A semiotic methodology for assessing the compositional effectiveness of generative text-to-image models (Midjourney and DALL·E)

Enzo D'Armenio[1,2], Adrien Deliege[1], and Maria Giulia Dondero[1,2]

[1] University of Liège, Belgium
[2] F.R.S.-FNRS, Belgium

Abstract. In this paper, we intend to propose a unified methodology in order to assess the effectiveness of text-to-image generation models. Existing evaluation methods are based on criteria such as fidelity and alignment between verbal prompts and produced images, but the human and CLIPScore evaluations mix together criteria of position, action, and photorealism. Instead, we intend to adapt the model analysis elaborated in visual semiotics in order to identify a set of discrete visual composition criteria upon which to establish the accuracy of the assessments. We will therefore distinguish three fundamental dimensions -plastic categories, multimodal translation, and enunciation- each articulated in multiple specific sub-criteria. We will then test these criteria on Midjourney and DALL·E while providing the abstract structure of the prompts in order to allow them to be used in future quantitative analyses.

Keywords: text-to-image · Midjourney · DALL·E · semiotic evaluation

1 Introduction

The scientific debate on artificial intelligence (AI) has exploded following the release of generative models. One of the most important dates in the recent history of AI is November 30, 2022, the day on which the first version of ChatGPT, a text-to-text generative model capable of reacting to users' verbal language requests, was released. In the same period, generative models capable of producing still images from users' verbal descriptions (text-to-image models) were released: to mention the best known, DALL·E was released in January 2021, Midjourney in July 2022 and Stable Diffusion in August of the same year. Although these dates might be moderately important from a technical point of view, because these models result from a refinement of existing processes, they are utterly important from an interdisciplinary point of view: thanks to these models, artificial intelligence stopped being a niche object whose functioning concerned expert communities, to become a set of tools within everyone's reach, potentially deployable in any social context.

The existing literature has identified and discussed some of the consequences of the dissemination of these models. Many of these contributions have focused on their social impact in order to isolate the most important and pressing issues arising from their use [7]. Within this strand of studies, particular attention has been paid to the social biases reproduced within the representations made with these models [10] and to the amplification of stereotypes derived from their easy use [1]. More specific contributions have questioned the consequences of the use of these models within specific social domains, in particular that of art [6].

A relatively unexplored axis of studies, which we propose to develop in this contribution, concerns the assessment of the compositional efficiency of text-to-image generation models. The general problem concerning existing studies is the lack of a coherent theoretical and methodological framework that can articulate evaluation criteria with respect to visual composition. Indeed, some of these studies are based on the proposition of benchmarks that can systematize the evaluations expressed by humans with respect to text-to-image generations [14]. However, human preferences are actually a macro-criterion that depends on the combination of several parameters such as the appropriateness of the represented elements with respect to prompts, their arrangement, and their photorealism. Other contributions focus on the perceptual criteria that underlie humans' evaluations in order to judge the fidelity of a computationally generated image [15]. Once again, perceived fidelity is an indirect criterion that depends on lower-level features belonging to the composition of the image itself. More generally, perceptual and fidelity evaluations made by humans are used in accordance with a quantitative criterion, adopting the implicit logic that a large number of concordant evaluations on a given compositional feature is adequate to assess the overall effectiveness of these models [9, 11]. However, although numerous, individual's perceptual evaluations are not derived from a knowledge of the criteria that articulate the compositional qualities of an image. Similar considerations can be addressed to studies that combine CLIPScore [5] and human ratings [8]: the criteria of fidelity and alignment are not supported by an articulation of the specific visual traits that allow for an objective assessment. Other studies choose to use specific criteria, such as the ability to count [13], but rarely organize these criteria according to a unified and coherent structure. Overall, the existing literature underestimates criteria concerning visual composition with respect to verbal utterance, to rely on general macro-criteria derived from common sense. A consequence of this shortcoming is that the tests conducted so far unintentionally associate multiple complex criteria within a single prompt, being unable to distinguish the effectiveness of a model on each of these criteria.

On the contrary, our contribution proposes to start from a unified methodology, based on the criteria of visual composition produced in the domain of Paris School Semiotics from the Eighties. Visual semiotics has proposed qualitative models for the analysis of images, considering them as meaning-producing totalities [2–4, 12]. Just as linguistics has analyzed the fundamental structures and criteria that articulate verbal languages, so visual semiotics has examined the criteria that allow an image to be meaningful through compositional configurations. It is only through the identification of as many relevant compositional criteria as possible that the fidelity of an image to a verbal prompt can be judged. In other words, in this contribution we intend to propose a unified methodology for evaluating the effectiveness of text-to-image generation models, using the qualitative parameters of semiotic analysis of visual composition as quali-quantitative criteria for evaluating text-to-image models.

