical formula that allows us to navigate the latent space (a space that is often compressed to reduce the model's size and the amount of computation required) and extract pixels from it [16]. The relationship between these steps must also be defined, because in a latent space where each point among millions of others is represented by several parameters (512, or even 1024), the number of possibilities for moving from one denoising step to another is vast. It is the scheduler's role to define the path to take by reintroducing a small amount of noise at each step to guide the sampler, which, in turn, will remove a little more noise at each step.

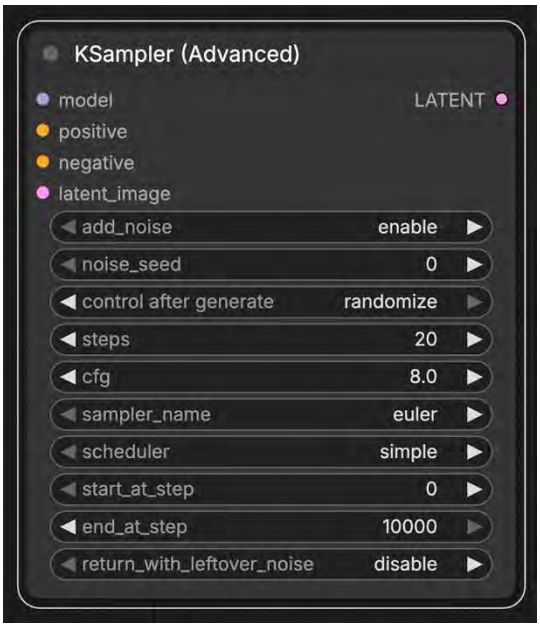

Figure 13. The main object used in ComfyUI, the KSampler.

Finally, there is the value assigned to Classifier-Free Guidance (cfg), which can be summarized as a semantic direction amplifier applied at each step between the model's prediction and each step of the sampler. In other words, it

is a value that allows for adjusting how closely each step adheres to the direction provided by the prompts. Another way to understand this value is as the margin of freedom the user allows for the random phases of the generation process (i.e., the application of noise at each step). The user prompt is transferred via a conditioning object, which encodes the prompt text into a CLIP multimodal space. The elements we want to see appear in the image must be specified in the positive prompt; conversely, elements that should not be included in the image can be specified in the negative prompt, which guides the process of extracting image elements from the latent space generated by the sampler. At the sampler's output, a compressed image in "latent" form is generated. To produce a pixel-based image, a Variational Autoencoder (VAE) must be used; its role is to convert the latent image—expressed in a compressed space—into a standard RGB pixel space, resulting in a final image.

Figure 14. A cfg effect test matrix we produced to better vizualize the influence of cfg values when generating an image from the very same prompt et seed (by the author).

For *Pierrot Lunaire*, we primarily used an open-source diffusion model called Chroma1, which we will discuss in more detail later. There are two major variants of the model: an HD version that adopts the technical principle of FLUX1 [17], which uses a compressed latent space, and Chroma1-Radiance, which offers an uncompressed pixel space. Starting from the same prompt, a diffusion model can produce images that are very different from one another. If the same "seed" is used and all parameters are identical for each generation process, the same image will be generated. It is therefore necessary to experiment with the various parameters one by one to understand their purpose and impact on the generation process. Added to this are the numerous limitations in these models' ability to interpret prompts.

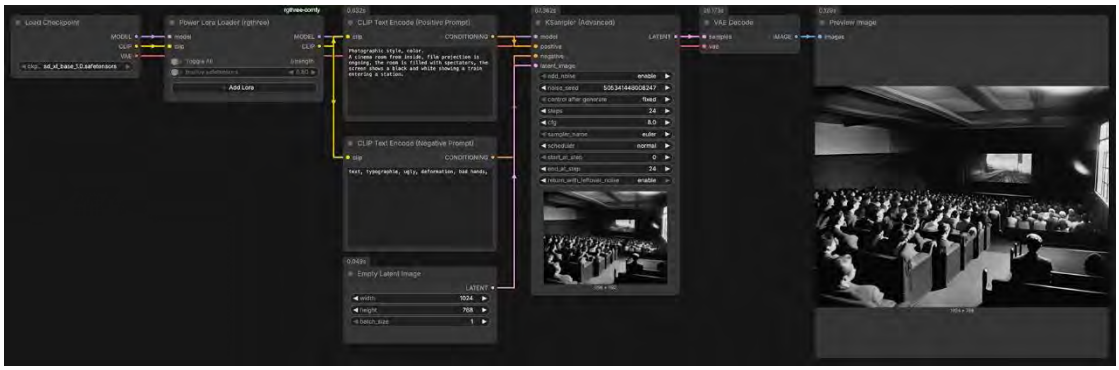

Figure 15. The simplest ComfyUI workflow to generate an image with SDXL.

