Semiotic Review 9: Images | article published April 2025 | DOI: https://doi.org/10.71743/ee5nrx33

For a Semiotic Approach to Generative Image AI: On Compositional Criteria

Big Dada

(Enzo D'Armenio
enzo.d-armenio@univ-lorraine.fr

Maria Giulia Dondero
mariagiulia.dondero@uliege.be

Adrien Deliège
adrien.deliege@uliege.be

Alessandro Sarti
alessandro.sarti@ehess.fr)

Abstract: This article analyzes the semiotic functioning of Midjourney and DALL•E, two generative AI models capable of producing images out of natural language prompts. The theoretical assumption of this article is that the images produced by these AIs are the results of a particular intersemiotic translation, realized through the collaboration of human and computational operators. Our research will show the specificity of the intersemiotic translation realized by AIs *vis-à-vis* more classical kinds of translation (e.g., from a novel to a movie) and will also analyze the different kinds of "visual reasoning" characterizing Midjourney and DALL•E. Our article has two goals: first, to study how these models perform intersemiotic translations; namely, what choices they make in order to translate the generality of the symbolic (and indexical) signs of verbal languages into the specificity of the visual composition. Second, we intend to verify the degree of control that one can have over the visual composition. Following this, we present the results of the tests carried out on Midjourney and DALL•E pertaining to two semiotic macro-criteria: plastic categories (eidetic, chromatic, and topological) and visual enunciation (gaze relations, visual translation of verbalized actions, temporality, and aspectuality). These criteria were developed by Paris School semiotics in order to analyze artistic images.

Here, they will be used as principles of composition and parameters for controlling the results. At the end, we demonstrate that through this experimentation with elementary parameters of visual composition, semiotics can provide an epistemological and analytical framework for understanding and assessing the intersemiotic translations realized through generative AIs. Reciprocally, these tests on AIs aid our understanding of the two semiotic macro-criteria used, notably leading to a multiplication of enunciative instances in image production. Databases, algorithms, prompts, and aleatoric elements act as discursive agents.

Keywords: semiotics, generative artificial intelligence, intersemiotic translation, visual studies, composition, enunciation

---

0. Introduction

This article focuses on generative artificial intelligences (AIs), in particular those involving the text-image transduction, such as Midjourney and DALL•E.[1] Our research examines the generation of images from natural language prompts addressing the text-to-image translation process. Specifically, we explore the following, crucial question: How can a generative AI construct a two-dimensional *compositionally coherent* image based on a verbal prompt? A second, closely related question, concerns what we could call "machinic perception": What kind of optics characterize AI-generated imagery? Does this way of seeing (which, note, is also a way of "hearing," or reading, verbal prompts)[2] depend on the dataset of images used in the training stage, and therefore on the multiple "ways of seeing" that are enunciated in images and distributed across the dataset? Each image is the staging of a perspective on the world (and values), as represented by the database. In this sense, the database contains thousands of points of view on the world.

Clearly, this "perception" is an aggregation of all the "stimuli" received from the different data present in the dataset, where these stimuli have been translated into numerical vectors known as "embeddings." To investigate this machinic way of "hearing" and "seeing," this article uses the so-called "plastic" categories that have been formulated in what is known as Paris School semiotics. These categories are relevant to explain the compositions produced by such AI, as well as the "compressed" and "averaged" way of seeing that is enunciated in every image generated from such datasets and by their prompts. In doing so, this article also tests what in Paris School semiotics are called "enunciative categories," that is, the indexical dimension of space, time and

intersubjectivity; doing so allows us to study the relational dynamics between the generated image and machinic discourse.

As to the first macro-concept, *plastic categories* (Greimas 1989; Floch 1985), we assess the machine's compositional capabilities through prompts involving: object and color positioning (topological category); relationships between colors (chromatic category); and the interplay of form outlines (eidetic category).

As to the second macro-concept, we will focus on the *uttered enunciation* (Greimas and Courtés 1982). This concerns the way the image includes or excludes the viewer or even addresses him/her; how the image welcomes or rejects the viewer within its surface; and how time is represented in the image. Regarding temporality, an action can be portrayed as belonging to the past, to the future, or be represented as unfolding in the present (an ongoing action). In addition to temporal orientation, verbal language can also challenge the machine's ability to visually represent the rhythms of an action: an action that lasts, begins, or ends (i.e., aspectuality).

We use these two macro-categories to examine the intersemiotic translation between verbal prompts and image generation: plastic components and enunciative configurations, thus, are employed here not as analysis parameters, but as compositional principles.

The following tests with Midjourney and DALL•E have two goals. First, we aim to verify the degree of control that one can have over the visual composition by activating prompts concerning abstract representations, i.e., the combination of eidetic (i.e. object outlines and distinctions between different rendering zones), chromatic, and topological characteristics, and visual enunciation (the circulation of gazes, the temporality and meta-pictorial devices regarding space). Indirectly, we also aim to understand the impact of the initiative of the human user of the AI on the results obtained; the impact of the aleatoric; and the impact of the rules and characteristics of the datasets. This means that we will also pinpoint when the images generated are *not* relevant (i.e., begin to lack fidelity) to the verbal prompts and in which ways.

In the field of computer science, these models are constantly evaluated on the basis of statistical criteria. Yet in order to assess whether images are fitting the prompts submitted by users, these evaluations contain an implicit theory of visual composition. Our belief is that semiotics can propose explicit criteria, related to visual composition, in order to assess the effectiveness of visual generative AI.

Our second goal is to understand how these models perform intersemiotic translations: namely, what choices they make in order to translate the generality of the symbolic signs

of verbal languages into the specificity of the visual composition.

To achieve these two goals, we produce a comparative analysis between the generation of images produced by Midjourney and the version of DALL•E available through ChatGPT-4. In the following, we illustrate the functioning of these models and detail the experiments carried out with Midjourney and DALL•E through prompts referring to the criteria outlined above.

The main findings of our tests concern the different enunciation modes characterizing Midjourney and DALL•E. As we'll see, the first of these models produces images trying to simulate stereotypical artistic styles pertaining to visual materialities (i.e., oil painting), of inscription gestures (the irregularities due to the movements of an artist's hand) and surfaces (walls, canvas). In contrast, DALL•E adopts a more neutral, almost didactic style, but allows one for more precise control of the visual composition. A second result concerns the difficulty these models have in producing tasks that we might naively assume to be typical of a computational logic: counting objects, arranging them in space, respecting directionality. Overall, these tests not only allow us to explore the different ways in which AIs produce an intersemiotic translation from verbal to visual language but also to test the categories of semiotic analysis themselves. What happens to these categories when they are used as principles of composition and their realization relies on the collaboration of different enunciation entities, such as human agency, the computational logic of algorithms, and the systematic insertion of aleatoric elements?

In this sense, our article answers a question that has not yet been deeply studied in recent literature on generative AI, namely, a formal semiotic analysis of the intersemiotic translation that constitutes these generative AIs.[3] We are aware that in doing so we are making an *epoché* of the political and ethical issues that have deeply concerned recent writing in the humanities and social sciences on AI technologies. We see these as complementary projects, though we underline how the latter requires understanding of the former. In this article, we restrict our ambition to the formal analysis of the results of these translations.

1. Premises on Intersemiotic Translation

Our starting point is to consider the production of images by generative AI models as a particular kind of intersemiotic translation (Jakobson 1959; Eco 2008): the translation of verbal language (the natural language prompts introduced by a human user) into visual configurations (the images produced by the generative AI model). Generally speaking, this translation operation is complex because, as pointed out by Jean-François Bordron (2011:166), it involves two different ways of enunciating and meaning-making. Verbal language is based on *predication* and relatively specific grammatical and syntactical

structures. In contrast, images compose visual features in a non-predicative manner, exploiting a logic related to tabularity and mereology (the relationship between parts and between parts and the whole). The study of intersemiotic translation between these two modes of semiosis already bears an important tradition in continental semiotics (Dusi and Nergaard 2000; Dusi 2015; Basso 2000) and in linguistic anthropology (Silverstein 2003; Gal 2015).

Starting from these traditions, we intend to emphasize two aspects pertaining the production of images through AI. Firstly, the fact that verbal language operates starting from what Peirce called symbolic thirdness: a verbal utterance such as “a yellow color” is located at the level of the generality of linguistic categories. It is a generic yellow, which can certainly be verbally modulated as light, saturated, bright, opaque, darker than another yellow, et cetera. On the contrary, the yellow of an image embodies some particular qualia and thus exploits the functioning of the primary iconism of perception (Eco 2000:376–80). In short, it is a matter of the difference between a sequence of symbolic signs and visual hypo-icons.[4] Intersemiotic translations performed using generative AI must thus establish a correspondence between linguistic categories and visual elements. This process involves selecting specific marks from the vast array of marks engaged during the model’s training phase. For example, the prompt “a yellow page” necessitates choices in visual rendering, including the precise shade of yellow, illumination, and the represented materiality of the color (e.g., oil paint, photography). Additionally, multiple decisions must be made to translate the concept of a “page” into the image’s spatial organization.

Secondly, translation becomes more complex when verbal utterances not only impose a choice between multiple visual possibilities but also entail a relationship between the predicate entities. A prompt such as “a seated human” will be translated through the choice of the particular features of the represented human being—not the verbally predicated generic human, but a visually composed specific human, determined ultimately through a genre—and the multiple ways of sitting, including the configurations of the body, legs, the back, and shoulders. The visual translation of this prompt also implies the choice of a “style,” a specification of the materiality involved, and of a specific point of view. In short, it is a translation that must produce countless hypo-iconic details that are not contained in the verbal prompt.

It is for this reason that, faced with this great variety of parameters concerning intersemiotic translation, we have chosen to start from the two macro-criteria of analysis used by visual semiotics noted above: plastic categories and visual enunciation.

The notion of plastic categories was developed in Paris School semiotics in order to explain the functioning of artistic images and poetic discourse beyond the themes and the figures (denotata) of these kinds of texts. Plastic categories are relevant to study the visual composition, the relations between the various zones in an image, the gradual differences between them, and the global organization beyond already codified and recognized referents. Visual enunciation, for its part, allows us to analyze how visual configurations build different relationships with the viewer.[5]

Starting from intersemiotic translation implies examining the possible control over verbo–visual correspondences, the compositional and stylistic choices characterizing the functioning of AIs, as well as the inconsistencies. This is a field of meaning construction that is new to semiotics because of two factors. Firstly, the "technologies" themselves, the generative AIs, are new: in the current scientific landscape, a way of transforming these models into objects of scientific investigation has not yet been stabilized. We believe that visual semiotics can help identify relevant parameters for this purpose. Second, generative AI is beginning to colonize human practices, especially in the fields of design, translation, editing, and coding. These social transformations challenge semio-linguistic disciplines and invite a reconfiguration of their conceptual and methodological apparatus.

Before starting our tests, we wish to introduce generative models in a more technical way.

2. Text-to-Image Generative Models: Some Landmarks

As already mentioned, text-to-image generative models are AI programs that create images based on written descriptions, or "prompts," provided by the user. By processing an input prompt and soliciting the compositional rules and patterns learned from their usually extensive training data, these models can render detailed images that are supposed to align with (that is, are appropriate or relevant to) that prompt.

The process typically begins by converting the prompt into a high-dimensional vector representation, also called "embedding," which captures the semantic meaning of the text within a long list of numbers that are more easily manipulable by the machine. This step can be performed by, for example, a language model—usually based on the transformer architecture (Vaswani 2017)—that has been trained to process vast amounts of text, or by a multimodal model such as CLIP (Radford 2021) that has been trained to project images and corresponding textual descriptions into a shared embedding space. The actual generative component—which, since 2022, has generally been a type of diffusion-based model (Rombach 2022); before 2022, the generative component was typically an autoregressive model and earlier, before 2020, a Generative Adversarial Network, or GAN (see Figure 1)—then takes this embedding and begins producing an image. A simplified overview of the training and generative processes is provided in Figure 2. In the case of

diffusion models, the system starts with random noise and iteratively refines it, guided by the text embedding, until it forms an image after a predefined number of diffusion steps. The model can conduct this process because, during training, it has performed the same kind of denoising process on large datasets of noised text–image pairs, thus learning how linguistic concepts map to specific visual elements (see Figure 3).

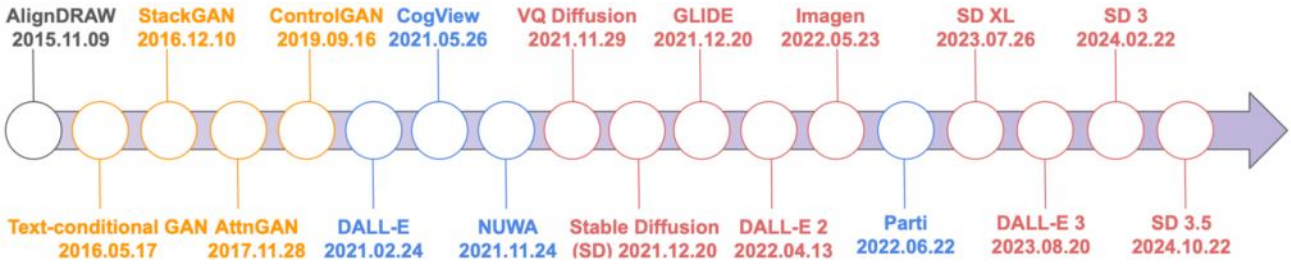


Figure 1. Timeline of some salient text-to-image models (from Zhang 2024). GAN-like models are in orange, autoregressive models in blue, and diffusion-based models in red, clearly indicating the current dominating paradigm.