In the first part of the paper, we provide a methodological description. We first present the criteria of semiotic image analysis relevant regarding their use as composition criteria for text-to-image generative models. These criteria are organized according to three fundamental dimensions: the plastic dimension, multimodal translation and the enunciational dimension. The plastic dimension concerns the organization of spaces, colors and visual forms within a representation; the multimodal translation concerns the relationship between spatial and temporal actions expressed verbally and spatial and temporal actions expressed visually; the enunciational dimension concerns the relationship that the image establishes with the viewer and between the different actors represented. These three dimensions of compositional analysis, developed by semiotics to analyze works of art, are transformed into an organized set of composition criteria to be applied to image generation. We aim to develop a repeatable qualitative-quantitative evaluation protocol: a unified methodology based on visual composition, a set of discrete criteria and prompt types related to each of them.

In the second part of the paper, we will apply the identified criteria to image generations, evaluating the degree of accuracy of Midjourney v6 and DALL·E 3.

Findings

- Prior human and computational criteria to assess the compositional effectiveness of text-to-image model is questionable because it is based on macro-criteria such as fidelity and alignment. We provide a theoretical and methodological framework based on the visual composition in order to articulate as many atomistic criteria as possible.
- Human perception and CLIPScore are unreliable criteria to test these models. We provide a set of criteria articulated organically around three dimensions of text-to-image generations: plastic, multimodal and enunciational.
- Current literature focuses on the number of prompts and rating in order to assess text-to-image model. We propose a quali-quantitative design pertaining to the type of prompt to be utilized to test different specific tasks.

2 Methodology

Since the 1980s, post-structuralist semiotics has developed a methodology for analyzing the meaning of any kind of image as structured visual discourse. Some of these criteria are particularly suitable to be transformed into composition and evaluation criteria to test the efficiency of text-to-image models. In this section we briefly present the selection of analysis criteria developed by visual semiotics, and how we adapted and operationalized them to test computational models. The choice of isolating as many criteria as possible serves to avoid an error in test design: if a prompt contains several criteria such as spatial arrangement, the textural materiality of an image (*e.g.* photographic or pictorial), and actions (static or dynamic), it becomes difficult to identify which task has troubled the model when it produces an image that is not faithful to the prompt. Discrete criteria make it possible to evaluate the model's effectiveness on each composition criterion, and avoid confusing the evaluation of a specific task with the evaluation of the combination of several tasks. For example, a model may be effective in spatial arrangement, but prove ineffective when it has to associate spatial arrangement with the representation of human actions.

In order to identify the criteria, we will start from three macro-dimensions elaborated by visual semiotics to analyze image composition: plastic categories, multimodal translation and enunciation.

Plastic categories. The first dimension concerns plastic categories. Alongside a reading that evaluates the recognizable figures of the world (trees, people, objects, landscapes), it is in fact possible and necessary to adopt another, more formal reading, which is interested in the features that, in an image, are not linked to the recognition of an object in the world, but to configurations specific to visual language: everything that concerns the organization of a two-dimensional composition unfolding within a frame that contains it and separates it from the environment. The most fundamental plastic category, which is the starting point for analyzing any image, is that of topology, because it allows us to study and evaluate the relationship between the frame and the center of the image, by making it possible to construct the axes of segmentation of the surface (left vs. right, top vs. bottom, center vs. periphery) as well as "forces" that determine the dynamics of the images (centrifugal forces that produce tension towards the center vs. centripetal forces that build tension away from the center).

The chromatic component, by contrast, concerns the differences, oppositions and similarities in the saturation and luminosity of the colors within an image.

Finally, the eidetic category focuses on the types of shape contours, more or less linear or curvilinear, more or less fragmented.

Multimodal translation. The second dimension relates to the multimodal translation of verbal discourse into visual configurations. Verbal language constructs its meaning in relation to predication, distributing subject and complement positions with respect to the verb. Visual language, on the contrary, does not construct meaning through predication, but through composition: by relating the parts of the image to other parts and to the whole. This dimension concerns the translation between the verbal and predicative logic of verbal language and the compositional logic of visual language. We have translated this principle in relation to the complexity of the action and its temporality. The first criterion therefore concerns the type of action and can be distinguished into prompts without actions ("a person" or "a dog"), prompts with simple actions and without object complements but with directionality ("a person looks"), prompts with complex actions that are gradually more dynamic and contain an object complement ("a person picks up an object", "a person chases another person"). The second criterion concerns the temporality of the action: verbs of actions that are finished ("an object has fallen"), of actions that are to begin ("an object is about to fall"), of durative actions ("an object is falling").

Enunciation. The third dimension pertains to enunciation, and in particular the way in which the image constructs a relationship with the viewer, or relates the characters represented. In semiotics, enunciative analysis concerns many complex criteria such as how the image constructs the spatiality of objects and characters, how it articulates different effects of temporality. For the purposes of this paper, we have selected the criteria best suited to be reduced to a simple, discrete prompt. The first criterion concerns two general configurations: the enunciation-discourse or the enunciation-story. In verbal language, these two regimes depend on the fact that the speaker can engage in first-person discourse - an "I" addressing a "you" (enunciation-discourse), or erase the pronominal marks by using "he" or "it" in order to construct an impersonal regime of enunciation: this is the case, for example, in historical and scientific discourse. In the case of images, these two regimes are articulated in relation to other elements, since images do not have a pronoun system. One of the configurations most studied in semiotics concerns the gaze: if a represented figure looks towards the viewer, this gaze configures a discourse enunciation regime, because it replicates an "I-you" dialogue through the gaze. If, on the other hand, the figures are not addressing the viewer, the events depicted fit into the enunciation-history regime: without an address to the viewer, the events seem to take place in an impersonal manner.