Because if the interest in generative models lies in their ability to generate images that are not strict reproductions of those used to train them, but rather highly sophisticated assemblages that produce new images indistinguishable from authentic photographs, when an element described in the prompt is completely absent from the training dataset, the model will be unable to make it appear and will attempt to generate a form based on the ones it knows: these are the so-called hallucinations, mentioned earlier, which can result in monstrous images [18]. This also means that what is commonly called hallucinations in AI has nothing to do with human hallucinations: it is simply a combination of data that produces images which we, as humans, consider errors due to cultural conditioning. Technically, these are images like any others; their generation process is the same as for those we would consider "successful." Reference is often made to this work on the text that guides generation as "prompt engineering." We used LLMs (locally, mainly via Ollama [19]) to help us formulate prompts tailored to the models we used, because who else but a program could correctly formulate a command for another program?

Two other major categories of workflows were central to the creation and completion of *Pierrot Lunaire*. First is the ability to generate short videos from a single image. This technique is an evolution of the method used to generate images, adding a spatiotemporal dimension to maintain consistency between frames. Second, we used several techniques to improve the quality of the images after their initial generation with Chroma1-Radiance. Some of these techniques have long been known in computer graphics, but they have taken on a new dimension with the advent of AI models.

Figure 16. The generation steps of one image of poem 4. Laundress Moon (by the author).

3.2.2. img2video

Video generation is a technique that is fundamentally transforming the creation of animated content. The general principle behind these techniques is very similar to that of image generation, as they involve the same two main training phases using annotated examples: denoising is learned by a neural network (DiT, for Diffusion Transformer), and inference is guided by prompts provided by the user. We used Wan2.2 [20], developed by Alibaba, the Chinese e-commerce and cloud hosting giant, whose models are open to use and allow users to refine their own models [21]. Directly usable in ComfyUI locally on a computer with sufficient hardware, Wan2.2 uses two sampling phases. The first works with high noise to organize the scene's structure (space, object positions), as well as the direction and speed of primary movements. The second phase

operates on finer noise to process movement details and textures (low noise). This implies that the denoising guidance retains an element of randomness, since—as with images—it involves guiding the generation of images based on the potential the model contains as a result of its training. Prompts must therefore match the text used during training so that the inference process produces the desired movements. However, in practice, it is common to obtain disappointing results that do not match the original intended outcome at all. It sometimes takes dozens of attempts to get a video that matches the prompt, and sometimes you also have to be satisfied with what you've managed to produce and accept the fact that the model is simply not capable of generating the desired video.

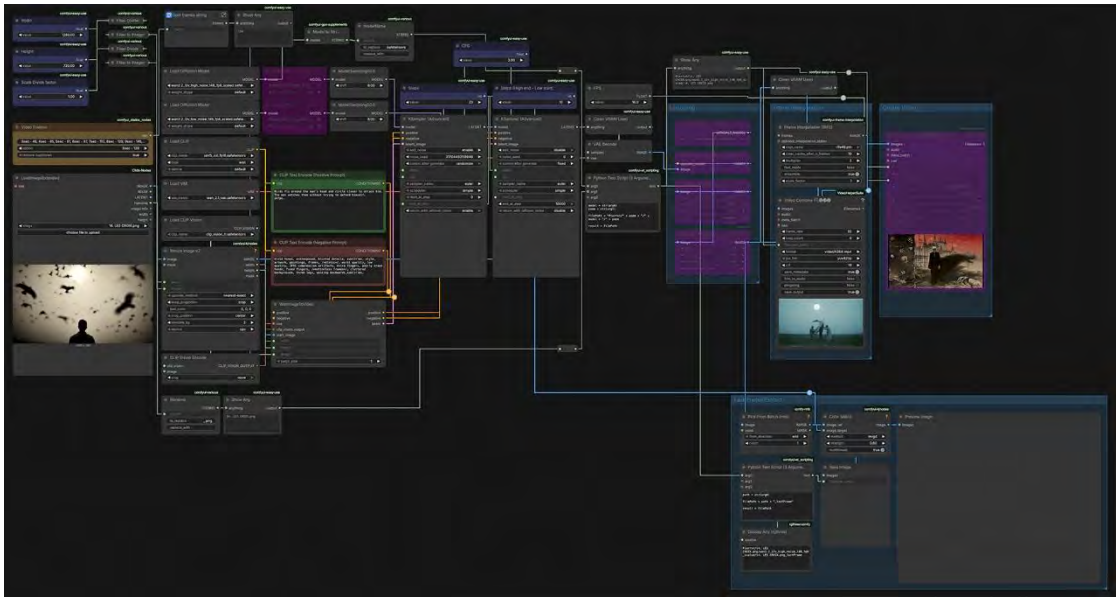

Figure 17. The ComfyUI Image-to-video workflow used for Pierrot Lunaire (by the author).