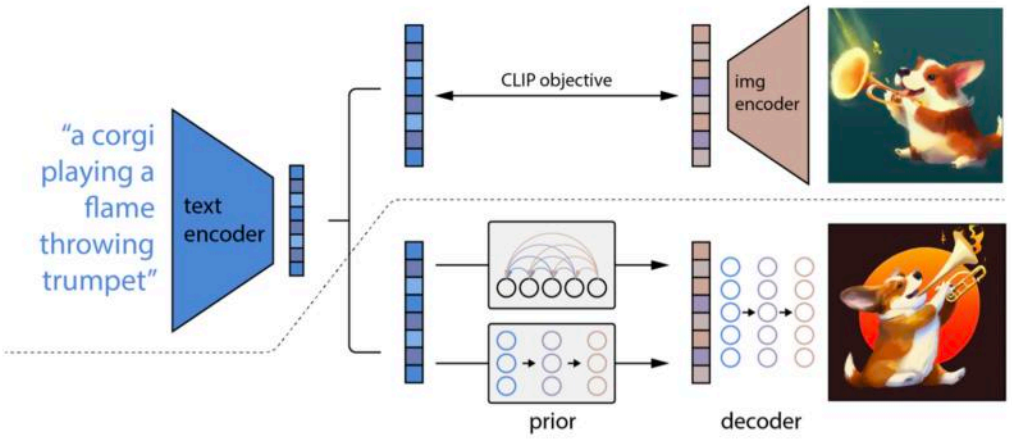


Figure 2. High-level overview of the training (over the dashed line) and generative (under the dashed line) processes underlying DALL·E 2 (from Ramesh 2022). During training, corresponding text-image pairs are encoded in a shared embedding space in a contrastive way, i.e., such that their representations are close to each other, and far from unrelated images and texts. During inference, the prompt is encoded, then used in a generative model that decodes it and generates an image, usually following a diffusion process (see Figure 3).

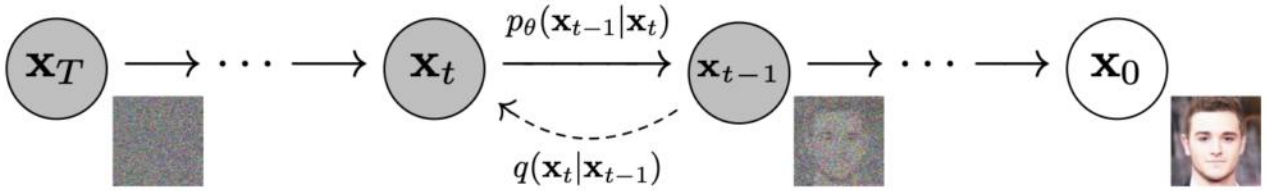


Figure 3. Illustration of a diffusion-based image generation process (from Ho 2020). During training, images are noised iteratively (from right to left in the figure) and the objective of the model is to reconstruct the original image from the noisy image (left to right). This allows, during inference, to start from a noisy representation and denoise it progressively to generate an image, guided by the prompt embedding (see Figure 2).

While these models can be incredibly powerful, they also have limitations: they may generate what—from the perspective of developers and users—are seen as inaccuracies or nonsensical details; they may inherit biases that are present in their training data; and they require considerable computational resources.[6]

The research underlying the technology of text-to-image models has rapidly evolved beyond models developed in academia to become consumer products developed and sold by private companies. While this field is moving fast, a few text-to-image programs seem to stand out and have received wide adoption by the public: DALL•E (currently version 3, from OpenAI), Midjourney (v6.1, Midjourney Inc.), Stable Diffusion (v3.5, Stability AI), Imagen (v3, Google), Runway (v3, Runway AI, Inc.), among others.[7] These programs are accessible as web-based or desktop-based software, or through general-purpose conversational chatbots such as the well-known ChatGPT (also from OpenAI), which can be queried to generate images (through the latest DALL•E version). Midjourney currently operates through the Discord platform by requiring the user to give specific commands through the prompt line, thus providing many more functionalities but making it more cumbersome to use for the lay public. A web-based interface has been made available recently for frequent users. Many of these text-to-image programs are also accessible to developers through an API (Application Programming Interface), which allows users to automate the development of applications around the programs (when authorized) and the algorithmic generation of many images through specific code snippets. Many of these platforms have free trials and their affordable pricing has made them increasingly widely used.

Importantly, programs like Midjourney and DALL•E, as used in this article, rely on proprietary text-to-image models, but are not “models” per se, as an entire infrastructure is built around the models to provide additional features or safety guardrails. A major example is the case of DALL•E 3, whose mother company OpenAI publicly states that

user prompts are first analyzed and revised by a large language model (currently such as their GPT-4 model that powers ChatGPT). The analysis aims at preventing the generation of, for example, adult content, overly violent scenes, discriminative representations, and so on. Also, OpenAI claims that longer, more detailed prompts yield better images, which motivates the automatic revision by ChatGPT of the user input into "better" prompts. This has major consequences that one must be aware of when trying to study such models: (i) the user prompt is not the one that is directly provided to the text-to-image model; (ii) the user has little to no control on the revision process of the prompt; (iii) studying the text-to-image model itself is not possible; only the whole system, including the analysis and revision of the prompt, can be studied, which might limit the scope of any study, in particular on sensitive subjects.[8]

While the prompt filtering might seem frustrating to some users, DALL•E's revised prompt can be obtained along the generated image, allowing at least a certain degree of transparency. On the other hand, its competitor Midjourney does not publish any scientific results, nor has it explained whether the prompt is revised or not, making the generation process even more opaque. To the best of our knowledge, this is also the case for all the other state-of-the-art programs mentioned above, except Stable Diffusion, whose initial process was detailed (Rombach 2022) and is speculated to outline roughly how every other model works.[9] Nonetheless, it is important to keep in mind that different models might behave differently, and that our conclusions are valid for the whole user-friendly "system" encompassing the text-to-image models themselves.

3. Testing Plastic Categories as Composition Parameters

In order to analyze in depth the intersemiotic translation between verbal prompts and visual compositions—and in order to identify the choices and degree of variability that the image, due to its specificity, is bound to make compared to the verbal description[10]—in this section we review a series of tests that we conducted using the plastic categories developed in Paris School semiotics to analyze visual texts. These tests allow us to understand the type of interactivity between the prompt and the machine's functioning.[11]

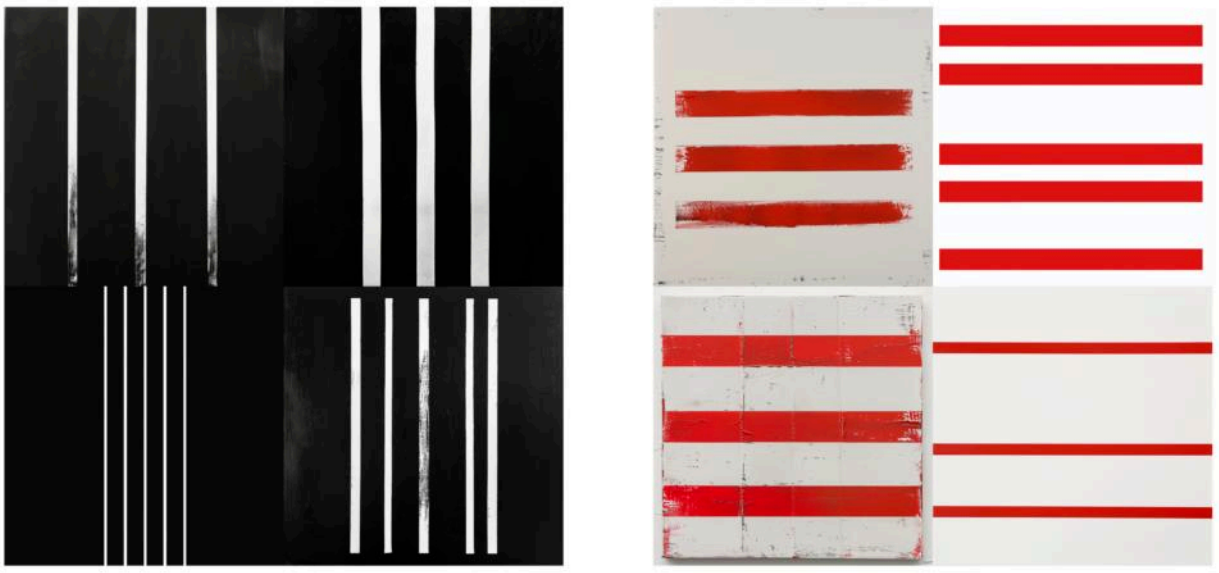
In Paris School semiotics, plastic categories correspond to the compositional characteristics that, in an image, are not related to the recognition of an object in the world but rather to configurations specific to visual language, namely everything concerning the organization of a composition unfolding within a frame that contains it and separates it from the environment.

The most fundamental plastic category, from which semioticians start to analyze any image, is the topological one because it allows the analyst to study (and "control") the relationship between the frame and the center of the image; in particular, it allows the

analyst to construct axes of surface segmentation (left vs. right, top vs. bottom, center vs. periphery) as well as the perceptual orientations of these segmentations (towards the left vs. towards the right, up vs. down, towards the center vs. towards periphery), and, finally, any "forces" that are determining the dynamics of the images (centrifugal forces vs. centripetal forces). The chromatic component, on the other hand, concerns the differences and similarities in the saturation and brightness of the colors present in the images. The eidetic category, for its part, focuses on the contours of shapes, which can be more or less linear or curvilinear, and more or less fragmented.

3.1. Eidetic Dimension

Let us begin with the eidetic dimension and test the composition of geometric shapes. We asked Midjourney to generate an image consisting of three white vertical lines on a black background, and then an image consisting of three red horizontal lines on a white background (Figures 4–5).

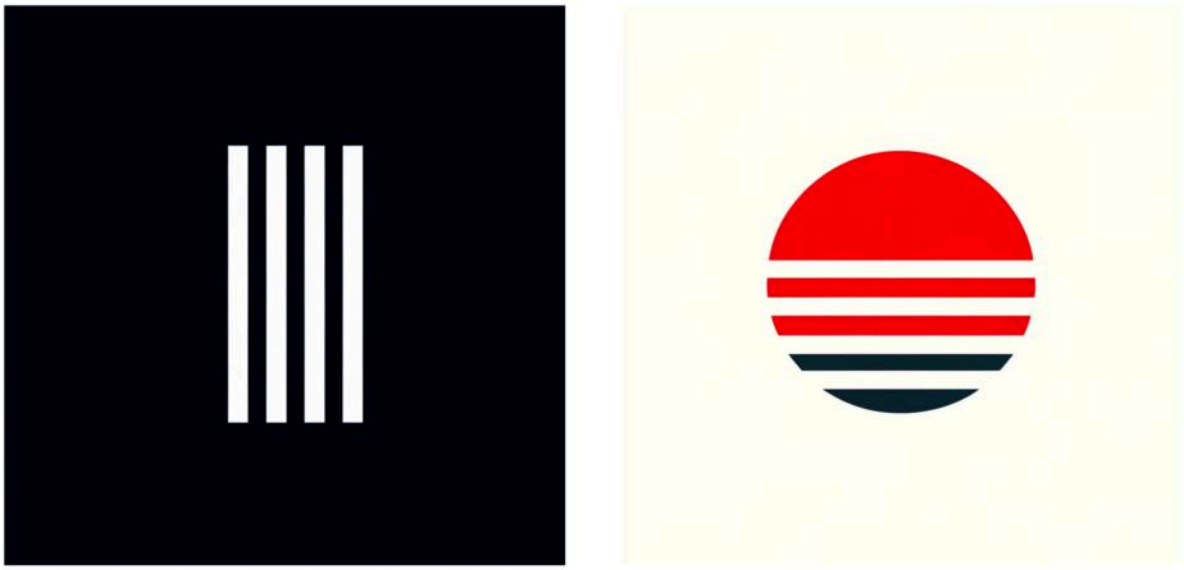


Figures 4-5. Midjourney 6. Four outputs from the prompts: "three vertical white lines on a black background" (left) and "three horizontal red lines on a white background" (right), April 2024.

In the first generation (Figure 4), only two of the four images obtained contain three lines. We notice that in the first and fourth images, the white lines show chromatic irregularities, a texture effect that goes beyond the scope of mechanical graphic composition to refer to a gesture of inscription. The same applies to the second generation (Figure 5): three of the four images comply with the instructions of the prompt, displaying the required eidetic elements. However, we find the same textural display in the first and third image, the latter also allowing us to see the borders of the visual object, that is, the contour typical of an artistic painting. It can also be noted that except for the second image generated in Figure

5, the particular tone of white can be traced back to the simulation of photographic color distortion: This is the typical “off-white” resulting from the photographic capture of a material object.