The second criterion concerns the pictorial devices studied in the history of art [12]: particular visual objects within the representation (mirrors, windows, curtains, doors) that allow the space of the representation to be structured and direct the path of the gaze, concealing, allowing a glimpse, inviting one to look beyond or away. The window, for example, invites the viewer to look beyond it towards the horizon. This device helped stabilize the landscape genre. The mirror, and any reflective surface, allows several points of view to be added to the same image. This device helped stabilize the self-portrait and portrait genres. An open door creates an effect of discovery in relation to a space, drawing the eye through it. The curtain, for its part, hides, reveals and modulates vision. These are objects that configure general procedures for articulating space, and although they have been developed in the context of the study of artistic images, they are used in any type of image, whether photographic, pictorial or realized in computer graphics, and in multiple social domains such as art, scientific representations, advertising, and visual narratives

Overall, these criteria allow for a randomization of tests: For plastic criteria, it is sufficient to substitute the type of geometric object, its color and position to construct a statistical survey. For multimodal translation, these are variations and reductions of the following verbal structure: "a person/animal/object + action verb + temporality + aspectuality + spatial circumstantials". An overview of these criteria is provided in Tab. 1.

Table 1: Evaluation criteria and test prompts for visual composition.

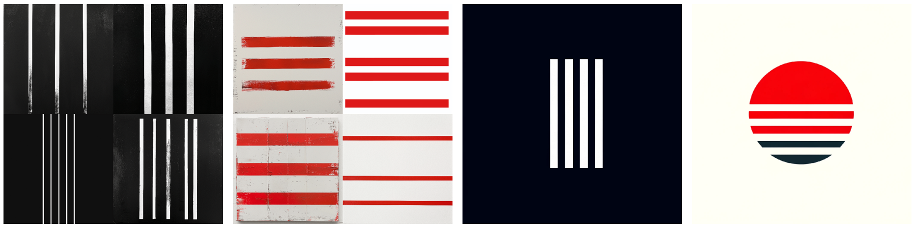
Dimension of the visual composition	Evaluation criteria	Type of test prompt
Plastic categories	Topological criterion: organization of the image spaces according to parameters such as left/right, top/bottom, center/periphery	"An abstract object disposed in a section of the image"
	Chromatic criterion: the oppositions and similarities in the saturation and luminosity of the colors represented	"One or more abstract objects characterized by specific colors"
	Eidetic criterion: lines, shapes, and contours organized according to linear, curvilinear, fragmented strokes, etc.	"One or more abstract objects shaped in linear/curvilinear/fragmented manner"
Multimodal translation of actions	Type of action: absent, static, dynamic, structured	"An animal/person"; "An animal/person watching"; "A person picking an object"
	Actantial distribution: verbs with only a subject, with a subject and one or several object complements + spatial or other circumstantials	"A person watching something"; "A person picking up an object"; "A person picking up an object from the ground"; "A person giving an object to another person"
	Aspectuality: an action in its initial moment, an action during its unfolding, an action in its final phase	"A person about to eat a meal"; "An artist is finishing painting a picture"; "A person has finished reading a book"
Enunciation	The direction of the gazes of the character represented	"A person staring at us"; "A person staring at us while doing something"
	Represented devices that direct the gaze of the spectator and articulate the represented space	"The reflection of one or more persons through a mirror"; "One or more persons doing something behind an open door"; "A view of one or more persons doing something behind the windows"

3 Results

3.1 The plastic dimension

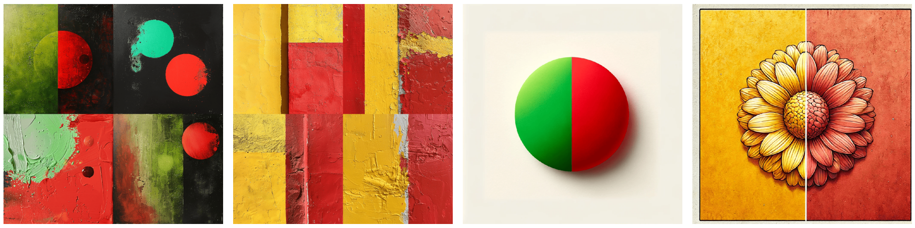
To test the composition of geometric shapes - called in visual semiotics the eidetic component of plastic categories - we asked Midjourney to generate an image of three white vertical lines on a black background, and then of three red horizontal lines on a white background (Figs. 1a and 1b). In the first generation[3], only two images contain three lines. In the first and fourth images in Fig. 1a, the white lines show chromatic irregularities, a textural effect that goes beyond the framework of a machine graphic composition and refers to the gesture of inscription. For the second generation: three images display the eidetic elements requested. However, the same textural display is found in the first and third images, with the latter also giving a glimpse of the limits of the visual object, *i.e.* the outline of a painting with an artistic purpose. Also, the background is not blanked, except for the second image in Fig. 1b. The same prompts with

[3] The numbering follows the Western reading order: left to right, and top to bottom.