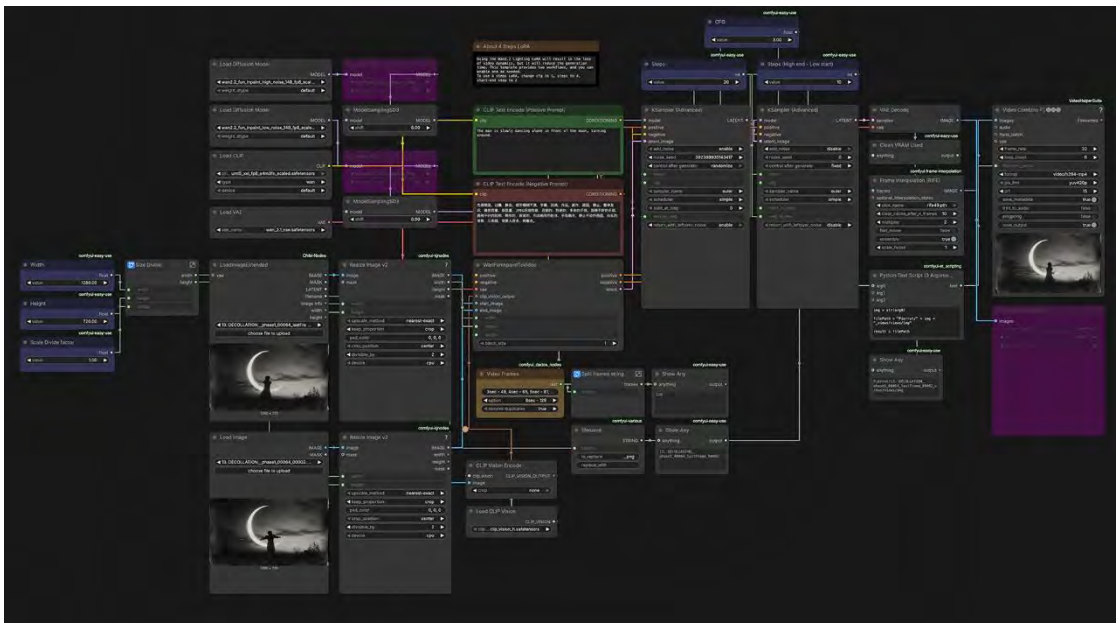

Figure 18. The start & end frame Image-to-video workflow (by the author).

Two techniques enabled by specialized Wan2.2 models allowed us to minimize errors. The first involves Wan2.2 Fun Inpaint [22], a model fine-tuned by Alibaba's internal engineering team. This model does not make predictions based solely on a still image, but rather on a start frame and an end frame, hence the name of the workflow: start & end frame. In the case of Pierrot Lunaire, because we made the decision to use no other software than ComfyUI, we had to find methods to create the images corresponding to the end of the video sequences we wanted to produce. This involved using techniques related to modifying existing images (whether generated or not), which we will discuss below. Defining the final image of the video to be generated allowed us to significantly improve the model's adherence to our prompts: denoising is guided much more effectively, since comparing the start and end images enables more efficient conditioning that better respects the user's prompt.

The second technique is particularly impressive: Wan2.2 Fun Control-Camera. This is also a refined version of the original model, but this time it is designed to manipulate the camera within a specific scene in an input image. Several camera movements are available (vertical, horizontal, forward, and backward travelling; clockwise or counterclockwise rotation), and the speed and offset point within the image to guide the movement can be configured.

For *Pierrot Lunaire*, this allowed us to create a precise scene for Poem 15: *Nostalgia*. This poem opens the third part of the musical work, in which Pierrot abandons the anger, violence, and darkness of the previous part and yearns for rest, for a return to the roots of the art from which he emerged as a theatrical character: the commedia dell'arte in Italy and its evolution during the Baroque-Rococo transition in European art in the 18th century. For the first sequence of this poem, our intention was to present a Pierrot who is more colourful than in the previous parts, yet trapped inside a crystal, as suggested by the first stanza. Among the tens of images we generated, the idea of moving the camera forward in a completely frozen scene emerged, in order to give viewers the feeling of entering the crystal imprisoning Pierrot.