We tested the same prompts in ChatGPT-4, which uses the DALL•E generator model. The results show a comparable difficulty in composing these plastic configurations, and they display a neutral, graphical texture rather than one meant to appear “handmade.”



Figures 6-7. DALL·E. Output from the prompts: “three vertical white lines on a black background” (left) and “three horizontal red lines on a white background” (right), April 2024

Note that image generation, for a given prompt, is not deterministic but stochastic. This implies that there is a *distribution* of generable images for a single prompt. In this sense, the generative action pursued here (i.e., prompting the model to produce images) is more significant as an act of exploration of a part of the database than as a production of final results.[12] Of course, accessing the entire distribution is impossible with finite material and temporal resources. So, the images we generated are just samples, which we assume to be representative of the much wider set of possible outputs from the prompts we submitted.

In that vein, we validated our observations through the generation of 50 images with the same prompt, as for Figures 4 and 6, as displayed in Figures 8 and 9. For Midjourney (Figure 8), we observe 5 images (images 35, 38, 41, 43, 45) that perfectly align with the prompt with no or barely noticeable aesthetic effects, and 28 others (such as images 2, 8,

31, 42) that display the correct number of lines but also add such unsolicited effects. Among the remaining 17 images, 13 contain 4 or 5 white lines, and one image contains 2 lines, 6 lines, 7 lines, or 12 lines, showing the relative consistency of Midjourney in terms of lines produced. We also count aesthetic artifacts on 38 images; this high number is not surprising from Midjourney, which typically aims at generating visually pleasant results, regardless of the prompt.

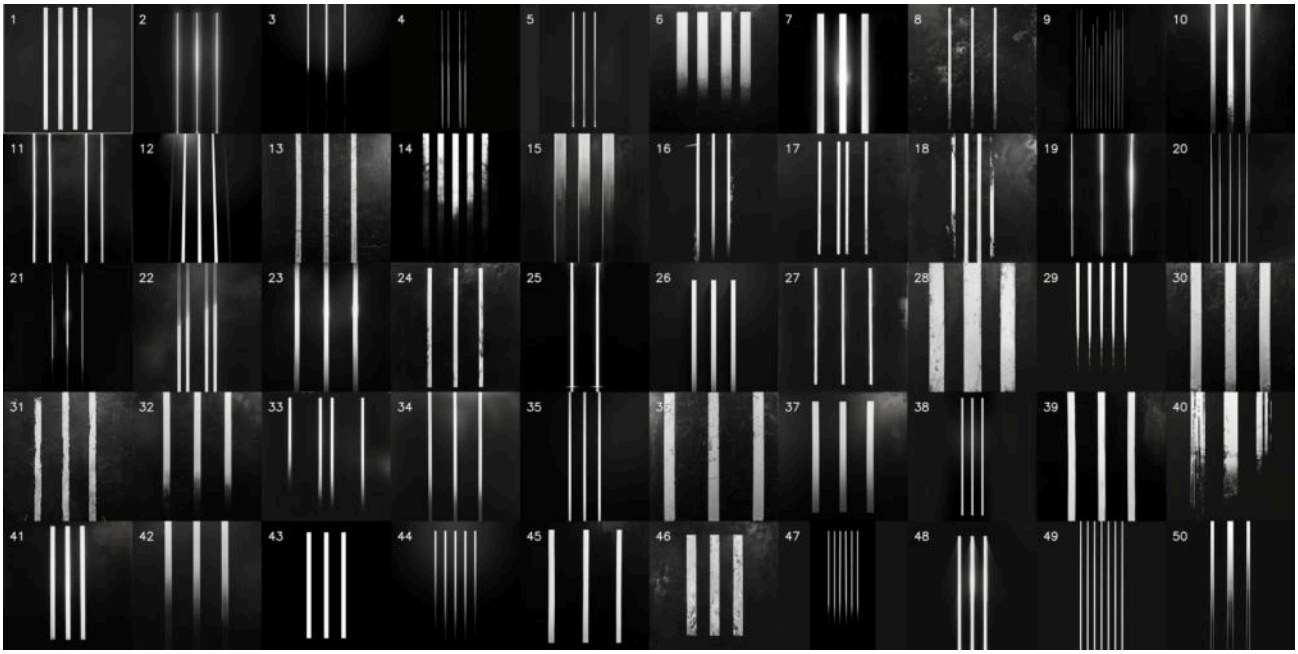


Figure 8. Midjourney 6. Prompt: "three vertical white lines on a black background," February 2025, repeated 50 times, for validation of our smaller-scale observations.
(for enlarged image: https://semioticreview.com/attachments/BigDada/)

In the case of DALL•E (Figure 9), we observe 4 images perfectly aligned with the prompt (17, 35, 42, 43), with 4 others with aesthetic effects added (image 12 has light beams instead of lines, images 18 and 40 have a black background on a grey canvas, image 38 has a greyish background). Contrary to Midjourney which represented only vertical white lines, DALL•E produced 19 images containing diagonals (e.g., image 23), horizontal lines (e.g., image 1), or curved lines (e.g., image 19). The number of vertical lines also varies much more among the 23 images containing only vertical lines (but not 3 of them), as 20 images have between 4 and 10 vertical lines, with some having up to 30 (image 36). Unsolicited aesthetic artefacts are present in 23 images, confirming the trend to a generally more neutral texture (in comparison with Midjourney). We can also note that, when more than 3 vertical lines are present, the concept of having "3 entities" is present in 12 images, for example, under the form of showing 3 rectangles (e.g., images 11, 28, 29, 37), or structures (e.g., images 8, 20, 23).

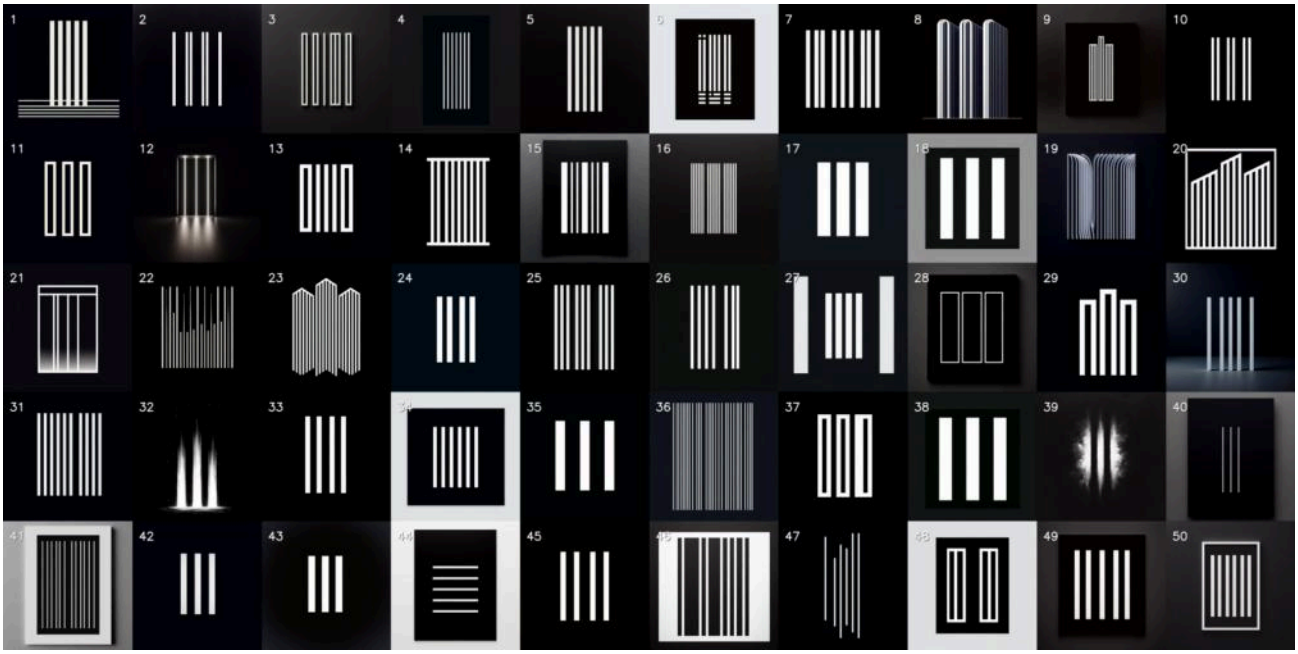


Figure 9. DALL·E. Prompt: “three vertical white lines on a black background,” February 2025, repeated 50 times, for validation of our smaller-scale observations.
(for enlarged image: https://semioticreview.com/attachments/BigDada/)

3.2. Chromatic Dimension

Let us now turn to the chromatic dimension through prompts that focus on minute color differentiations.

Figures 10-12. Midjourney 6. Four outputs from the prompts: “a bright green spot next to an opaque red spot” (left); “a saturated yellow next to a desaturated yellow” (center); “a saturated yellow colour next to a desaturated red colour” (right), April 2024.

Surprisingly, the second generation (Figure 11) produces images showing the face of a woman (the images in the top-left and bottom-right), even though such a figure was not

mentioned in the prompt. In all the images generated, we find the same treatment of pictorial painting that is typical of Midjourney, both in terms of the thickness of the chromatic material and of the inscription surfaces. The images generated do not just show colored areas, but these are made to appear textured in a recognizable manner: in the first generation, it is oil painting, whereas in the last, the texture of the inscription surface is irregular, as if it were a wall. In general, our compositional control remains imperfect, as the qualities regarding color saturation are partly independent of the indications given in the prompts.

The test we conducted using the same prompts on DALL•E shows an opposite configuration style: except for Figure 13, these are simple, neutral images, almost exemplifiers of the colors requested.

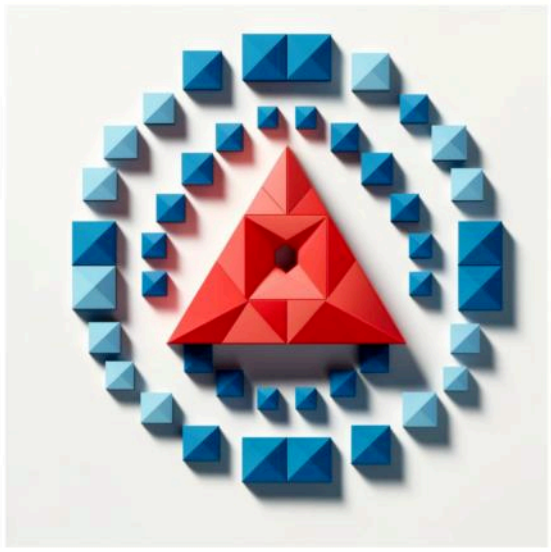
Figures 13-15. DALL·E. Outputs from the prompt: “a bright green spot next to an opaque red spot” (left), “a saturated yellow next to a desaturated yellow” (center) “a saturated yellow colour next to a desaturated red colour” (right), April 2024.

3.3. Topological Dimension

Moving on to the topological dimension, we have to clarify that this category never presents itself in an autonomous way, as it concerns the arrangement of shapes, figures and colors, and their relationship with the space of representation as a whole. We submitted two prompts: the first (Figure 16) requesting an image consisting of a circle in the top right part of the image, and the second (Figure 17) requesting a triangle surrounded by squares arranged in a regular way.

Figures 16–17. Midjourney 6. Four outputs from the prompts: “a small circle at the top right of the image on a neutral background” (left), “a triangle at the center of the image, surrounded by squares arranged in a regular manner” (right), April 2024.

In the first generation, the circle position is consistent with the prompt in two of the four images, but due to the chromatic and textural treatment of the background, the images do not show geometric shapes against an empty background as we had expected. In the second generation, three images follow the composition described in the prompts, displaying the same “choice” of composition among other possible ones: for example, the image could have staged a series of small squares that are not concentrically nested within one another, but that are instead arranged around the triangle. This is exactly the result that is obtained by using the same prompt on DALL•E.[13]



Figures 18-19. DALL·E. Output from the prompts: “a small circle at the top right of the image on a neutral background” (left), “a triangle in the center of the image, surrounded by squares arranged in a regular manner” (right), April 2024.

3.4. “Aesthetic Pressure” in Generation: Material and Figurative Tensions

A number of considerations are necessary in the light of these experiments on these plastic categories. The most obvious concerns a generalized “aesthetic pressure” in the images produced by Midjourney: the search for an artistic-mimetic composition, for a particular balance, and for an effect of painting typical beauty. This pressure seems to be fueled by two main compositional tensions. On the one hand, a strong tension towards material effects is produced through a processing of textures that refer to a particular gesture and inscription surface. The colors often present an irregularity that involves a representation of bodily gesture and manual work on the pigments (Figures 10–12); the background is almost never uniform, due to being often already made up of a represented materiality that is sometimes recognizable as belonging to an artistic tradition. This pressure is illustrated in an exemplary way by Figure 5: the third image generated, exposing the edges of a painting, looks beyond the simple execution of a visual rendering to configure what presents as an artistic object.

The second pressure, which is weaker, is carried out in relation to figurativity (that is, to the image as representing recognizable objects). Although it is possible to generate abstract images, there is a tendency in Midjourney towards the representation of recognizable figures of the world. In Figure 16, for example, the arrangement of the circle relative to the bottom, as well as the treatment of the latter, builds a landscape effect. The circle is no longer merely endowed with its meaning as a geometric shape but generates a figurative

tension towards what appears as a representation of a celestial body (the sun or the moon).