(a) Midjourney (b) Midjourney (c) DALL·E (d) DALL·E

Fig. 1: Prompt: "Three vertical white lines on a black background" (a,c), "three horizontal red lines on a white background", (b,d).



(a) Midjourney (b) Midjourney (c) DALL·E (d) DALL·E

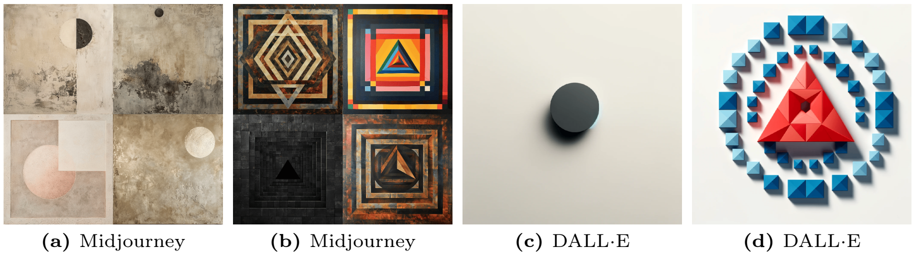
Fig. 2: Prompt: "A bright green spot next to an opaque red spot" (a,c), "a saturated yellow colour next to a desaturated red colour" (b,d).

DALL·E 3 (Figs. 1c and 1d) show a comparable difficulty in composing these plastic configurations, and they display a neutral, graphic texture[4].

We then tested the chromatic dimension with prompts involving fine color differentiation. The images generated by Midjourney in Figs. 2a and 2b, next to Figs. 2c and 2d for DALL·E, do not simply display colored patches, but patches of color whose texture can be recognized: in the first generation it is oil paint. In the last images pertaining to Fig. 2b, the texture of the inscription surface is irregular, as if it were a wall. Generally speaking, our control over the composition remains imperfect, because the qualities on the saturation of the colors are partly independent of the indications contained in the prompts. The same prompts submitted to DALL·E 3 showed the opposite pattern: the images were simple, neutral, almost exemplifying the colors requested.

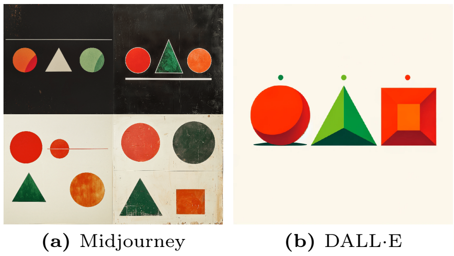
Finally, we tested the topological components. This category never stands alone, as it concerns the arrangement of shapes, figures and colors and their relationship to the overall representation space. We submitted two prompts: the first requiring an image composed of a circle in the upper right-hand part of the image, the second a triangle surrounded by squares arranged in a regular pattern (see Figs. 3a and 3b for Midjourney and Figs. 3c and 3d for DALL·E). In the

[4] Note that image generation is not deterministic: there is a large, yet inaccessible, set of images that can be generated. Our images are therefore only samples, which we assume to be representative, of the distributions associated with each prompt.



(a) Midjourney (b) Midjourney (c) DALL·E (d) DALL·E

Fig. 3: Prompt: "a small circle at the top right of the image on a neutral background" (a, c), "a triangle at the center of the image, surrounded by squares arranged in a regular manner" (b, d).

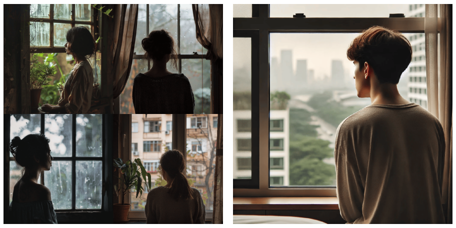


(a) Midjourney (b) DALL·E

Fig. 4: 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".

first generation, the position of the circle conforms to the prompt in two out of four images, but the chromatic and textural treatment of the background does not respect the required neutrality. DALL·E produces a different configuration for the second prompt: the image features a series of small squares that do not fit into each other, as they are arranged around the triangle, but adds an unrequested 3D effect.

We therefore decided to increase the complexity of our queries on the plastic categories: a sequence of three 2D objects - a red circle, a green triangle and an orange square - crossed by a white horizontal line. None of the images generated displays all the elements contained in the prompt. Midjourney also sought a particular balance, an effect of beauty (Fig. 4a). The colors often show an irregularity that implies a bodily gesture and work on the pigments; the background is almost never uniform or neutral, because it is already made up of a specific materiality. The tests we then ran with DALL·E produced objects that are not 2D because of a perspective effect and the presence of shadows: instead of a circle, DALL·E displays a sphere; the triangle is a tetrahedron, and the square is more a kind of square-based pyramid with a cut-off vertex (Fig. 4b).