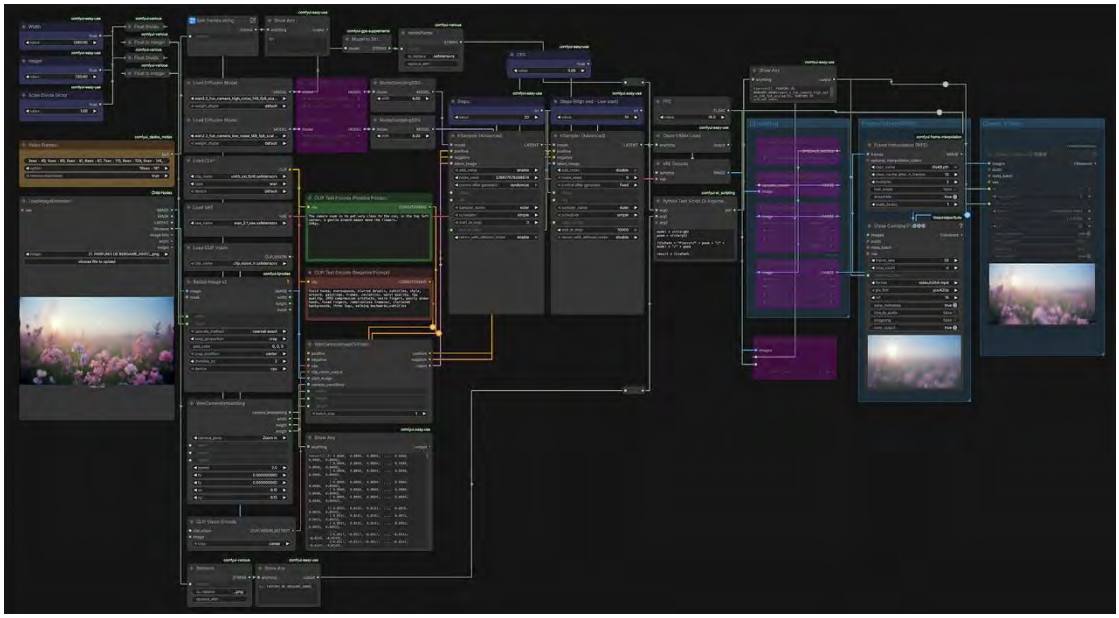

Figure 19. The start & end fram Image-to-video workflow (by the author).

Figure 20. Some of the images generated for poem 15 Nostalgia, first part.

Figure 21. Still captures of the zoom in video for poem 15 Nostalgia, first part, made with Wan2.2 Fun Camera Control.

Figure 22, 23. Still captures of a camera traveling video made with Wan2.2 Fun Camera Control, and an attempt to derivate it to 3D Gaussian Splatting scene, showing the inconstencies of the space made by the video generation.

Other tests we conducted using this same technique demonstrate the model's ability to fill in the blind spots of a landscape scene during a camera movement. Visually, the spatial coherence is impressive, even if it is merely an illusion. Indeed, when attempting a 3D reconstruction from a sequence of images captured

with Wan2.2 Fun Camera Control, we observed that the perspective is not precisely maintained from one image to the next and that several architectural elements appear or disappear during camera movements. In a photogrammetric reconstruction, the space captured by numerous images (photographic or synthetic) of the same scene is used to determine identical points in space and gradually reproduce a 3D point cloud model. One of the recent advances in this field of computer graphics is the technique of Gaussian Splatting, a set of coloured spots of varying sizes and orientations that create the illusion of photographic quality without the need to reconstruct a textured mesh.

3.2.3. Image-Space

While we were exploring these technologies, a new technique emerged, and we were able to test it immediately upon its public release: Sharp Monocular View Synthesis [23], which converts an image into a navigable 3D volumetric object in real time. Developed by Apple's R&D team, it is a neural network pre-trained on photographs and their depth map representations to project points into space from the viewpoint of the shot (i.e., the camera, when one exists). From the resulting point cloud, Gaussian splats are generated based on the colour of the image's pixels and an estimation of illumination derived from the distance of each point relative to the camera. The result is a bas-relief-like shape that remains surprisingly consistent when viewed near the original vanishing point and allows for slight variations in viewpoint in real time. In a way, this is photogrammetry from a single viewpoint, whereas it is preferable to use hundreds of images with classical reconstruction algorithms (though for efficiency and extremely high precision).

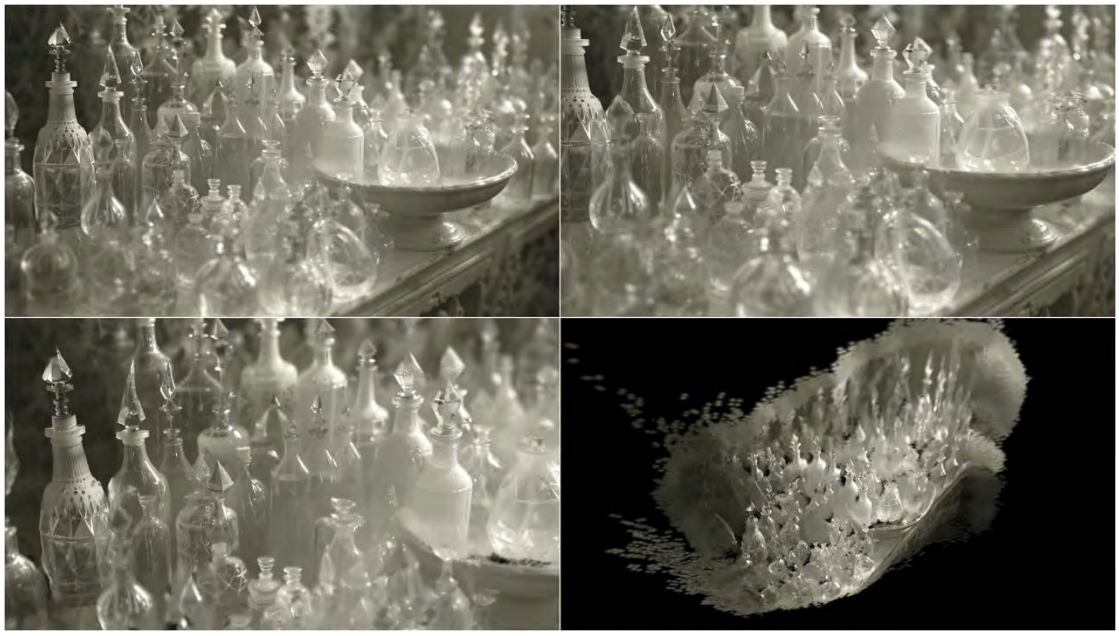

Figure 24. Still captures of the opening scene of poem 3 Dandy Pierrot, 3D Gaussian Splatting made with ML-Sharp and integrated in our realtime mixing environnment.