The tests that we then conducted with DALL•E show that these two tensions are not present in this generative model, even if the image generated does not comply with the prompt. Objects are no longer two-dimensional, owing to a perspective effect and the presence of shadows: instead of a circle, DALL•E displays a sphere; the requested triangle is a tetrahedron, and the square is rather a kind of square-based pyramid with a cut top. It can be noted, however, that DALL•E adopts a more neutral style compared to Midjourney; in so doing, it seems to prefigure a use that is not related to the artistic field but that would be suitable for different purposes, for example educational purposes. The rendering of colors, textures and, more generally, of the "style" displayed seems to aim at adopting an anonymous style, what we can call a neutral enunciation, as if no hand were involved in this drawing, as shown in Figure 20.

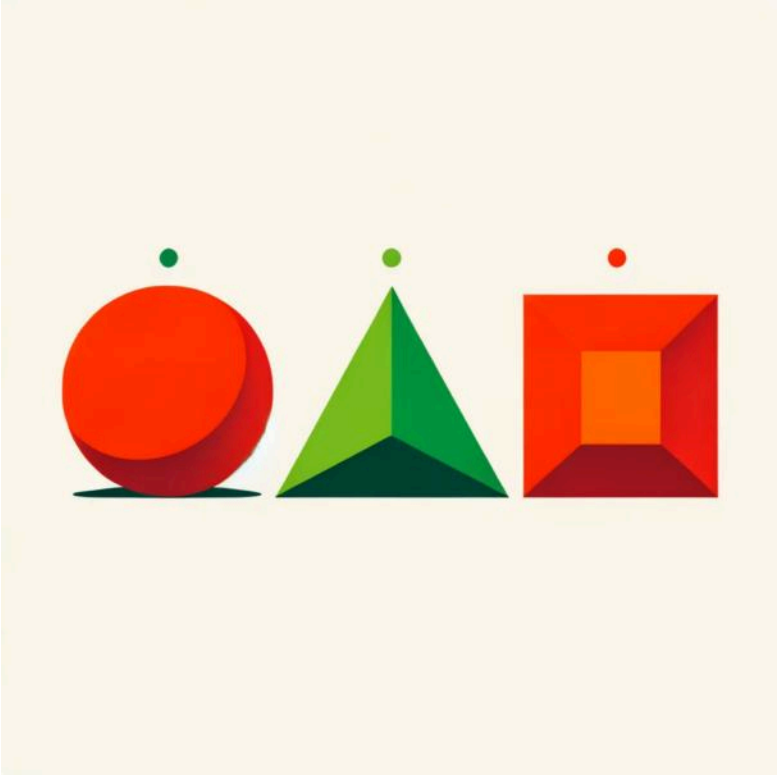
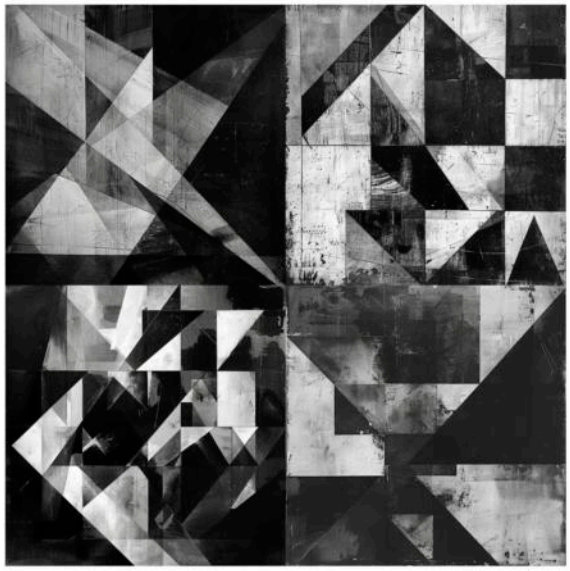


Figure 20. DALL•E. Output from the prompt: "a sequence of three two-dimensional geometric objects: a red circle, a green triangle and an orange square crossed horizontally by a white line," April 2024.

3.5. Coherence of Variations: "Vary Region" Function

Generally, Midjourney appears to have a limited efficacy in structuring compositional details. If we try to control the composition according to the plastic criteria of the spatial arrangement and the chromatic rendering of the elements, the results show a lack of

adaptation between the verbal prompt and the visual configurations. However, if more generic prompts are used, the image hews more closely to the verbal input (Figures 21–22).



Figures 21-22. Midjourney 6. Four outputs from the prompts: “an abstract image in black and white showing triangles and squares” (left), “a tension between abstract elements and figurative elements” (right), April 2024.

During these experiments, we also used the “zoom out” and “vary region” functions to act directly on the image and bypass the prompt. The “zoom out” function allows the user to extend the image beyond its initially given boundaries, while otherwise keeping the content of the original image. “Vary region” allows the user to select an internal region of a given image and to modify it. Through these two functions, Midjourney seems to adopt a different compositional logic than that of intersemiotic (verbal to image) translation. When producing an image from a prompt, the process of generation begins with random visual noise-that includes aleatoric element-so it is impossible, repeating the same prompt many times, to get each time similar images in terms of composition. On the other hand, it seems clear to us that if the user modifies a region of an already generated image, the AI seems to have incorporated a criterion of visual coherence.[14] Indeed, we realized that for some requests, such as negation or subtraction of an object or a color from an image, and for other requests, it was crucial to operate directly in response to the image and not by solely operating on the translation from verbal language to the image (this being very limiting in terms of compositional strategy). [15] These functions enable a workaround of this limitation.

Here is a sequence of operations that we performed on the same image. We first asked for a sequence of three geometric objects (Figure 23).

Figure 23. Midjourney 5. Output from the prompt: “Craft a visual sequence featuring three geometric objects,” December 2023.

Surprisingly, a human hand appears, which was not contained in the prompt, and which confirms the figurative tension that animates image generation. Following this, we highlighted the large rectangular section that contains the hand without giving an extra prompt, using the “vary region” function (Figure 24).

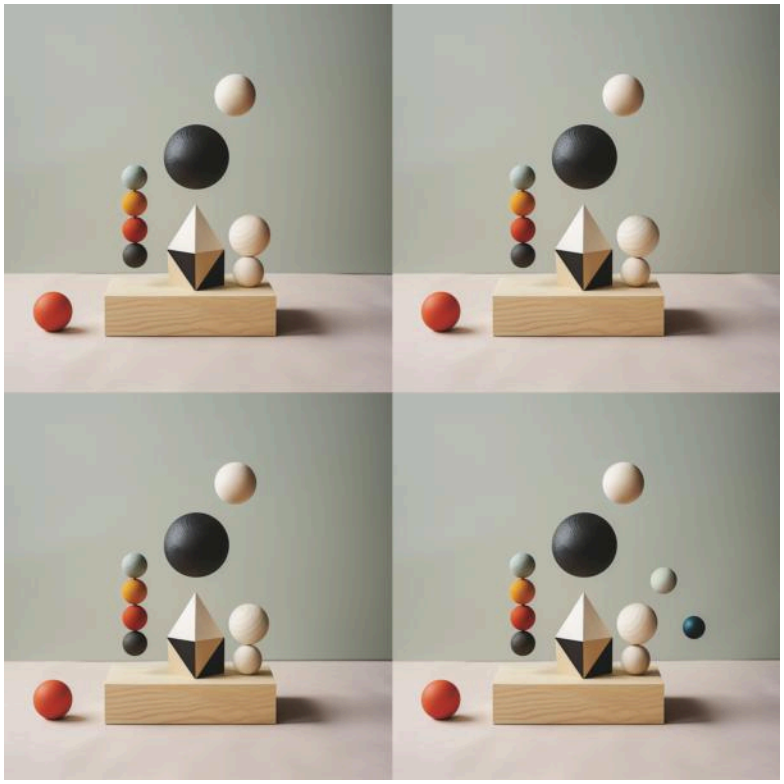


Figure 24. Midjourney 5. Four outputs from using the "vary region" function on a same image (Figure 23), applied to the part of the image that represented the human hand and without adding any further prompting, December 2023.

The pressure to produce figurative elements from verbal prompts is resolved in this image through the "vary region" function. Its action shows the existence of a criterion of visual coherence, which allows to eliminate the elements that are deemed by the user to be inconsistent—in this case, the hand that belongs to the figurative dimension—in relation to the rest of the composition. In the fourth image, two spheres have been added in a manner consistent with the composition's other elements. This function allows the user to obtain more complex images where the collaboration between AI and human operators is relatively more balanced.

4. Enunciative Articulations

The second set of tests we carried out concerns the second macro-criterion to analyze the compositional work of Midjourney and DALL•E: *enunciation*. Developed from the theorization of verbal discourse—and in particular, from the study of the pronominal and deictic marks of language mobilized in a text—visual enunciation is expressed in images through specific configurations. These configurations concern the way images articulate spaces, times, and actors (Fontanille 1989; Dondero 2020) in relation to the operations of production and of observation of these images.

4.1. The Gaze Addressed to the Viewer

Among the configurations that concern these three parameters of the enunciation theory, we find Benveniste's (1971[1966]) opposition between *discours* and *histoire* to be particularly productive. In verbal language, these two regimes of enunciation depend on whether the utterance (*énoncé*) explicitly marks or effaces the presence of its speaker and addressee (e.g., through the use of personal pronouns), as an "I" addressing a "you" in discursive enunciation or, in historical or "impersonal" enunciation, through the bleaching of deictic reference to the event of enunciation (the shift to third-person pronouns, non-present tense, non-deictic adverbs of time and space, etc.). If face-to-face oral conversation is the paradigm of the former, historical and scientific discourse are examples of the latter.

In the case of images, these two regimes are articulated in relation to other elements, the images not having a system of stabilized pronouns or other such deictic form classes (tense, evidentiality, etc.). One of the most studied configurations in semiotics concerns the gaze: if a represented figure looks toward the viewer, this gaze configures a regime of discursive enunciation, because it replicates an "I-you" type dialogue through the gazing. On the contrary, if the figures are not addressed to the spectators, the events represented are in accordance with the regime of the historical enunciation: in the absence of an address to the viewer, events seem to take place in an impersonal way and in a space-time that is represented as not shared with the viewer.

The tests we have carried out on these two types of addressing strategies pose a huge challenge to AI, as we are asking the software to take into account the image's meta-discursive dimension. This can be done through a prompt that focuses on the discourse itself (intransitive dimension) and not on the representational one (transitive dimension). If Midjourney is asked to produce the image of a man looking at the spectator, it will tend to produce an image that presents humans and spectators in the image (Figure 25), whereas DALL•E seems to be able to handle direct, dialogical references to the viewer (Figure 26).

Figure 25. Midjourney 6. Four outputs from the prompt: “a man looking at the spectator,” May 2024.

Figure 26. DALL·E. Output from the prompt: “a man looking at the spectator," May 2024.

However, if pronouns are used, Midjourney’s image generation is effective. Prompts that point to a man looking at “us,” or a woman stretching her hand towards “us,” or directing a knife towards “me” make it possible to obtain formally correct results (Figures 27–28).

Figures 27-29. Midjourney 6. Outputs from the prompts: “a man who directs his gaze towards us” (left), “a woman extends her hand towards us” (center), “a woman holds a knife towards me” (right), May 2024.

In Figure 29, only the third and fourth images comply with the enunciative configuration expressed by the prompt; this is a remarkable result, because it implies that the AI can process pronouns that refer outside of the representation space, that is, to the spectator’s space. However, this is a simple configuration, which mobilizes only one player in space. If one tries to replicate the scheme of the enunciation on interacting actors, a relevant correspondence is more difficult to obtain (Figures 30–32).

Figures 30-32. Midjourney 6. Outputs from the prompts: “two men fight each other. one of them looks at us” (left), “two men embrace. they are looking towards me” (center), “two women embrace. they are looking towards me” (right), May 2024.

In the second generation in Figure 31, only the second and third images comply with the prompt. As for Figure 32, only the third image shows a direct look towards the viewer.

Generally, these tests show that control over the syntax of the eyes and the glance is limited. Midjourney appears to be able to visually translate pronouns involving the viewer, but difficulties of manipulation arise when the prompt describes multiple actors, as observed with regard to plastic categories.

The same prompts submitted to DALL•E allow more precise control over composition, but display the same neutral style already observed in previous generations.

Figures 33-35. DALL•E. Output from the prompts: “two men fight each other. one of them looks at us” (left), “two men embrace. they are looking towards me” (center), “two women embrace. they are looking towards me” (right), May 2024.

The first image shows a more relevant translation between verbal prompts and visual configurations compared to the results obtained on Midjourney: although the two men are not engaged in a fight with each other, the different articulation of the gazes is respected, allowing control over the composition of the different parts of the image.