(a) Midjourney **(b)** DALL·E

Fig. 5: Prompt: "a person looks out of the window".

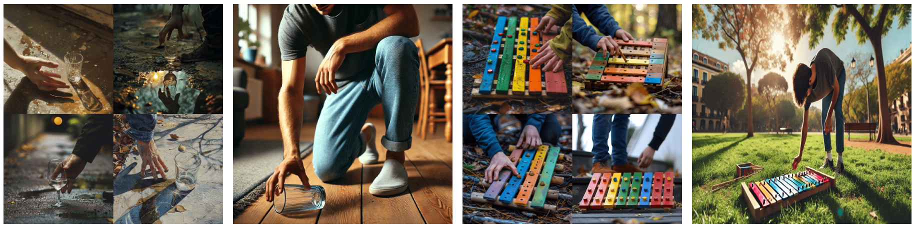
3.2 The multimodal translation

Midjourney and DALL·E produce adequate images for prompts that do not present actions, so we will not elaborate further on these tests. Indeed, if we ask for an object, a human, an animal, or a specific breed of animal, the visual results are adequate for the prompts. Generation may not be as effective in the production of groups of objects, people or animals, especially in light of the problems already found in the existing literature concerning counting competence. However, this task concerns rather the plastic dimension and not the multimodal translation of verbal actions into visual actions.

We tested different action verbs, which involve different verbal structures. As first prompt: "a person looking out of the window", without specifying what the person is looking at. We use the verb "to look" as a kind of zero degree of action, because in this case it does not involve a second actant and articulates an action in the present tense: our prompt has not specified the object of the look (Fig. 5). As in the examples from the experiment on spatiality, the act of looking, especially when involving a single human figure, poses no problems.

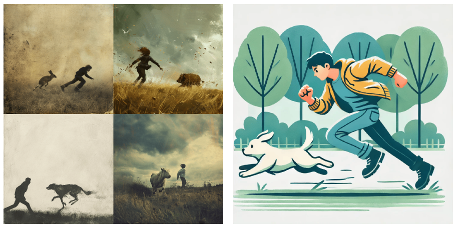
We tested transitive verbs that involve two actantial roles. The verb "to take", for example, involves a subject and an object complement. A prompt such as "a person picks up a glass from the ground" requires the coordination of two actants (a person and a glass) in relation to a spatial organization of the action (from the ground). Midjourney and DALL·E produce an image that respects the prompt (Fig. 6), and Midjourney offers several perspectives of the gesture to the observer. So we wondered whether the AI was just as effective at picking up a less common object, such as a xylophone (Figs. 6c and 6d), which moreover involves logistical difficulties: it is not picked up in the same way as a glass.

In the images produced by Midjourney, the person 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 carried out on glass: glass has a vertical configuration that is better suited to the gesture of picking up. It is also a smaller object, suited to a single hand gesture. The xylophone, on the other hand, imposes a horizontal configuration on the image, and its larger dimensions require a gesture structured in several stages, and a more complex configuration of the human body, requiring the



(a) Midjourney (b) DALL·E (c) Midjourney (d) DALL·E

Fig. 6: Prompt: "a person picks up a glass from the ground" (a,b), "a person picks up a xylophone from the ground" (c,d).



(a) Midjourney (b) DALL·E

Fig. 7: Prompt: "a person chases an animal".

actions of two hands. The same result was obtained by DALL·E , but it should be noted that this AI produced an image with a different, horizontal format.

The biggest difficulties arise when it comes to taking into account two animated actants linked by an action verb, as shown in Fig. 7. Midjourney is able to distinguish the direction of the action, but often transforms the action itself. This is the case of a person chasing an animal. In the first image produced by Midjourney, it is the animal that is chasing the person, in the second it is more of a confrontation between the two, in the third it is a reciprocal chase or game, in the fourth the two figures do not necessarily seem to be linked by any action. The same prompt submitted to DALL·E produces an adapted visual composition.

As part of the experiment on temporal articulation and aspectuality, we had to take into account an additional dimension of temporality: not just the time expressed by the verb, but the temporal point of view on the action as a whole. Generally speaking, we can distinguish between punctual, durative, iterative, inchoative (the beginning of an action) and terminative (the end) temporality.