As with all emerging technologies, limitations are quickly reached. We found that landscapes were not rendered well in 3D space, likely because the training dataset did not include them. Architectural interiors, buildings in urban or natural environments, and forests work very well, but more abstract images produce only fairly flat surfaces that lack any visual interest. We therefore carefully selected the images we converted, and in some cases took these limitations into account when generating the images themselves.

We decided to integrate these spatial images into our *Pierrot Lunaire*, creating a kind of multimodal 3D real-time mixing that allows us to transition freely and seamlessly from a still image to a video to a 3D model using opacity fades. To control the viewpoint within these

spaces, we use a video game controller, which allows for high precision in the movements we can define on the fly. These technologies are bound to advance, with their applications primarily envisioned in augmented and extended reality, but also in robotics and autonomous vehicles: for if we can reconstruct the environment with a high degree of fidelity, we can analyse it to define the actions to be taken in the near future and endow non-human agents with this "predictive intelligence" [24].

3.2.4. Post-Processing Workflows

To conclude this overview, we would like to highlight some workflows that may seem more modest at first glance but are particularly useful in the process of creating images such as those we developed for this artwork. Having taught computer programming for digital media creation for just over 15 years, I often had to explain to younger students the fundamental principles of the digital image—its computational construction from discrete elements, encoded according to certain standards that enable interoperability between devices (such as screens and printers). One of the fundamental principles concerns the amount of information constituting the digital image, the processing possibilities this opens up, but also its limitations. One of the least intuitive limitations concern scaling: when we want to increase the dimensions of a digital image, we change its original matrix and thus increase the number of pixels that compose it. We must therefore use the available pixels to produce the same image, but larger.

While various image resizing algorithms have been around for a long time, one thing was true until recently: a program cannot create information that is not present in the original image, so there is a loss of quality (resolution) when the image is enlarged. For students learning the basics of the graphics pipeline, this knowledge is essential for effectively managing image production. This is no longer entirely true today. Image upscaling and cleaning techniques have made a considerable breakthrough. Many GAN (Generative Adversarial Networks) [25] models exist today and enable upscaling that infers new pixels using AI prediction techniques (rather than "simple" interpolations).

Since these can be used directly in ComfyUI, it is now possible to create minimal workflows for processing older images, but workflows can also be chained together: generate an image at a medium resolution, then upscaling it with a suitable model (such as UltraSharpV2 [26]), and feed it back into a new sampling phase using the same model as the one used for generation. Thus, an image is produced in several stages through successive refinements, which also helps reduce the pressure on the graphics card's RAM. At each stage, it is possible to modify parameters to get closer to the desired result.

It should be noted here that some diffusion models now directly incorporate existing image processing capabilities. Flux2 supports a wide range of uses while remaining a single, unified model. Below is an example of cleaning an image of its artifacts for Poem 14, "The Crosses," whose prompt is surprisingly simple: "Keep image 1 exactly as it is except for one thing: remove the square texture effect throughout the image. Keep all details as they are."

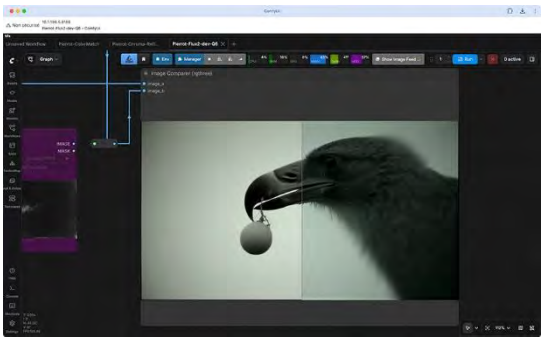

Figure 25. A simple cleaning process using FLUX2 to remove the square-patch effect on images frequently produced by Chroma1-Radiance.

3.2.5. Technical Setting Specifications

To run these workflows, we used three different hardware configurations, the first of which—listed below—was phased out as soon as we were able to order new equipment.

1) MacBook Pro M3 Pro with 18 GB of unified RAM

2) Zi Artist 9 Studio PC

- Intel Core Ultra 9 285K processor (3.7 GHz / 5.7 GHz)
- 64 GB of DDR5 RAM

- NVIDIA GeForce RTX 5090 graphics card, 32 GB

3) NVIDIA DGX-Spark

- 20-core Arm processor: 10 Cortex-X925 + 10 Cortex-A725
- 128 GB unified LPDDR5
- Blackwell GPU

The current AI technology market is largely dominated by NVIDIA. The CUDA architecture (Compute Unified Device Architecture, which encompasses both hardware and software) is the contemporary standard for tensors computations and managing memory allocation between the CPU and GPU. Without a CUDA-compatible graphics card, computation times become extremely long, as our initial experiments on a high-end Apple laptop demonstrated. With Chroma HD, using the same ComfyUI workflow, generating a 1280 x 720-pixels image takes the following time:

Table 1. 1280 x 720 px image generation duration on our 3 hardware settings.