We carried out an analysis of the distribution of outputs pertaining to the prompt “two men fight each other. one of them looks at us,” in order to show the generality of our observations (see Figures 36 and 37). Midjourney (Figure 36) correctly keeps showing two men fighting on the 50 generated images and, as already pointed out above, consistently fails at directing the gaze of one man toward us. DALL•E (Figure 37) mostly produces images depicting two men but sometimes adds a third one (6 images out of 50) as a spectator of the scene (images 2, 11, 12, 14, 25, 42), as in Figure 33. Contrary to Midjourney, and as mentioned already, the two men are not always both engaged in a fight; instead, they are often ready to fight but do not yet touch each other, which happens in 34 images. The gaze of only one man is correctly directed at us for 6 of the remaining 16 images, which thus perfectly align with the prompt (images 4, 5, 17, 21, 31, 49). It can also be noted that the gaze of the two men is often directed at us as well (e.g., images 28,

47) but at the expense of showing a fight between the men; this observation supports our previous comment on Figure 33.

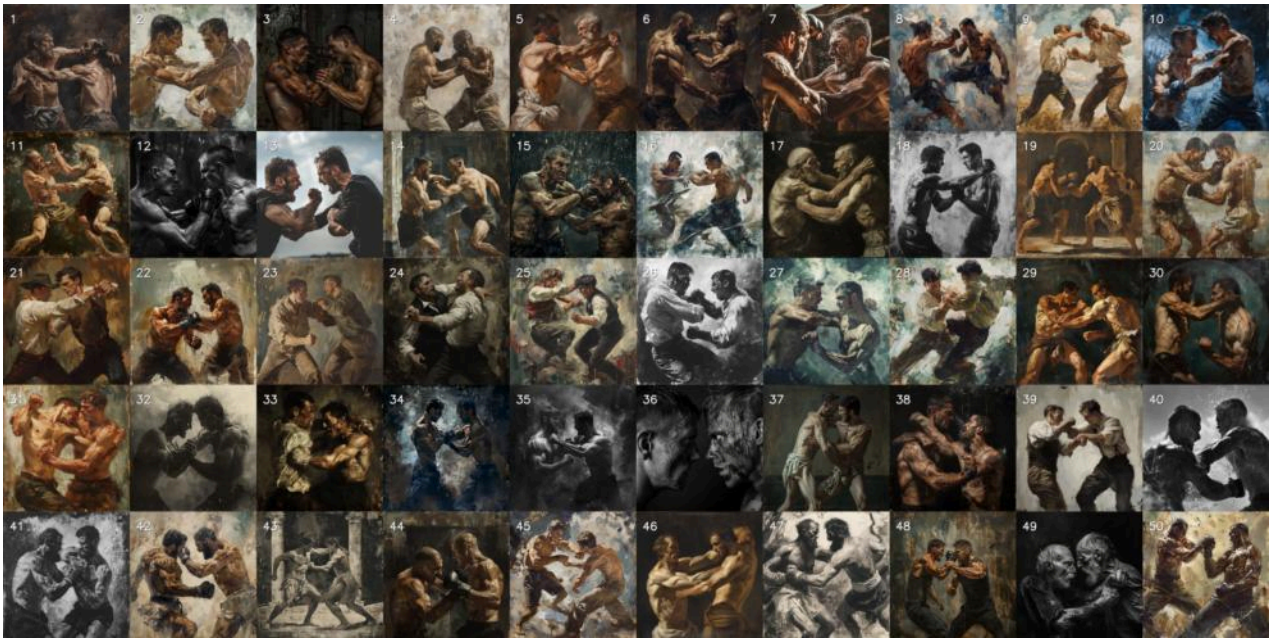


Figure 36. Midjourney 6. Outputs from the prompt: "Two men fight each other. One of them looks at us," February 2025, repeated 50 times for analysis of the distribution of outputs.

(for enlarged image: https://semioticreview.com/attachments/BigDada/)

Figure 37. DALL·E. Outputs from the prompt: "Two men fight each other. One of them looks at us," February 2025, repeated 50 times for analysis of the distribution of outputs.

(for enlarged image: https://semioticreview.com/attachments/BigDada/)

4.2. Expressing Space: Metapictural Devices

Another way to articulate intersubjectivity and space in images is to organize a path of the gaze through meta-pictural devices. The work of art theorist and art historian Victor Stoichita (1997) formalized a series of procedures that explain how images articulate space in relation to the viewer. The window, for example, invites the viewer to go beyond the represented world in the foreground and to look towards the horizon. Such arrangement has been concomitant with the stabilization and autonomization of the landscape genre. The mirror allows the addition of multiple points of view within an image. This device has been concomitant with the stabilization of the portrait and self-portrait genres. An open door builds an effect of discovery in relation to a space, pushing the gaze through it. The curtain, on the other hand, allows one to hide, to intersect and to modulate the vision and pertains to the autonomization of the "*scene de genre*" in the painterly tradition. Finally, the niche blocks the view towards the horizon and creates a tension towards the space below the representation, towards the viewer. This device has been concomitant with the stabilization of the still-life genre.[16] More generally, these formal arrangements and procedures benefit from a common configuration, that of the "frame within the frame," of which the various forms can be observed in an exemplary way in *Las Meninas* de Velázquez.

We tested some of these formal devices. Images generated from prompts depicting a woman hiding behind a curtain (Figure 38), or a woman reading at the bottom of a room, behind an open door (Figure 39), present in some cases suitable spatial configurations. It can be noted that in the case of curtains, the woman is in the foreground and the curtains are not used to articulate the space in depth. However, in this case, the prompt did not explicitly refer to depth of field. The word "curtain" in the prompt is perhaps not explicit enough for the AI model to construct an intersemiotic translation requiring a complex inference such as: "if there is a curtain, there is also a contiguous interior space." In the second generation (Figure 39), on the contrary, the position of the woman was specified in the verbal prompt.

Only in two of the four images in Figure 39 is the visual result aligned with the description contained in the prompt. In the first and third images, the woman is not behind the door but in front of it. In the second and the fourth images, the woman is actually positioned behind a door and captured visually while reading. Some peculiar visual issues appear in these images, however. In the third one, the threshold dividing the two rooms features two doors instead of one. Meanwhile, in the second, although the spatial configuration is respected, the door seems to occupy only half of the threshold dividing the two rooms, as if a second panel of the door were missing.

Figures 38-39. Midjourney 6. Outputs from the prompts: “a woman hides her face behind a curtain” (left), “a woman reads at the back of the room, behind a half-open door” (right), May 2024.

On the other hand, prompts requiring the presence of a mirror, as well as those requiring the presence of several actors interacting behind a door, are difficult to visually translate (Figures 40–41).

Figures 40-41. Midjourney 6. Outputs to the prompts: “on the right side of a large room, a mirror reflects the image of two men arguing” (left), “two women talk secretly at the back of the room, behind a half-open door” (right), May 2024.

There are few mirrors in the first series of images (Figure 40), only in the first and in the third generation. In the first image, a reflective surface is placed in front of the two men who are not arguing. What's more, the mirror is not placed towards the right but towards the left, although there are plenty of reflections in the room. As for the second series (Figure 41), in none of the four images produced is there an articulation of space that is perfectly suited to the prompt, that is, an image where the two women are talking in a space opened onto by a door. Midjourney seems to have difficulties in executing visual commands pertaining to spatial orientations such as "right" or "left," or "back of the room," as well as other instructions such as "foreground, background." However, we can see that in the second and fourth images, the open door articulates the space of the image according to the functions that Stoichita (1997) attributed to such a meta-pictorial device: increasing the depth of vision of the image and allowing observation beyond the foreground.

DALL•E produces visual translations that conform more to the verbal utterances expressed by the prompt, but nevertheless, the representations are not fully adapted to the articulation of space that has been described by the prompt. This suggests that these generative models, Midjourney in particular, have been trained on descriptions of represented and objectified situations and not on deictic nor intersubjective and inter-objective relations, that is, according to intersemiotic translations of enunciative perspective across linguistic prompt and visual image. In fact, reciprocal and respective positions of objects and actors are not respected in the visual representations.

Figures 42-43. DALL•E. Output from the prompts: "on the right side of a large room, a mirror reflects the image of two men arguing" (left), "two women talk secretly at the back of the room, behind a half-open door" (right), May 2024.

In Figure 42, the mirror is not positioned on the right side of the room if we base our reference coordinates with respect to the observer's viewing position, but the men are indeed arguing. In Figure 43, the AI has included a very special door—a transparent door—in order to ensure the visibility of the two women. We note that the instructions concerning perspectives ("behind" from the observer's standpoint) are correctly executed by the model: two women are visible and, at the same time, are placed behind an obstacle which disrupts the open visibility of a frontal representation. However, they appear to be placed towards the front of the room in which they are located, not in the back.

4.3. Visual Articulation of Temporality (Verbs of Action)

To test the articulation of temporality in still images, we used different verbs of action in the verbal prompts, as well as various actantial structures (i.e., of the arguments of the verbs). These tests place us at the very heart of intersemiotic translation: the translation between symbolic predication, a distinctive feature of verbal languages, into visual configurations, that is, configurations that obey a mereological logic (relationship between part and part, and between parts and the overall whole).

Figure 44 shows the results of our first prompt: a woman looking out a window, without specifying what she's looking at.

Figure 44. Midjourney 6. Outputs from the prompt: "a woman looks out of the window," May 2024.

As in the examples experimenting with spatiality, the act of gazing, especially if described in a prompt that concerns a single human figure, is visually translated in a relevant way. We used the verb “to look” as a kind of zero degree of action, because, in this case, it doesn’t involve a second actant and articulates an action in the present tense—our prompt does not specify the object of the gaze.

We also tested transitive verbs that involve two actantial roles. The verb “to take,” for example, implies a subject and a direct object. A prompt such as “a man picking up a glass from the floor” involves not only two actantial roles, but also the orientation of an action and a punctual temporality (Figure 45).

Figure 45. Midjourney 6. Outputs from the prompt: “a man picks up a glass from the ground,” June 2024.

Midjourney generates an image that respects the direction of the prompt, offering the user several perspectives on the gesture. We note that when the glass has not yet been picked up, there is a staging of the man’s movement to raise the glass (a gesture that enacts verticality), whereas when the glass is already in the man’s hands, the man assumes a more passive and non-vertical position: the man is crouching (i.e., his action of picking up is being done from a position on/near the ground).

Given this, we wondered whether the AI could visually translate an ongoing action on a less common object, such as a xylophone, which incidentally has its own logistical

difficulties: it is not picked up in the same way as a glass, because it demands a gesture that employs two hands instead of one, and because the object adheres in a less suitable manner to the act of being picked up than the glass.

Figure 46. Midjourney 6. Outputs from the prompt: "a man picks up a xylophone from the ground," June 2024.

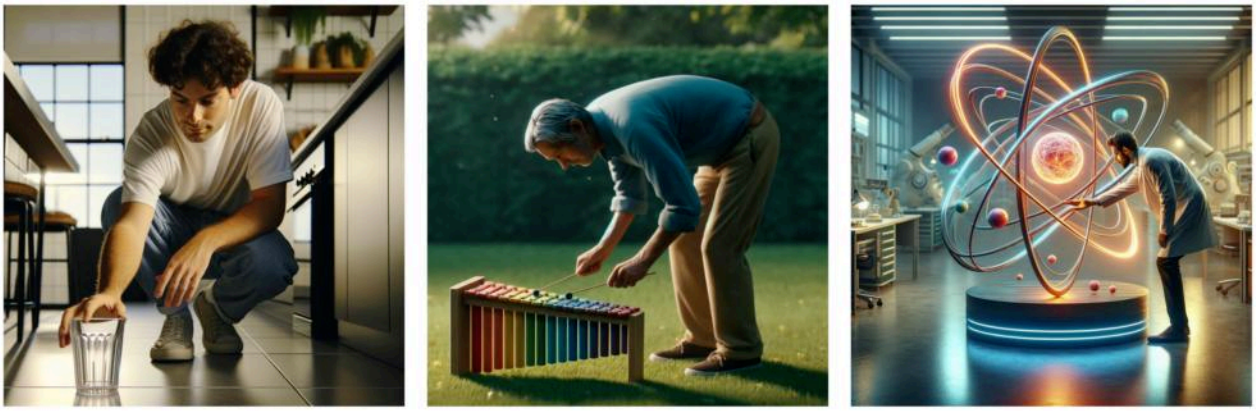
In these images, the man is never actually picking up the object. What's more, the visual configuration is more static. This difference can be explained in terms of dimensionality, compared with the previous test with a glass: a glass has a vertical configuration that is better suited to the gesture of picking up. It is also a smaller object, suited to a one-off hand gesture. In contrast, the xylophone imposes a horizontal configuration on the image, and its larger dimensions call for a more complex gesture and the need for a more complex configuration of the human body, often requiring the action of two hands. In this sense, we can speculate that Midjourney was developed by relying on a certain idea of "perceptual affordance," according to which the instructions to represent positions of the human body concern single, punctual gestures, and not to involve a broader action involving the whole body such as the one requested here (a man faced with a very large, horizontally developed object).[17]

We also tested a paradoxical configuration: picking up an atom (Figure 47), that is, an entity infinitely smaller than a glass, to experiment with the spatial and bodily

configurations that would be represented. Midjourney performs a stylization and dimensional change to make the atom proportionate to the human body.

Figure 47. Midjourney 6. Outputs from the prompt: “a man picks up an atom from the ground,” June 2024.