First, we tested a terminative aspectuality through a prompt about a person who has finished eating a meal (Figs. 8a and 8b). In the images produced by Midjourney, temporality and aspectuality are not respected, because in none of the four cases are there any elements signaling that the meal is over. In the case of DALL·E, on the contrary, a boy looks towards the viewer smiling at an almost empty plate. We obtain similar results for a prompt describing an inchoactive action: a person about to start eating a meal (Figs. 8c and 8d). In the images produced by Midjourney, only the first image could fit the prompt. However,

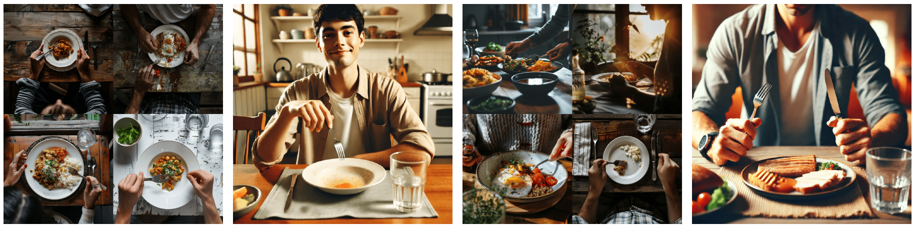

(a) Midjourney (b) DALL·E (c) Midjourney (d) DALL·E

Fig. 8: Prompt: "a person has finished eating a meal" (a,b), "a person about to start eating a meal" (c,d).

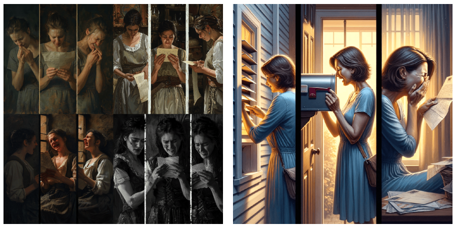

(a) Midjourney (b) DALL·E

Fig. 9: 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".

there is not enough evidence to say with certainty that the person is starting to eat, and the image is only adequate because we know the content of the prompt. In contrast, the image produced by DALL·E produces a configuration that visually translates the act of starting to eat: we see a person holding a fork and knife and then preparing to use them in front of a plate full of food.

We then tried to understand whether it was possible to generate a sequence of shapes capable of dividing the image into several sections, as shown in Fig. 9. We asked Midjourney to compose images showing three actions: in the first a woman picks up a letter, in the second she reads the letter, in the third she weeps with joy. The images generated by Midjourney effectively divide the representation space into three parts, but the action sequence is not adapted to the prompt. The image produced by DALL·E is sharper in the division of actions, but the process of finding the letter is extended to the first two sections, while the reading and emotional reactions are merged into a single image.

3.3 Enunciation

In relation to the dimension of enunciation, we first tested the degree of control with respect to the gazes of the characters represented, in accordance with the regime of enunciation-discourse. If we ask Midjourney to produce an image of a man looking towards the spectator, it will tend to produce an image that presents a man and spectators in the representation Fig. 10a. DALL·E , on the other hand, seems to be able to handle references to the spectator Fig. 10b.

If pronouns are used, Midjourney's image generation proves effective. A prompt that focuses on a man looking towards "us" produces formally correct results: in Fig. 10c, only the third and fourth images respect the enunciative configuration expressed by the prompt, but this is a remarkable result, because it implies that the model can handle pronouns that refer to the spectator's space outside the representation. We obtain the same result with DALL·E (Fig. 10d).

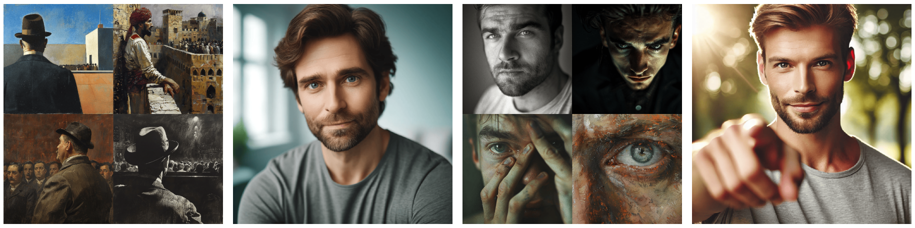
However, this simple configuration involves a single actor in the space. If we replicate the same enunciation-discourse regime involving interacting actors, the results are rarely appropriate (Figs. 11a and 11c). In the second generation in Fig. 11c, only the second and third images respect the prompt. And none of the images produced in Fig. 11a show an actor looking towards the viewer. Overall, these tests show that control over gazes is limited. Midjourney seems effective in handling pronouns involving the viewer, but managing visual configurations seems to encounter obstacles when the prompt describes several actors, probably as a result of the same limitations observed in relation to plastic categories.

The same prompts with DALL·E allow more precise control over the composition (Figs. 11b and 11d). The first image is remarkable in comparison with those obtained with Midjourney: although the two men do not necessarily display elements related to a fight, the different articulation of the glances is respected, allowing control over the composition of the different parts of the image.

The second criterion we tested concerned meta-pictural devices. We limited ourselves to testing the door device. The images generated from a prompt describing a woman reading behind an open door (Fig. 12a), present in two cases adapted configurations of space. If we try to increase the complexity of the actions, with a prompt describing two women talking in secret behind a half-open door (Fig. 12c), none of the four images produced shows a perfectly adapted articulation of space. However, we can see that the open door structures the space of the image according to the functions that Stoichita [12] attributed to this device: increasing the depth of vision of the image and allowing observation beyond the foreground. The same prompts with DALL·E produce comparable results: in Fig. 12b the prompt is respected, while in Fig. 12d a very special door is generated, with a void at the top to ensure the visibility of the two women.