MacBookPro M3	~1800 s (30 min)
PC Zi Artist 9 Studio	24 s
DGX-Spark	~180 s (3 min)

Although Apple has developed a hardware acceleration solution dedicated to inference (MLX and Metal Performance Shader—*mps*), this framework is not compatible with CUDA, and since the training pipeline for diffusion models uses the CUDA architecture, the inference pipeline requires it as well. Model conversion systems exist, as do platform-agnostic formats (such as ONNX, Open Neural Network Exchange), but performances remain low nonetheless on the diffusion models we tested.

With WAN2.2, the size of the models—even in their quantized versions it's more than 20 GB for both models (quantization reduces the numerical precision down to lower-bit representations to reduce the memory footprint, making some models usable on systems limited by memory bandwidth)— the MacBook Pro does not have enough memory to even run the inference, since the entire pre-trained models, text encoder and VAE must be loaded into memory for the inference to be executed. For an 8-seconds video generated from a 1280 x 720 pixel image, the following processing times are required:

Table 2. 1280 x 720 px 8 s Img2video generation duration on our settings, WAN2.2-14B base model.

MacBookPro M3	Out of memory error
PC Zi Artist 9 Studio	~30 min
DGX-Spark	60~70 min

With acceleration techniques using LoRA, these processing times can be significantly reduced at the cost of lower image quality, which is generally offset by refinement and upscaling techniques. Their main drawback is that the movements in the videos are less aligned with the prompts, which is why we did not use them for *Pierrot Lunaire*. With "Turbo" LoRAs, we achieve the following processing times:

Table 3. 1280 x 720 px 8 s Img2video generation duration on our settings, WAN2.2-14B base model with turbo LoRA.

MacBookPro M3	Out of memory error
PC Zi Artist 9 Studio	4~6 min
DGX-Spark	12~15 min

Competing models, particularly the LTX-2, achieve significantly shorter generation times, which is very impressive, but the model's adherence to the prompt and the artifacts it introduces are too pronounced. For example, with a starting image that has a 1900s aesthetic (analog, black & white with sepia shades), even when using negative elements in the prompt to prevent it, the model automatically applies an old-film aesthetic by adding unwanted artifacts like dust or smoke. Thus, even though it takes only 98 seconds to generate a video from a 1280 x 720-pixel source image on our PC Zi Artist 9, the results are so far from what we're visually aiming for and so inconsistent—even after more than 10 attempts—that we've excluded this model from our workflows.

Since the primary goal of this creative project is aesthetically-oriented, we did not implement a system for comparing the models with one another, as such a comparison could never be objective. Our assessment is purely subjective, just as the very nature of artistic creation is. Thus, we tested the following models: *Z-Image*, *Flux1*, *Flux2*, numerous derivatives of *SDXL* (*Juggernaut-XL*, *DreamShaper*), *SD3.0*, *SD3.5*, and *Qwen Image*. All these models perform similarly in terms of generation speed, with the exception of the base version of *Flux2*, which takes about 4 minutes to generate our images on the Zi Artist 9 Studio PC. It was

indeed the aesthetic qualities—which we evaluated subjectively and for this particular project—that led us to choose the *Chroma* model series, in addition to their openness and high transparency, which we discuss below in our conclusion.

The technical requirements for these technologies are evolving very fast. The situation we faced during the six months we worked on *Pierrot Lunaire* changed as the project progressed, mainly due to the emergence of new models. But advances in computing performance are also to be expected.

4. Conclusions

Figure 26. Exploring latent spaces?
Image generated with Flux2 (by the author).

The illustration above captures the feeling we experienced repeatedly while creating the images that make up *Pierrot Lunaire*. Approximately 20,000 images were generated to get the 120 or so that were actually used in the show, in the form of still images, video, or image-space. Guidance techniques, prompt engineering, and conditioning seem highly advanced when described technically and when we observe their first results. But implementing them in a large-scale project intended for public presentation and expected to demonstrate high visual and conceptual artistic quality made us fully realize their limitations.

Starting with a given poem, we break it down into scenes and, for each one, we write prompts—sometimes with the help of LLMs—and then observe the results. The initial images open up many possibilities and stimulate our imagination, especially when we don't have any preconceived images in mind. Gradually, we make choices regarding the image's composition, the elements it should contain, their placement within the image, and their precise appearance. A shift occurs here, because the more detailed our prompts become, the less the model is able to generate an image that matches our imagination—as if our instructions had become too precise and were ultimately ignored. Sometimes, after 10 or 20 iterations based on the same prompt, we get a result that is both very close to what we intended yet still somewhat surprising: it's a kind of miracle produced by chance that we try in vain to control and that is difficult (if not impossible) to reproduce. We go pixel fishing in an ocean of possibilities with very few tools to make something of it...