The same prompts as Figures 45–47 submitted to DALL•E produce different results, as shown in Figures 48–50. In the case of the glass, the visual configuration is very close to that produced by Midjourney. The xylophone, on the other hand, is not being picked up from the ground, as the man is actually playing it (perhaps translating the idiomatic sense of “to pick up” meaning the learning of a habit or skill). As for the atom, we find the same process of stylization as that carried out by Midjourney, but DALL•E, while adapting the size of the atom to make it commensurable with human size, generates a version that is too large and therefore unsuited to the gesture of picking it up. Here, it seems that DALL•E has little training on this kind of object, or rather, that the problem has to do with the training for perceptual affordances. The size is right for the atom to be in front of the man as its equal actor, but the consistency of the gesture respective to these specific objects is not correct.



Figures 48-50. DALL·E. Outputs from the prompts: “a man picks up a glass from the ground” (left), “a man picks up a xylophone from the ground” (center), “a man picks up an atom from the ground” (right), June 2024.

The main difficulties pertaining to the visual translation of verbal predication arise when two animated agents are linked by an action verb. Midjourney is occasionally able to distinguish the direction of the action, but often transforms the action itself. Figure 51 shows the output from a prompt concerning a woman punching a man.[18]

Figure 51. Midjourney 6. Outputs from the prompt: “a woman punches a man,” June 2024.

These images seem to represent a generic struggle rather than a precise direction of the struggle. And in the third image, the struggle appears like a dance. We can make the hypothesis that the configurations of visual traits associated with body gestures that require interaction between several humans are located in a relatively specific area of the latent space of these models. This hypothesis remains to be verified, but it would explain why the verb “to punch” is translated visually as if it were the verb “to dance.”

Figure 52. Midjourney 6. Outputs from the prompt: “a woman falling pushed by a man,” June 2024.

Testing more complex action sequences that are supposed to be represented in a single image shows that none of the images produced conforms to the actantial configuration denoted by the verbal prompt, nor the direction of the man’s action (of pushing) towards the woman. The figure of falling is present only in the first and fourth images, while in the other two, the figures are suspended vertically or appear to be flying. It’s also worth noting that these images abstract from their environment: The bodies seem to act in a void characterized by an aestheticizing atmosphere due to the textural treatment of the background.

As tested with yet another prompt, even the verb “to chase” produces a visual translation that does not abide the direction and overall coherence of the action verbally expressed (Figure 53).

Figure 53. Midjourney 6. Outputs from the prompt: “a group of men and women are chased by a car,” June 2024.

The four images do not display a car chasing a group of people but simply show a car and people running in the same direction, though the faces of the running people express haste. This is made clear by the presence of people behind the car, also running in the same direction. Also note the disproportionate size of the bodies in the first image, which appear larger than the car itself.

The same prompts submitted to DALL•E show other characteristics of the translation.

Figures 54-56. DALL•E. Output from the prompts: “a woman punches a man” (left), “a woman falling pushed by a man” (center), “a group of men and women are chased by a car” (right), June 2024.

Figure 54 captures what is denoted in the verbal predication: The direction and punctuality of the action are aligned to the prompt. In Figure 55, the woman is indeed falling to the ground, in an urban environment, but the man seems to be trying to help her, not push her. Finally, Figure 56 constructs an appropriate spatial and temporal configuration: the group of people is running in front of a car, though the two figures on the right are not quite placed in the direct path of the car's trajectory line. In any case, the faces of the running people express neither haste nor fear—they do not seem to be involved in or paying attention to the pursuing entity.

4.4. Temporal Aspectuality

As part of our experimentation, we also considered an additional dimension of temporality: *aspectuality.* Unlike the deictic category tense, which concerns temporal relations of succession of events, aspect concerns the perspective on, or quality of, an action represented as a process. In general, we can distinguish between temporal perspectives that are punctual (point-like, completed), durative (ongoing), inchoative (the beginning of an action), or terminative (the end of an action), though these do not exhaust the kinds of aspect that we find representable in language (and images).

First, a test of terminative aspectuality was carried out using a prompt describing a person who has just finished eating a meal (Figure 57). We then carried out a test on a prompt describing an inchoative action: a person about to start eating a meal (Figure 58).

Figures 57-58. Midjourney 6. Output from the prompts: "a person has finished eating a meal" (left), "a person about to start eating a meal" (right) July 2024.

In the first four images produced by Midjourney in Figure 57, temporality and aspectuality do not seem to be translated visually, as in none of the four cases are there any elements indicating that the meal is over. In the test pertaining to inchoative actions (Figure 58), only the first image visually translates in a manner reflecting the submitted prompt. However, there is not enough evidence to say with certainty that the person is starting to eat, and the image is appropriate only because we know the content of the prompt.

In contrast, the images produced by DALL•E offer a configuration that visually translates the act of finishing and starting to eat.

Figures 59-60. DALL•E. Output from the prompts: “a person has finished eating a meal” (left), “a person about to start eating a meal” (right), July 2024.

In Figure 59, a young man looks towards the spectator, smiling at an almost empty plate in a satisfied manner. In Figure 60, we see a person holding a fork and knife, ready to use them in front of a plate full of food. This position of the hands communicates a projective perspective. Overall, the translation of a verbal utterance endowed with a durative or inchoative temporal aspectuality into a visual configuration is a complex operation. It requires not only the translation of general verbal elements (a person) into specific configurations (the specific person represented), but also the visualization of solutions that can suggest, through a still image, the imminent unfolding of an action or its conclusion. Compared to Midjourney, DALL•E seems to perform this complex visual translation with more ease because it is able to position the body of the person represented in poses that are recognizable to a culturally Western eye.

4.5. Representing the Succession of States

We also tried to understand if it was possible to generate a sequence of shapes dividing the image into several sections. We asked Midjourney to compose images representing three different emotional states: joy, melancholy, and sadness. The prompt is very long because we tried to compose it using the suggestions provided by DALL•E.[19]

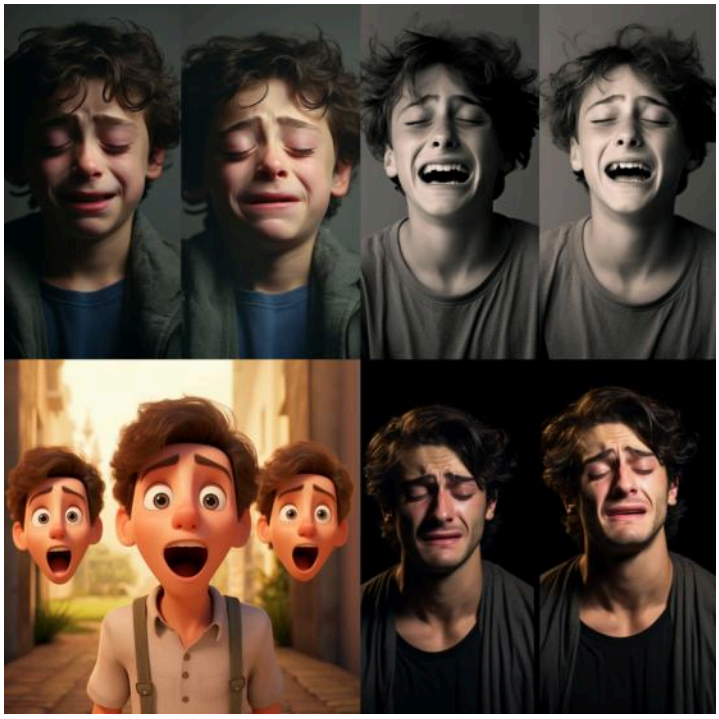


Figure 61. Midjourney 5. Outputs from the prompt: "Capture the emotional journey of Luca in three phases within a single image. Begin by portraying Luca in a moment of joy, where a genuine and radiant smile reflects his happiness. Transition to a second phase, illustrating the onset of sadness as Luca's expression changes to one of melancholy. Finally, depict the depth of emotion in the third phase, showcasing Luca in a tearful state, conveying the raw and powerful experience of crying. Utilize lighting, composition, and facial expressions to convey the nuanced emotions across these three distinct phases of Luca's emotional spectrum," December 2023.

First, the images generated are not divided into three parts, except in the third, and partially so. Secondly, the three emotions are not represented, as the majority of them only show a state of sadness.

The same prompt gives better results in DALL•E (Figure 62).

Figure 62. DALL·E. Image generated from same prompt as Figure 61, December 2023.

It's clear that sequence is important here, especially in the modulation of shadows and facial states, which change quite smoothly, whereas Midjourney creates rather abrupt transitions between one state and another.

After several experiments, we can conclude that Midjourney is not very effective in visually composing a spatial articulation—that is, a coherent sequence of shapes described by prompts. While DALL•E is more successful at "understanding" prompts, Midjourney excels in rendering textures and adhering to pictorial styles.

A second test allowed us to confirm the processing performed by Midjourney and DALL•E with respect to the articulation of a sequence composed of several states. In this case, we asked for an image divided into three parts: In the first part, a woman picks up a letter; in the second, she reads the letter; and in the third, she cries with joy (Figure 63).

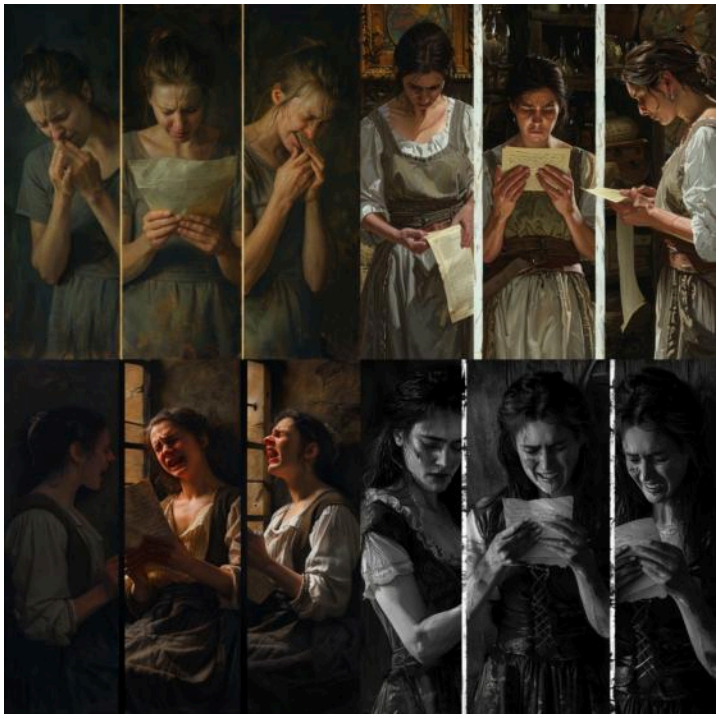


Figure 63. Midjourney 6. Four outputs from the prompt: “An image divided into three parts. In the first part of the image a woman takes a letter. In the second part of the image the woman reads the letter. In the third part of the image the woman cries with joy,” April 2024.

Figure 64. DALL·E. Output from the Prompt: “An image divided into three parts. In the first part of the image a woman takes a letter. In the second part of the image the woman reads the letter. In the third part of the image the woman cries with joy,” April 2024.

The images generated by Midjourney (Figure 63) effectively divide the representation space into three parts, but the action sequence is not adapted to the prompt. We can indeed notice a discrepancy between the emotions and the phases of the action represented. The image generated by DALL•E (Figure 64) separates the actions more clearly, but the process of finding the letter is extended to the first two sections, while the act of reading and the emotional reactions are merged into a single image. We can make the hypothesis that verbs related to emotions are not treated as isolated actions by the model. According to this hypothesis, the AI was trained to use, by default, the direct object of action verbs, even in the case of an emotional verb such as “to cry”; in other words, verbs related to emotions are treated as modifiers of actions, in this case reading, and are more difficult to understand as intransitive actions.

4.6. Image Negation

Another set of experiments on the enunciation of such images focused on linguistic negation. In current models, each word in the prompt is treated as a potential element of the image, without consideration of the interaction between words. Thus, a word preceded by “without” is actually more likely to appear in the image. This limitation is specified in the Midjourney documentation (https://docs.midjourney.com/docs/no).

Figure 65. Midjourney 6. Outputs from the prompt: “a dog without a tail,” March 2024.

In Figure 65, instead of composing a dog without a tail, the AI arranges the dog so that the tail is clearly visible. The same problem occurs when composing a Van Gogh-style landscape and specifying in the prompt that the sun should *not* appear (Figure 66). If we simply write the negative form in the prompt, the sun remains present and is even multiplied.

Figure 66. Midjourney 5. Outputs from the prompt: “Landscape in Van Gogh’s style without sun,” September 2023.

On the other hand, we can use a specific functionality: the “--no” parameter, which modalizes the generation by excluding the elements that follow. By adding “--no sun moon” to the prompt devoted to Van Gogh’s landscape, the sun disappears (Figure 67).

Figure 67. Midjourney 5. Outputs from the prompt: “Landscape in Van Gogh’s style -- no sun moon,” September 2023.