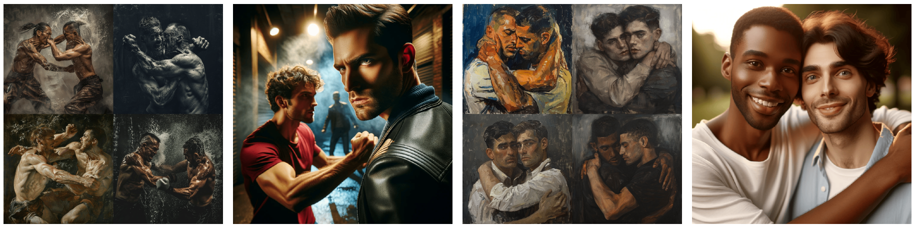
Throughout these tests, we chose to limit ourselves to the criteria that articulate the fundamental dimension of images and to test them discretely, so as to assess the effectiveness of two text-to-image models on each of these criteria. However, they can be combined to construct more complex operations: for example, a punctual action on an object located in a specific part of the image, and a terminative action set in another section. Currently, these articulations seem to us excessively complex because these models are not yet able to produce satisfactory images concerning the plastic and especially topological dimension, the most fundamental of the criteria of visual composition.

Remarks on CLIPScore. We also computed the CLIPScore of some of the generated images and their corresponding prompt. For instance, the CLIPScores



(a) Midjourney (b) DALL·E (c) Midjourney (d) DALL·E

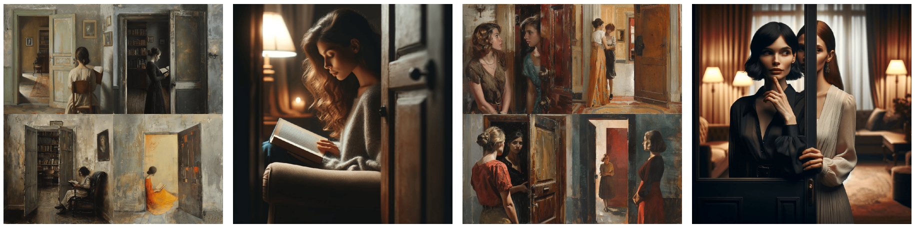
Fig. 10: Prompt: "a man looking at the spectator" (a,b), "a man who directs his gaze towards us" (c,d).



(a) Midjourney (b) DALL·E (c) Midjourney (d) DALL·E

Fig. 11: Prompt: "two men fight each other. one of them looks at us" (a,b), "two men embrace. they are looking towards me" (c,d).

between Figs. 11d and 12b and their prompts were respectively of 27.79 and 27.84. Despite the fact that these images are formally correct with respect to the prompts, these are low values. These values are also lower than other images that do not comply with the prompts, as in the case of Fig. 1c (30.58) and Fig. 1d (29.24). Such tests indicate that the correlation values produced by CLIPScore are unreliable. And even if they were reliable, the numerical value would not allow the identification of the formal trait that does not match the prompt.



(a) Midjourney (b) DALL·E (c) Midjourney (d) DALL·E

Fig. 12: Prompt: "a woman reads at the back of the room, behind a half-open door", "two women talk secretly at the back of the room, behind a half-open door".

4 Conclusion

We proposed a unified methodology for evaluating text-to-image models. Contrary to methods based on *e.g.* fidelity and alignment between prompts and images or CLIPScore evaluations, we used the concepts of visual semiotics. This allowed us to mobilize three fundamental dimensions of visual composition, articulated in a series of discrete criteria, in order to evaluate these models.

The first dimension is the plastic one: it concerns the formal reading of the image, beyond the possible presence of concrete figures (*e.g.* objects, people, landscapes, *etc.*). This category is divided into three fundamental criteria: the topological criterion concerns the organization of the image spaces according to parameters such as left vs. right, top vs. bottom, center vs. periphery; the chromatic one refers to the oppositions and similarities in the saturation and luminosity of the colors; the eidetic one concerns lines, shapes and contours, organized according to linear, curvilinear, fragmented strokes, *etc.*

The second dimension concerns the multimodal translation of actions expressed through verbal language into actions represented through images. We have identified three criteria: the type of action (absent, static, dynamic, structured), the actantial organization (verbs with only a subject, with a subject and an object complement, with a subject and several object complements plus spatial or other circumstantials), temporality and aspectuality (an action in its initial moment, or during its unfolding, or in its final phase).

The third dimension pertains to enunciation and is articulated in two criteria, the first of which concerns the control of the character’s gazes and the possibility of them looking toward the viewer. The second criterion concerns the meta-pictorial devices: elements such as doors, windows and mirrors, capable of structuring the space and inviting the spectator to a specific visual exploration.

We proposed a series of tests to evaluate the effectiveness of Midjourney and DALL·E relative to each of these criteria of semiotic analysis. Although the limited number of tests performed does not allow for statistical treatment, the criteria we used can be articulated to support statistical analysis: we described the types of prompts that could be used, presenting their general structure and a number of examples and variations.