Ultimately, what prevails is the impression of a constant struggle against the entropy inherent in these generative models. The more these technologies improve, the more refined our ability to control them becomes; yet these tools remain very complex to use at present, with a single image sometimes requiring three or four additional processing steps after generation to achieve the desired result. In short, we are far from the magic wand that the major stakeholders in these technologies are currently promising us to justify their thirst for conquering every possible market—creativity being just one among many, perhaps a slightly newer one [27]. But we must also acknowledge the extraordinary technical ingenuity of these systems and their power to drive creativity.

Although *Pierrot Lunaire* was specifically designed to artistically explore new AI technologies, we can say with certainty that without these technologies, we would never have been able to produce such a large volume of images and video with so few technical resources: creating all the images and films using more traditional methods (filming, set and costume design, computer-generated effects, etc.) would require considerable human, technical, and financial resources. This radically changes our relationship to artistic production: it is not a matter of depriving certain sectors of their business or even replacing jobs with autonomous AI agents, but rather of exploring a new aesthetic potential that opens up new horizons when the technologies that support them are understood and used with intelligence. For in the terminology of Artificial Intelligence, within the context of image creation, we have primarily encountered the illusion of intelligence in technical tools, but a genuine legacy of intelligence embedded within the images themselves,

and a potential intelligence to be developed and activated when these images are assembled through inference.

We discovered and used many other techniques and models to create this work, but the essence is captured here: a new field of image technology has now opened up and is within everyone's reach, albeit at the cost of significant technical learning efforts. For *Pierrot Lunaire*, a practice-based research project, we deliberately exaggerated the challenge by aiming to use only open-source tools and locally deployable, self-hosted models. The two computers we acquired (a high-end PC equipped with an NVIDIA RTX-5090 graphics card and an NVIDIA DGX-Spark mini-supercomputer) allowed us to experiment without limits with the most advanced consumer technologies.

Among the models we tested and used, all are freely available under license terms generally open to research and non-commercial production. They are said to be "open weights," as the weights comprising their neural networks can be freely used for inference. However, not all of them are equally transparent, particularly when it comes to their training datasets. That is also why *Chroma1-HD* and *Chroma1-Radiance* were our top choices. The models in the *Chroma* series are developed by a single person (a one-man project) known as lodestones [28], and they provide direct access to the datasets used for training. The model datasheets tell us that the images come from *Pexels* [29], a free and royalty-free image bank, whose descriptions were automatically generated by lodestones using *CogVLM* [30]. The Chroma1-HD diffusion model uses the same architecture as *FLUX1* via a variant, *FLUX.1-schnell*, as a starting point for large-scale training. *Chroma1-HD* is therefore the closest thing to open-source technology for image generation models that we have found that meets our expectations.

Chroma1-Radiance uses the same datasets but radically changes the architecture by not using a latent space: the model uses a principle derived from 3D space reconstruction, Neural Radiance Fields (NeRF), which are neural networks dedicated to calculating the luminance of a point in space from multiple viewpoints; Gaussian Splatting, mentioned above, is an evolution of this technique. The architecture of this model remains poorly documented, with only a technical diagram available on its distribution page [31].

But here we get a glimpse of what is truly significant about these contemporary technologies that are so profoundly transforming our relationship to creation: while the expert and thoughtful use of image and video generation models opens new pathways for artistic creation, it is certainly the invention of the technologies themselves that we should link more explicitly to artistic projects, as the creation of technical tools should be considered part of the artistic act itself and, by extension, a political act. By deliberately keeping our distance from the major proprietary models (*DALL.E*, *Midjourney*, *Sora*, etc.)—which are online and opaque regarding the use of the data we share on these platforms—we also seek to maintain a certain independence and autonomy in the creation of artworks. Yet we recognize and advocate for the fact that the creation of tools and technologies must meet the same requirements of critical positioning.

Figure 27, 28. Pierrot Lunaire, LUX Valence, May 12 2026 (©Acronie).

Funding: This work is co-funded by the European Union's Digital Europe Programme under Grant Agreement No. 101226207 (AI-SECRETT — AI Supported and Enhanced Creativity for the Triple Transition). Views and opinions expressed are however those of the author only and do not necessarily reflect those of the European Union or the European Health and

Digital Executive Agency (HaDEA). Neither the European Union nor the granting authority can be held responsible for them.

Acknowledgements: This work was made by *Acronie*, artist duo of Dominique Cunin & Mayumi Okura, with the support of Esad Grenoble•Valence management team.

Didier Vadrot, director the Music and Danse Conservatory of Valence, initiated the relationship with Catherine Rossi-Batôt, director of LUX, Scène National de Valence, to program this event and conducted the live concert with the professors-musicians:

- Ondi Forte, violin and viola
- Marie-Joelle Lecorre, cello
- Jean-Frédéric Perraud, flute and piccolo
- Moneim Brini, clarinet and bass clarinet
- Glawdys Wild, piano
- Laetitia Cattier, vocals

Pierre Bassery, artist and trombone professor participated actively in the artistic direction and the general scenography.

We express here our deepest thanks to all of them for their trust in this exceptional adventure.

Conflict of Interest: The authors declare no conflict of interest. The funder had no role in the design of the study; in the collection, analysis, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References and Notes

[1] https://aisecrett.eu, [Accessed: June. 02, 2026].