However, repeating this approach with images of the dog doesn’t work the same way as with Van Gogh’s landscape, where the AI performs a rather “clever” operation by representing a dog in a specific area of the image so that the tail is arranged out of frame. This operation is interesting because it exploits the expressive qualities of images to modulate the presence of the represented objects, translating the negation not as predicating about the object (that the dog has no tail) but with respect to the enunciative scene of seeing an image of the object (a dog where no tail is visible). With the exception of conventionalized signs of negation or interdiction (like an X over an image, or a red circle with a slash across it), visual language, unlike verbal language, does not possess discrete operators dedicated to the expression of negation. One of the configurations studied in visual semiotics to construct negation effects is the use of the “off frame” (Badir and Dondero 2016), exactly as in the images produced by Midjourney (Figures 68-69).

Figures 68-69. Midjourney 6. Outputs from the prompt: “a dog --no tail,” “a dog shot from a distance --no tail,” June 2024.

The results obtained using DALL•E concerning negation exhibit the same characteristics.

Figures 70-71. DALL•E. Output from the prompts: “Landscape in Van Gogh’s style without sun” (left), “a dog without a tail” (right), June 2024.

These cross-tests confirm some conclusions already made about these two AIs: DALL•E offers finer compositional control in terms of spatial and temporal articulation, while Midjourney is devoted to the aesthetic rendering of materials and to the simulation of pictorial styles.

4.7. Narrative Modalizations and Substances of the Plane of Expression

A final set of experiments concerns modality: *to want*, *to be able*, *to have to*, and so on. In Paris School semiotics, the focus is on the articulation of modalities pertaining to the actions of one or more characters in a narrative program. It often regards the coupling or the competition between different modalities, for instance, the case of a person who wants to perform a certain action but lacks the competence to do so, or the case of a person who doesn't want to do something but must. In other words, the focus is on applying modal verbs—*want*, *know*, *can*, *must*—to human actions. As for images, this criterion relates to a more global dimension of the image: the possibility for a still image to express a micro-narrative made up of different modalities embodied in the representation of one or more actors' gestures and objects' bodies. For these reasons, narrative modalities concern enunciation, the figurative dimension, as well as the plastic dimension of images. First of all, they concern enunciation because they establish specific dialogical relations between represented characters and viewers. Secondly, they concern the figurative dimension because they involve displaying actions that can be recognized and named using verbal language. The difference with the experiments described above is that these actions are modulated to produce micro-narratives in the prompted images. Finally, they pertain to the plastic dimension, because they require a particular composition of sensible qualities, one capable of displaying actions driven by contradictory forces. In other words, experimenting on the intersemiotic translation of narrative modalities makes it possible to assess the ability of AIs to generate images that imply the germs of narration.

Let's take the example of the verb *to refuse*: a woman refusing an apple offered by a man. We designed a prompt enabling us to test the way in which verbal refusal is translated into images through bodily posture.

Figure 72. Midjourney 6. Outputs from the prompt: “a woman refuses an apple offered to her by a man,” May 2024.

Figure 73. DALL·E. Output from the prompt: “a woman refuses an apple offered to her by a man,” May 2024.

None of the images in Figures 72 and 73 unambiguously depicts a woman refusing an apple. The challenge of displaying a complex sequence in a still image is considerable, of course; in the third and fourth images of Figure 72, the woman's hesitation seems to display a moment preceding or following the refusal.

We tested the same prompt for the exchange of another object, asking to produce not a generic image but a photograph (Figures 74–75). We did this to test the "sensitivity of the machine" to the problem of production techniques and media specificities.

Figure 74. Midjourney 6. Outputs from the prompt: "photographic image of a woman refusing an apple offered to her by a man," May 2024.

Figure 75. DALL·E. Output from the prompt: “photographic image of a woman refusing an apple offered to her by a man,” May 2024.

Midjourney generated black-and-white images (Figure 74), again characterized by the lack of a clear gesture of refusal. We might speculate that only in the last image, where the woman pouts and is three-quartered, a sign of rejection is suggested, whereas in the second image, the fact that the woman is averting her gaze could be interpreted as a sign of reluctance. DALL·E, on the other hand, translated the verbal refusal into a gesture (Figure 75), condensing it in the woman’s hand. However, in this image, the photographic texture is not perfectly reproduced, as the woman’s face appears to be composed of a mix of photographic techniques and CGI.

More generally, narrative modalizations and depicted substances of the plane of expression represent two additional criteria pertaining to the plastic dimension and to enunciation, which are the two parameters that are the focus of this article. Visual modalizations concern images that articulate the gestures of characters (figurative dimension), express a narrative tension between them (enunciation), and are intended to produce specific spatial articulations (plastic dimension). Similarly, the depicted substances of expression, relating to a variety of materials and production techniques (photography, painting, drawing), pose an additional challenge in terms of compositional control, to be tested in subsequent experiments.

Summary and Discussion

The tests above represent a systematic analysis of the compositional possibilities of the Midjourney and DALL•E generative AIs. Our aim has been to examine how these models translate verbal prompts into visual compositions in collaboration with human operators (in this case, ourselves). Indeed, the generation of images from a prompt is an intersemiotic translation between verbal texts and the concepts expressed through them and specific visual traits in images. In a way, every verbal prompt requires that the visual translation is accomplished through a wide range of computational choices. In order to set the parameters of the intersemiotic translation realized by the AI model and analytically explore the potentially infinite multiplication of visual marks that can be employed, we carried out a series of tests pertaining to two semiotic macro-criteria, the plastic dimension, and enunciation. Our tests combined a qualitative methodology with elements of quantitative analysis. On the one hand, we adapted the criteria of semiotic analysis in order to transform them into prompts that would allow us to qualitatively study the translation process. On the other hand, we chose two prompts, one relating to the plastic dimension and the other to enunciation, in order to probe the distribution of outputs from verbal prompts over a larger range (in this case, 50 images generated from a single prompt), validating our qualitative observations on a few samples.

At present, the translation realized by Midjourney and DALL•E doesn't allow users to fully control the plastic composition of the output image, especially in relation to tasks such as counting and arranging objects in specific positions in the representation space and with respect to the viewer or to other actors represented in the image. While DALL•E gives more scope for controlling plastic composition than Midjourney, Midjourney produces images conforming to widely circulated standards of aesthetic value by simulating pictorial matter and irregular inscription surfaces.

The parameter of enunciation tested the way in which the images that were generated address—or don't address—the viewer, or construct and suggest a path of gaze to him or her. We started with simple prompts, which concern the gazes of the characters towards one another and towards the viewer, and then tested meta-pictorial devices: elements such as curtains, doors, and mirrors are capable of constructing specific paths of the gaze through visual obstacles that impede or liberate vision. Enunciation also concerns different typologies of action, as well as the temporality and aspectuality characterizing them. First, we tested actions linked to verbs involving a single actant (a gaze without specifying the object of the gaze); next, verbs involving several actants (picking up objects); and finally, action verbs involving complex actions and a direction (fighting, pursuing). We explored different temporal points of view in relation to an action in progress (i.e., aspectualities): durative, punctual, inchoative (beginning), or terminative (ending) actions.

The images generated show that simple actions and temporalities impacting isolated figures are adapted in the images produced by Midjourney and DALL•E. However, prompts describing complex actions and a precise direction are often transformed in surprising ways, as in the case of fighting that turns into dancing.

Finally, we tested narrative modalities and proposed a few examples of prompts concerning the substances of the plane of expression of images. Narrative modalities concern gestures capable of expressing a micro-narrative: this is the case of the refusal of an object offered by one actor to another. As for the represented substances of expression and production techniques (photography, painting, etc.), this is an additional criterion that needs to be tested more systematically. Until now, the scientific literature has analyzed AIs primarily in terms of visual styles, particularly painting (see Manovich and Arielli 2021–2024; D'Armenio, Deliège, and Dondero 2024), to check whether they are capable of reproducing the features and elements that make up the specificity of an artist or a stabilized trend in the history of art. It appears that the criteria taken up in this article —relating to the plastic dimension and to enunciation—enable us to test more objectively the degree of control of the user in relation to the compositions of generative AIs, as it goes beyond a specific social domain.

Conclusions and Future Directions

In these last lines, we return to the issue of intersemiotic translation in order to present our findings and sketch out future avenues of research for semiotics in relation to generative AI.

With regard to the intersemiotic translation studied by semio-linguistic disciplines, image production by means of AI presents some considerable differences from other forms of translation and transduction. Image-generating AI involves an intersemiotic translation that is a multi-actor enunciation realized in a computational manner. Human operators and computational models collaborate on several levels: the computational model is trained on large datasets associating images and verbal descriptors, based on human archives extracted from different domains (art, social networks, comics, etc.). Such datasets are themselves constructed and organized with metadata that must be provided through human conceptual labor. In the same way, in using such models, human operators must necessarily adapt their semiotic logic to comply with the computational functioning of the models and exercise their agency starting from this awareness.

With respect to the combination between these different logics, a seemingly banal conclusion is that AIs not only enunciate differently than humans do; but, more interestingly, each model is itself characterized by a different enunciative style. We have seen how Midjourney produces images with a considerable degree of stylistic elaboration,

and that it seeks an aesthetic effect by working not only on the composition but also on the simulated materiality of the substrates and the inscription gestures. In contrast, DALL•E adopts a more neutral, almost didactic style, but allows for better control of the spatial composition and of the actions depicted.

The most unexpected results, however, concern exactly the supposed difference between human and computational logic. Both Midjourney and DALL•E seem to struggle exactly with tasks that are generally associated with computation and spatiality. Tasks such as producing an exact number of objects or visual features, the arrangement in specific points of the representation space, are in fact rarely accomplished in the produced images. In the case of actions pertaining to a specific directionality, their translation into images is often not adequate to the prompt. Or again, verbs involving interaction between two humans, such as fighting, are visually translated into other actions involving similar but distinct body configurations, such as dancing. In other words, our tests appear to raise questions about the relationship between computational logic and the tasks typically associated with it, such as counting, spatial arrangement, and directional orientation. Working on low-level plastic features such as the organization of space, colors, and shapes allows us to identify these features that the models struggle with which might be naively associated with computational logic.

Looking forward, more categories belonging to the post-structuralist semiotic tradition could be tested. For instance, it would be possible (and seems necessary) to test the tensions between the figurative and plastic dimensions, or plastic categories that have been left out of our examination, such as visual transparency. Visual semiotics has also extended its conceptual apparatus to consider different media: a future series of tests could study the control of blurring associated with photographic genres such as portraits, war photography, and sports photography. In this way, it would be possible to articulate low-level parameters such as colors, shapes, and spatial arrangement with genre configurations and cultural stereotypes. For this, it would be necessary to conceive an experimental protocol capable of identifying how stereotypes are expressed through the specific resources of images, and how they are manipulated by the AI. In short, it seems to us that the theoretical gesture we have proposed in this article, that of adapting the criteria of semiotic analysis into principles of composition, could be extended to other analytical tools and other semio-linguistic traditions (e.g., one wonders if categories from social semiotics, linguistic anthropology, or other traditions could similarly be operationalized in such elicitation experiments).

A final remark that we feel it is important to emphasize concerns the role that semiotics could play in the larger field of the study of generative AI. Readers will likely have noticed that some of our remarks stem from an evaluative posture. We aim to reaffirm the

importance of this stance, even if we have shied away from overly normative judgments: in the field of computer science, AI models are continuously evaluated for their performance through statistical inquiries, which provide an indirect understanding of visual composition. We believe that semiotics can—and should—offer parameters of analysis for evaluating AI, beginning with the expertise of semiotics regarding the expressive articulations of images. That is, semiotics provides a form of expertise in visual analysis that is missing in existing metrics for evaluating AI performance.[20]

In turn, these kinds of experiments and results contribute to semiotic theory in at least two crucial ways: the first concerns the notion of enunciation and notably of "enunciative instances" (Fontanille 2004; Coquet 2007). If, traditionally, "enunciative instances" pertain to humans and common objects of everyday life, or non-subjective agencies such as the unconscious and automatized body movements (Fontanille 2004; Coquet 2007), through generative AI it becomes possible to include and analyze algorithms and aleatoric systems in our accounts of enunciative instances. These instances are agencies semiotics has to take into account to modify its enunciative theory. The second way in which these experiments contribute to semiotic theory again pertains to enunciation theory but from a diachronic point of view: these generative models make it possible to concretely analyze what is called *enunciative praxis* (Fontanille 2006; D'Armenio, Deliège, and Dondero 2024; Dondero 2025, n.d.), that is, the constant modification of culturally inherited virtualities in terms of styles, taste, and cultural trends.

Of course, such evaluations cannot be understood strictly on the basis of statistical-informatic functioning nor on purely textual or aesthetic analysis; rather, they must also take into account two other factors, which are at the beginning and at the end of the chain of such translations: (1) First, the skills of programmers in establishing the architecture of the different models, but also the annotations already contained in a given database, which includes cultural judgments and social biases. In this sense, various cultural (and so ideological) operations are being carried out in the transition between word and image. There are numerous intersemiotic translations, in other words, that are already baked into and presupposed by such models. Indeed, the database is itself not only a machine that translates but is the product of translation. (2) Second, the objectives and practices of users with respect to these models. Such objectives and practices can be aesthetic, artistic, commercial, and scientific, among other possibilities. The social life of these automatically generated images has not been sufficiently studied. While one common statement is that these images have no history, no cultural roots, and no aura, for the authors of this article, the most important question is one within an anthropological frame, namely, about the future of these automatically generated images: Are they thrown away and forgotten? Are they used for commercial, educational, or scientific purposes? How and with what effects do they circulate from one social domain to another?