Our results show that Midjourney produces images that are more aesthetically pleasing, often simulating materials and surfaces even at the expense of adherence to the prompt. In contrast, DALL·E is more effective in the control of composition: it performs slightly better with respect to plastic criteria, while it is much more effective with respect to the representation of temporalities (actions that have ended or that are about to begin). More generally, both models seem incapable of allowing accurate control of visual composition with respect to the plastic dimension, which is the most important dimension in images.

We believe that the dimensions, criteria and types of prompts we have identified might be highly useful for setting up quantitative studies. In addition, our paper lay the groundwork for the adaptation of other dimensions studied in semiotics and art history with respect to visual composition, which can be translated into criteria for evaluating text-to-image models.

Acknowledgments

E. D'Armenio is funded by the F.R.S.-FNRS research project "KineticEgo : Les performances identitaires dans les jeux vidéo et la réalité virtuelle. Une généalogie des médias visuels fondée sur le concept de mouvement. — Identity performances in video games and virtual reality. A genealogy of visual media based on the concept of movement", A. Deliege is funded by the F.R.S.-FNRS research project T.0065.22 - R.FNRS.5453 entitled "Pour une généalogie des formes visuelles. Approches sémiotique et computationnelle de larges collections d'images — Towards a Genealogy of Visual Forms: Semiotic and Computer-Assisted Approaches to Large Image Collections" of M.G. Dondero, who is funded permanently by the F.R.S.-FNRS. Besides, all the authors of the paper are affiliated with the University of Liège, Belgium.

References

1. Bianchi, F., Kalluri, P., Durmus, E., Ladhak, F., Cheng, M., Nozza, D., Hashimoto, T., Jurafsky, D., Zou, J., Caliskan, A.: Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale. In: ACM Conference on Fairness, Accountability, and Transparency (2023)
2. D'Armenio, E.: The Mediatic Dimension of Images. Visual Semiotics Faced With Gerhard Richter's Artwork. Visual communication (2022)
3. Dondero, M.G.: The Language of Images: The Forms and the Forces. Cham: Springer (2020)
4. Greimas, A.J.: Figurative Semiotics and the Semiotics of the Plastic Arts. New Literary History **20**(3) (1989)
5. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In: EMNLP (2021)
6. Jiang, H.H., Brown, L., Cheng, J., Khan, M., Gupta, A., Workman, D., Hanna, A., Flowers, J., Gebru, T.: Ai art and its impact on artists. In: AAAI/ACM Conference on AI, Ethics, and Society (2023)
7. Katirai, A., García, N., Ide, K., Nakashima, Y., Kishimoto, A.: Situating the social issues of image generation models in the model life cycle: a sociotechnical approach. arXiv (2023)
8. Lee, T., Yasunaga, M., Meng, C., Mai, Y., Park, J.S., Gupta, A., Zhang, Y., Narayanan, D., Teufel, H.B., Bellagente, M., Kang, M., Park, T., Leskovec, J., Zhu, J.Y., Fei-Fei, L., Wu, J., Ermon, S., Liang, P.: Holistic Evaluation of Text-to-Image Models. In: NeurIPS Datasets and Benchmarks Track (2023)
9. Li, B., Lin, Z., Pathak, D., Li, J., Fei, Y., Wu, K., Ling, T., Xia, X., Zhang, P., Neubig, G., Ramanan, D.: GenAI-Bench: Evaluating and Improving Compositional Text-to-Visual Generation. CVPR (2024)
10. Luccioni, A.S., Akiki, C., Mitchell, M., Jernite, Y.: Stable bias: evaluating societal representations in diffusion models. In: NeurIPS (2023)
11. Otani, M., Togashi, R., Sawai, Y., Ishigami, R., Nakashima, Y., Rahtu, E., Heikkilä, J., Satoh, S.: Toward Verifiable and Reproducible Human Evaluation for Text-to-Image Generation. In: CVPR (2023)
12. Stoichita, V.: The Self-Aware Image: An Insight Into Early Modern Meta-Painting. Cambridge University Press (1997)

13. Sukkar, A.W., Fareed, M.W., Yahia, M.W., Abdalla, S.B., Ibrahim, I., Senjab, K.A.K.: Analytical Evaluation of Midjourney Architectural Virtual Lab: Defining Major Current Limits in AI-Generated Representations of Islamic Architectural Heritage. Buildings **14**(3) (2024)
14. Wu, X., Hao, Y., Sun, K., Chen, Y., Zhu, F., Zhao, R., Li, H.: Human Preference Score v2: A Solid Benchmark for Evaluating Human Preferences of Text-to-Image Synthesis. arXiv (2023)
15. Zhou, S., Gordon, M.L., Krishna, R., Narcomey, A., Fei-Fei, L., Bernstein, M.S.: HYPE: A Benchmark for Human eYe Perceptual Evaluation of Generative Models. In: NeurIPS (2019)