[2] https://sshrc-crsh.canada.ca/fr/financement/terminologie.aspx#25. See also "Définir la recherche-création ou cartographier ses pratiques ?", Louis-Claude Paquin et Cynthia Noury, Université du Québec à Montréal https://www.acfas.ca/publications/magazine/2018/02/definir-recherche-creation-cartographier-ses-pratiques, [Accessed: June. 01, 2026].

[3] acronie.org, [Accessed: June. 01, 2026].

[4] 1912 was also the year Marcel Proust's novel In Search of Lost Time was published—a work widely regarded as marking the birth of the modern novel—the year Marcel Duchamp created his first ready-mades (including the famous Fountain), marking the beginning of conceptual art, and the year Picasso and Georges Braque produced their first Cubist works.

[5] For more on this topic, see Ioan Roxin, Generative AI: the risk of cognitive atrophy, https://www.polytechnique-insights.com/en/columns/neuroscience/generative-ai-the-risk-of-cognitive-atrophy, [Accessed: June 4, 2026].

[6] Imagine a drop of ink falling into a glass of water. The drop gradually spreads until the water is uniformly colored—a state that corresponds to pure noise in this context.

[7] See the LAION-5B dataset used to create numerous diffusion models, https://laion.ai/blog/laion-5b, and its predecessor ImageNet, https://www.image-net.org/about.php, [Accessed: June 3, 2026].

[8] Alec Radford et al., Learning Transferable Visual Models From Natural Language Supervision, 2021, https://doi.org/10.48550/arXiv.2103.00020, [Accessed: June 3, 2026].

[9] https://huggingface.co/docs/transformers/model_doc/t5, [Accessed: June 3, 2026].

[10] https://stability.ai/stable-image & https://stablediffusionweb.com, [Accessed: June 4, 2026].

[11] https://github.com/automatic1111/stable-diffusion-webui. This tool has played a major role in making local GenAI more accessible, but development has currently stalled. Alternatives do exist, such as Invoke (http://invoke.ai/), whose open-source version remains available and under active development despite Adobe's acquisition of the version directly accessible online.

[12] https://huggingface.co/.

[13] https://comfy.org/.

[14] There is a long tradition of visual programming languages, found not only in a branch of computer science focused on programming interfaces for non-specialists (on this topic, see Henry Lieberman et al., *End User Development*, Springer-Verlag, 2006), as well as in media creation software, the classic example being the music creation software MaxMSP (https://cycling74.com/products/

max) and its open-source equivalent PureData (https://puredata.info/).

[15] It is important to remember that true randomness does not exist in computer science; random functions generate long sequences of numbers calculated from a starting value (seed). The calculation is deterministic: rerunning the calculation with the same starting value will produce the same sequence of numbers. These are therefore referred to as pseudorandom numbers.

[16] The “K” in KSampler comes from K-diffusion samplers, which are number solvers for ODEs (Ordinary Differential Equations) or SDEs (Stochastic Differential Equations) used in the denoising process.

[17] FLUX1 and its variants were developed in 2024 by Black Forest Labs, a German company founded by former employees who left Stable Diffusion. Today, its successor, FLUX2, has taken over, offering multimodal versions suitable for a variety of applications.

[18] The SD3 and 3.5 models were widely criticized shortly after their release because the level of censorship applied to human bodies to avoid nudity was too strict, causing the model to generate grotesque images whenever a human body was included in the prompt. https://arstechnica.com/information-technology/2024/06/ridiculed-stable-diffusion-3-release-excels-at-ai-generated-body-horror/, [Accessed: June 5, 2026].

[19] https://ollama.com/.

[20] https://arxiv.org/abs/2503.20314 et https://huggingface.co/Wan-AI/Wan2.2-TI2V-5B.

[21] Fine-tuning refers to the process of adding custom datasets to a pre-trained base model using a technique called LoRA (Low-Rank Adaptation), which adds a new layer to an existing neural network to give it new “capabilities” (such as using a specific image style, a person's face, etc.).

[22] https://github.com/aigc-apps/VideoX-Fun.

[23] Sharp Monocular View Synthesis in Less Than a Second, https://arxiv.org/abs/2512.10685.

[24] This is the project behind NVIDIA's Cosmos model series, which aims to develop a physical AI that enables autonomous robotic agents to operate in our physical world: https://huggingface.co/nvidia/Cosmos3-Super.

[25] GANs are a deep learning technique that has been around for a while (2014) and is highly effective for many tasks involving translation, image generation, or image evaluation. The basic principle involves having two neural networks compete against each other, with one generating data and the other verifying its quality. After several iteration cycles, the network can no longer distinguish the generated data from that in the training dataset.

[26] https://openmodeldb.info/models/4x-UltraSharpV2.

[27] For more on this topic, see Laba, Natalia. “AI is not a tool.” AI & Soc 41, 4157–4159 (2026). https://doi.org/10.1007/s00146-025-02784-y.

[28] https://huggingface.co/lodestones.

[29] https://www.pexels.com.

[30] https://arxiv.org/abs/2311.03079.

[31] https://huggingface.co/lodestones/Chroma1-Radiance.