While these final questions fall outside the scope of this article, we believe that such an approach will be essential for a comprehensive semiotic and anthropological analysis of AI-generated images. The experiments in this study are one first step towards these posing, and answering, these questions.

Acknowledgments. The authors wish to warmly thank Constantine Nakassis for his unwavering help with the language and the ideas expressed in this text, as well as Meghanne Barker for her valuable suggestions.

Endnotes

1. Midjourney is a program for generating images from prompts, produced by the company of the same name. Website: https://www.midjourney.com/home DALL•E is a family of programs for generating images from prompts, currently available in three versions, the latest being DALL•E 3, produced by the OpenAI company. Website: https://openai.com/index/dall-e-3/↩

2. This conception of hearing in relation to images has been particularly developed in Nakassis in press.↩

3. There are some exceptions to this, though they are relatively rare; see, for example, Somaini 2023; Manovich and Arielli 2021–2024. ↩

4. We take up here the typology of signs elaborated by Charles Sanders Peirce and adopted by Eco (1999): for the purpose of this article, we use the term iconism for perception and the term of hypo-iconism for visual signs.↩

5. Overall, the analysis of the uttered enunciation could also be useful for future studies regarding the way enunciative praxis (Fontanille 2006) characterized by the collaboration/assemblage between human operators and generative AI can work without relying on criteria, widely used in computer science, such as fidelity and photorealism (Lee et al. 2023; Li et al. 2024). ↩

6. The case of biases is particularly interesting to debate, yet this is not the scope of this article, so we will keep the discussion short. In fact, by the training process itself of AI models, a good model is, in a way, a model that learns biases efficiently. Indeed, if one trains a model by always showing it, for example, men as CEOs of companies, then there is absolutely no reason for the model to envisage that a CEO should be depicted as a woman. However, this will be perceived by us, users of these AIs, as a “undesirable bias” of the model that should probably be avoided, that is, we would expect either a man or woman when prompted to generate a CEO with no other specification. While the claim of bias is a human judgment on the model, it is actually also a criticism indirectly addressed to the database used to train the model, wherein that database is a representation of our world and, of course, its own biases. This helps us to understand

that current AI models are mostly statistical models, in the sense that they predict or generate the most plausible output based on what they have processed during their training. ↩

7. Note that a smaller version number does not indicate anything on the quality or maturity of the product. Some companies choose to release new versions more often, delivering incremental improvements on a regular basis, while others opt for rarer but more significant changes from one version to the next.↩

8. For instance, in the case of CEOs, while not being the topic of this article, we asked DALL•E to generate 10 images of "CEO giving a speech"; we received ten images of women from various ethnicities. This finding is surprising at first sight as one would expect roughly five men and five women or a bias towards more men, as more represented at this position in our society, thus likely in the databases. However, in an attempt —failed in this case—to be more "fair," OpenAI likely revised the prompt to empower women, a practice that perhaps they anticipated receiving more positive feedback from the public. Let us note that Midjourney and Stable Diffusion both generated ten white, middle-aged men, well-dressed on a blue background. Beyond the discussion on biases, this experiment indicates that the owners of the models may choose to influence the output of their models, and thus might amplify existing biases, or introduce new ones, imposing their own vision and judgment on the world. ↩

9. Some low-tier versions of the Stable Diffusion family of models are free and open for developers and scientists, yet our early experiments showed relatively poor performance, making it necessary to switch to top-tier more expensive models for higher quality results. This price factor, balanced with the reported quality and availability of the models at the time of writing this article, and the fact that we do not aim to compare all existing text-to-image programs but rather examine semiotic properties of this technology in general, made us opt for selecting DALL•E and Midjourney.↩

10. We know that even in cases where the prompt is very precise, the image translation must always "add" details, something that is the case with all transductions and transformations (Silverstein 2003), be they image translation or adaptations (such as from a novel to a film). What face will this character have? What clothes? What "style"? In a way, every verbal prompt requires the image to make "choices."↩

11. The initial experiments were conducted in September 2023 and the analyses of the wider distributions from prompts in February 2025, both with the version of Midjourney 5 and Midjourney 6, through the Discord platform.↩

12. On this topic, see Meyer 2023 and Dondero n.d. The main idea is that the results of a prompt are relevant only if they give some information about the extent of a region of the database linked to the terms used in the prompt.↩

13. It should be noted that positioning problems are well known to the creators of DALL•E 3, as indicated in the model's technical report, in Section 5, "Limitations": https://cdn.openai.com/papers/dall-e-3.pdf.↩

14. We already noted this tendency towards coherence in Midjourney's composition in another test focusing on blurriness and sharpness in the fusion of the styles of Leonardo and Rothko. On this topic, see D'Armenio, Deliège, and Dondero 2024.↩

15. On the debate regarding the semiotic autonomy of visual language, that is, the fact that an image can comment and reflect on itself, against the language centrism of Benveniste and Barthes's work, see Fabbri 1998; Dondero 2020; Dondero, Beyaert, and Moutat 2017; and Lagopoulos et al. 2024.↩

16. For some comments on these procedures studied by Stoichita, see Dondero 2020, which tests these procedures on scientific images in astrophysics and non-invasive archaeology.↩

17. In order to eliminate possible lexical ambiguities concerning the verb "to pick up," we tested the same prompt but with a different object. The prompt "a human picks up a guitar from the ground" produced an appropriate image. These tests allow us to confirm that it is the object, in this case the xylophone, that perturbs the visualization of the action.↩

18. Midjourney's user guidelines (https://docs.midjourney.com/docs/community-guidelines) prohibit the generation of violent content. Therefore, although the images have been generated, it is possible that the act of violence has been automatically toned down.↩

19. Given the disappointing results of the images produced, we asked DALL•E to provide us with a more effective prompt for use in Midjourney. The version proposed by ChatGPT is long and shows an obvious dramatization of the three emotions.↩

20. Further, such models are themselves designed with such normative criteria (e.g., of fidelity, appropriateness, etc.) in mind which means that, as part of the phenomenon itself, semiotic analysis cannot but include this dimension. A fuller study would include empirical attention to the range of such evaluative postures and practices (e.g., among designers, regulators, diverse users), though this is beyond the scope of this article.↩

References

Badir, Sémir and Maria Giulia Dondero, eds. 2016. *L'image peut-elle nier?* Liège: Presses universitaires de Liège.

Basso, Pierluigi. 2000. Fenomenologia della traduzione intersemiotica. *Versus Quaderni di studi semiotici* 85–87:199–216.

Benveniste, Émile. 1971[1966]. *Problems in General Linguistics.* Florida: University of Miami Press.

Bordron, Jean-François. 2011. *L'iconicité et ses images: Études sémiotiques.* Paris: PUF.

Coquet, Jean-Claude. 2007. *Phusis et Logos: Une phénomenologie du langage.* Paris: Presses Universitaires de Vincennes.

D'Armenio, Enzo, Adrien Deliège and Maria Giulia Dondero. 2024. Semiotics of Machinic Co-Enunciation. About Generative Models (Midjourney and DALL•E). *Signata* 15. https://doi.org/10.4000/127x4

Dondero, Maria Giulia. 2020. *The Language of Images: The Forms and the Forces.* Cham, Switzerland: Springer.

Dondero, Maria Giulia. 2025. Semiotics of Artificial Intelligence: Enunciative Praxis in Image Analysis and Generation. *Semiotica* 2025(262):111–46. https://doi.org/10.1515/sem-2024-0195

Dondero, Maria Giulia. N.d. Enunciative Praxis and Enregisterment in the Domain of Generative Artificial Intelligence. *Semiotic Review*, in preparation.

Dondero, Maria Giulia, Anne Beyaert-Geslin and Audrey Moutat, eds. 2017. *Les plis du visuel : Réflexivité et énonciation dans l'image.* Limoges, France: Lambert Lucas.

Dusi, Nicola. 2015. Intersemiotic Translation: Theories, Problems, Analysis. *Semiotica* 206:181–205. DOI: 10.1515/sem-2015-0018

Dusi, Nicola and Siri Nergaard, eds. 2000. Sulla traduzione intersemiotica. *Versus Quaderni di studi semiotici* 85–87:3–54.

Eco, Umberto. 2000[1997]. *Kant and the Platypus: Essays on Language and Cognition.* San Diego, CA: Harcourt.

Eco, Umberto. 2008[2000]. *Experiences in Translation.* Toronto: University of Toronto Press.

Fabbri, Paolo. 1998. *La svolta semiotica.* Bari-Roma: Laterza.

Floch, Jean-Marie. 1985. *Petites mythologies de l'œil et de l'esprit. Pour une sémiotique plastique.* Paris-Amsterdam: Éditions Hadès-Benjamins.

Fontanille, Jacques. 1989. *Les espaces subjectifs. Introduction à la sémiotique de l'observateur.* Paris: Hachette.

Fontanille, Jacques. 2004. *Soma et séma. Figures du corps.* Paris: Maisonneuve et Larose. (Expanded edition. 2011. *Corps et sens.* Paris: PUF.)

Gal, Susan. 2015. Politics of Translation. *Annual Review of Anthropology*, 44:225–40.

Greimas, Algirdas J. 1989[1984]. Figurative Semiotics and the Semiotics of the Plastic Arts. *New Literary History* 20(3):627–49.

Greimas, Algirdas J. and Joseph Courtés. 1982[1979]. *Semiotics and Language: An Analytical Dictionary.* Bloomington: Indiana University Press.

Ho, Jonathan, Ajay Jain and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. *Neural Information Processing Systems Proceedings.* https://doi.org/10.48550/arXiv.2006.11239

Jakobson, Roman. 1959. On Linguistic Aspects of Translation. In R. Brower, ed. *On Translation,* pp. 232–239. Cambridge, MA: Harvard University Press.

Lagopoulous, Alexandros, Karin Boklund-Lagopoulou, Maria Giulia Dondero, Jacques Fontanille, Rea Walldén, and Maria Katzaridou. 2024. *Semiotics of Images. The Analysis of Pictorial Texts.* Berlin: De Gruyter.

Lee, Tony, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Benita Teufel, Marco Bellagente, Minguk Kang, Taesung Park, Jure Leskovec, Jun-Yan Zhu, Li Fei-Fei, Jiajun Wu, Stefano Ermon and Percy Liang. 2023. Holistic Evaluation of Text-to-Image Models. *NeurIPS Datasets and Benchmarks Track.* https://arxiv.org/abs/2311.04287

Li, Baiqi, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Tiffany Ling, Xide Xia, Pengchuan Zhang, Graham Neubig and Deva Ramanan. 2024. GenAI-Bench: Evaluating and Improving Compositional 451 Text-to-Visual Generation. *Conference on Computer Vision and Pattern Recognition.*https://arxiv.org/abs/2406.13743

Manovich, Lev and Emanuele Arielli. 2021–2024. *Artificial Aesthetics: Generative AI, Art and Visual Media.* http://manovich.net/index.php/projects/artificial-aesthetics

Meyer, Roland. 2023. The New Value of the Archive: AI Image Generation and the Visual Economy of "Style." *IMAGE. Zeitschrift für interdisziplinäre Bildwissenschaft,* Jg. 19, Nr. 1, S., pp. 100–111. DOI: http://dx.doi.org/10.25969/mediarep/22314

Nakassis, Constantine V. In press. Voicing, Looking, Perspective. *Current Anthropology.*

Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger and Ilya Sutskever. 2021. Learning Transferable Visual Models from Natural Language Supervision. *International Machine Learning Society.* https://arxiv.org/abs/2103.00020

Ramesh, Aditya, Prafulla Dhariwal, Alex Nichol, Casey Chu and Mark Chen. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. https://doi.org/10.48550/arXiv.2204.06125

Rombach, Robin, Andreas Blattmann, Dominik Lorenz, Patrick Esser and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. *Conference on Computer Vision and Pattern Recognition.* https://arxiv.org/abs/2112.10752

Silverstein, Michael. 2003. Translation, Transduction, Transformation: Skating 'Glossando' on Thin Semiotic Ice". In P. Rubel and A. Rosman, eds. *Translating Cultures: Perspectives on Translation and Anthropology,* pp. 75–105. Oxford, UK: Berg.

Somaini, Antonio. 2023. Algorithmic Images: Artificial Intelligence and Visual Culture. *Grey Room* 93:74–115.

Stoichita, Victor. 1997[1993]. *The Self-Aware Image: An Insight into Early Modern Meta-Painting.* Cambridge: Cambridge University Press.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser and Illia Polosukhin. 2017. Attention Is All You Need. *Neural Information Processing Systems Proceedings*. https://arxiv.org/abs/1706.03762

Zhang, Chenshuang, Chaoning Zhang, Mengchun Zhang, In So Kweon, and Junmo Kim. 2024. Text-to-image Diffusion Models in Generative AI: A Survey. ArXiv. https://doi.org/10.48550/arXiv.2303.07909